Proceedings of PAMDAS 2025 International conference on Physical Asset Management and Data Science

Thursday, July 17, 2025 - Friday, July 18, 2025

Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Portugal



Organization

Endorsements

ISBN: 978-989-8331-19-9

Proceedings of PAMDAS 2025 International conference on Physical Asset Management and Data Science

This book is available at `https://pamdas.rcm2.pt`

Edited by:
Mateus Mendes, mmendes@isec.pt
José Torres Farinha, tfarinha@isec.pt
Ana Rita Malta, amalta@rcm2.pt
Hugo Raposo, hugo.raposo@isec.pt

RCM^{2+} — Research Centre for Asset Management and Systems Engineering
`https://www.rcm2.pt`

Coimbra, Portugal, 13 July 2025

ISBN: 978-989-8331-19-9

Organization

Executive Commission

Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Guilherme Ramos Pereira, DSPA - Data Science Portuguese Association, Portugal
Beatriz Santana, DSPA - Data Science Portuguese Association, Portugal

Steering Committee

Ana Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Annemieke Meghoe, University of Twente, Netherlands
António Paulino, Polytechnic University of Coimbra, Portugal
Beatriz Santana, DSPA - Data Science Portuguese Association, Portugal
Bernardo Tormos, Universitat Politècnica de València, Spain
David Baglee, Sunderland University, United Kingdom
Fernanda Brito Correia, Polytechnic University of Coimbra, Portugal
Guilherme Ramos Pereira, DSPA - Data Science Portuguese Association, Portugal
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
João Costa, Polytechnic University of Coimbra, Portugal
Jyoti Sinha, University of Manchester, United Kingdom
José Martinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jorge Almeida, Polytechnic University of Coimbra, Portugal
Luís Margalho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Nuno Lavado, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Pascoal Silva, Polytechnic University of Coimbra, Portugal
Rita Malta, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Wojciech Golebiowski, University of Life Sciences in Lublin, Poland

Scientific Committee

Alexandre Martins, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ana Camanho, University of Porto, Portugal
A. Paulo Coimbra, University of Coimbra, Portugal
Akbayan Bekarystankyzy, Narxoz University, Kazakhstan
Ana Alves, Polytechnic University of Coimbra, Portugal
Ana Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Anabela Gomes, Polytechnic University of Coimbra, Portugal
Annemieke Meghoe, University of Twente, Netherlands
António Paulino, Polytechnic University of Coimbra, Portugal
Bagdat Yagaliyeva, Satbayev University, Kazakhstan
Balduíno Mateus, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Batyrkhan Omarov, Al Farabi University, Kazakhstan
Beatriz Fidalgo, Polytechnic University of Coimbra, Portugal
Bernardo Tormos, Universitat Politècnica de València, Spain
Bruno Oliveira, Railway Competence Center, Portugal
Carlos Pereira, Polytechnic University of Coimbra, Portugal
Dammika Seneviratne, Tecnalia Research and Innovation, Spain
David Baglee, Sunderland University, United Kingdom
Felippe Souza, University of Beira Interior, Portugal
Filipe Amaral, Polytechnic University of Coimbra, Portugal
Felipe de Miguel Diez, University of Freiburg, Germany
Fernanda Brito Correia, Polytechnic University of Coimbra, Portugal
Fernanda Coutinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Gulfarida Tulemissova, Suleyman Demirel University, Kazakhstan
Hugo Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jânio Monteiro, University of Algarve, Portugal
Janet Lin, Lulea Technical University, Sweden
João Barata, University of Coimbra, Portugal
João Costa, Polytechnic University of Coimbra, Portugal
João Durães, Polytechnic University of Coimbra, Portugal
João Pombo, Huddersfield University, United Kingdom
João Rodrigues, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Jorge Almeida, Polytechnic University of Coimbra, Portugal
Jorge Barreiros, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
José Martinho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
José Sobral, Lisbon Polytechnic, Portugal
Jyoti Sinha, University of Manchester, United Kingdom
Kiumars Teymourian, Lulea Technical University, Sweden
Leonor Melo, Polytechnic University of Coimbra, Portugal
Luís Andrade Ferreira, Railway Competence Center, Portugal
Luís Margalho, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal Luís Oliveira, Tomar Polytechnic University, Portugal
Lyazzat Atimtayeva, Suleyman Demirel University, Kazakhstan

Manuel M. Crisóstomo, Institute of Systems and Robotics, Portugal
Marcin Kaminski, Lodz University of Technology, Poland
Mateus Mendes, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Miguel Vieira, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Moldir Kuatova, Al Farabi University, Kazakhstan
Nuno Cid Martins, Polytechnic University of Coimbra, Portugal
Nuno Lavado, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Pascoal Silva, Polytechnic University of Coimbra, Portugal
Philippe Mathieu, University of Lille, France
Rafael Raposo, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ramin Karim, Lulea Technical University, Sweden
Raúl Salas, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Ricardo Mateus, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Rita Malta, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Sebastião Pais, University of Beira Interior, Portugal
Simão Paredes, Polytechnic University of Coimbra, Portugal
Telmo Silva, University of Aveiro, Portugal
Torres Farinha, RCM^{2+} - Research Centre in Asset Management and Systems Engineering, Portugal
Wojciech Golebiowski, University of Life Sciences in Lublin, Poland

Contents

Preface **viii**

1 Vibration monitoring using contactless technology case study **1**

2 A Comprehensive Digital Twin-Based Maintenance Model for Enhanced Asset Reliability **10**

3 Optimizing asset management through reliability data analysis **21**

4 Medical Equipment Classification and Normative Methodologies: A Complete Approach to Standards and Metrology **32**

5 Fault Analysis of a biomass generator **45**

6 The Transformative Role of FPGA Platforms in Accelerating AI tasks **65**

7 Milk Prediction Using Conformation Traits as Privileged Information **76**

8 Deep Metric Learning for Face Recognition with Focus on Twins: Triplet Network Approach **90**

9 Integrating Machine Learning for Predictive Maintenance in Urban Bus Fleets using Oil Analysis **97**

10 Analysis of the state of a plastic injection machine **108**

11 Development of a Machine Learning model to classify different postures for the sitting position **123**

12 Ozone pollution levels in Portugal (2023): geostatistical interpolation **135**

13 A comparative analysis of AR equipment for physical asset management applications **150**

14 Raw material waste reduction in a food industry manufacturing unit: a case study on equipment changes and process optimization **166**

15 3D Modeling of Railway Components for Augmented Reality-Based Maintenance Training **179**

16 Optimizing the Requisition System for the Purchasing Process in the Industrial Sector: A Case Study on Digital Transformation and Con-

tinuous Improvement **191**

17 Pareto-Optimized Model Predictive Control for Real-Time 3D Trajectory Planning of Collaborative Robots **200**

18 Prediction and Mitigation of Wild Forest Fires using AI - Literature review **211**

19 Combining Clustering and Predictive Algorithms for Real Estate Market Trends – Literature Review **222**

20 Case Study - Implementation of the 7S Philosophy in a Food Industry **233**

21 Trajectory Optimization for Collaborative Robots via the Deep Deterministic Policy Gradient Algorithm **261**

22 Implementation of an MVP for Healthcare Consumables Management: A Concept Paper for a Pilot Project in ULS Coimbra **272**

23 Multi-Head Loss-Driven Brain Stroke Segmentation Using Auxiliary Mask Supervised Swin Transformer **284**

24 Exploration of Textual Information in Points of Interest Recommendation **294**

25 Storm damage and planting success assessment in Pinus pinaster stands using Mask-RCNN **303**

26 Analysis in Social Media Applied to Business Intelligence and Analytics: A Literature Review **315**

27 Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning **327**

28 The Role of Real-Time Navigation APIs in Green and Smart Mobility **343**

29 Maintenance Management in Electrical Grids with High Renewable Energy Penetration: a review of challenges and trends **354**

30 An electric mobility software ecosystem: Lifecycle Management Implementation **362**

31 A ROS 2-Based Modular Control Architecture Integrating Visual Servoing and Model Predictive Control for Robotic Marking **370**

32 LSTM Encoder-Decoder Framework for Incremental Time Series Forecasting in Power Transformers **387**

33 Machine Learning Approach to Fault Detection in Microclimate system at Residential and Non-residential buildings **399**

34 Scalable Data Lake Ingestion using Apache Hudi **413**

35 Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care **426**

36 ISO 9001 and ISO 14001 Recertification: A Case Study in a Thermoplastics Injection Factory 438

37 Sentiment analysis of patient complaints in healthcare systems using VADER: Can it contribute to a better service? 443

38 Case Study - Failure analysis of a raw mill in the cement industry: FMECA with fault tree and Ishikawa diagram 454

39 Challenges in Preparing a Procurement Specification for Hospital Equipment 467

40 Exploration of vibration signatures and patterns for bearings failure modes 476

41 Rethinking Hospital Asset Maintenance: Integrating Artificial Intelligence into CMMS and EAM Systems 503

42 Impact of 5S tool on reducing setup times - case study in a metalworking industry 513

43 Towards an Automated System for Pig Aggression Detection and Tracking 522

44 Performance of Projects of Construction Companies in Portugal 539

Preface

As technology evolves, industries become more and more dependent on data for decision-making. Concepts such as the Internet of Things become more prevalent in modern life and industrial settings. Sensors provide data which need to be analysed. Analytics algorithms provide insights not only into the past and present of the systems and processes, but also highlight trends and probable future directions.

PAMDAS 2025 international congress took place in Coimbra, days 17 and 18 of July 2025, under the lemma "When Intelligence Enhances Asset Management". The goal was to bridge the gap between industry and data science. While modern industry needs powerful algorithms and data science for enhanced decision-making, data science researchers also benefit from industrial connections where they can gather real data and find qualified research partners and customers.

This congress attracted speakers and authors from different continents, interested in Physical Asset Management and Data Science. Keynote speakers Prof. Diego Galar, from the Lulea Technical University, Sweden, and Prof. Celso Azevedo, from AssetsMan, France, contributed with their experience and wisdom to the the event.

The proceedings are now made available to the public under open access, for use and proper citation.

The editors,

Mateus Mendes
Torres Farinha
Ana Rita Malta
Hugo Raposo

Chapter 1

Vibration monitoring using contactless technology case study

Contribution #14
Cite as:
Abdu Shaalan, David Baglee, Derek Dixon. Vibration monitoring using contactless technology case study. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-14,
  title ={Vibration monitoring using contactless
     technology case study},
  author={Abdu Shaalan and  David Baglee and  Derek
     Dixon},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Vibration monitoring using contactless technology, UK case study

Abdu Shaalan[1], David Baglee[2] and Derek Dixon[3]

[1] University of Sunderland, St.Peters campus, Sunderland, SR60DD
[2] University of Sunderland, St.Peters campus, Sunderland, SR60DD
[3] University of Sunderland, St.Peters campus, Sunderland, SR60DD
*Abdu Shaalan
Abdu.shaalan@sunderland.ac.uk

Abstract

Maintenance management in modern manufacturing organisations has become an increasingly important and complex activity, due to the enhanced use of automation within the production infrastructure. The significance of effective maintenance activities is compounded by introducing new technologies to further enhance competitiveness and profitability in the global marketplace. These new technologies may be classified by Industry 4.0 and have primarily been applied to production facilities. Importantly, the agility of this technology allows it to be included as part of a maintenance plan, yet the rate of implementation in maintenance tasks is much reduced. This has been due to several reasons, which include, amongst others, a return on investment, training, a lack of strategic planning, and mistrust within the traditional workforce. Despite this, the further recent proposal of Industry 5.0 attempts to broaden the terms of reference by acknowledging the importance of any application being both human-centric and sustainable. The paper will use this case study research to discuss the effectiveness of collaboration, support, and a targeted approach in the use of advanced technology to solve maintenance problems. Conclusively, this paper identifies the importance of incorporating a bottom-up approach when introducing technology to the traditional sectors of engineering maintenance.

Keywords: Advanced Maintenance, Condition Monitoring, Vibration analysis, Contactless monitoring

1 Introduction

Maintenance management in modern manufacturing organizations has evolved into an increasingly critical and complex activity, driven by the growing reliance on automation within production infrastructure. The importance of effective maintenance practices is further amplified by the integration of new technologies aimed at enhancing competitiveness and profitability in the global market. These technologies, often categorized under the umbrella of Industry 4.0, have primarily been applied to production processes [1]. However, their potential extends to maintenance activities, where they can significantly improve efficiency and reliability. Despite this potential, the adoption of Industry 4.0 technologies in maintenance has been slower than anticipated, due to challenges such as return on investment concerns, training requirements, and resistance from the traditional workforce [2].

The recent emergence of Industry 5.0 has introduced a new dimension to this discussion by emphasizing the importance of human-centric and sustainable applications. While these high-level strategic concepts hold promise, their practical implementation remains a subject of debate.

Vibration analysis is considered a cornerstone in manufacturing assets' health assessment, enabling early failure detection and preventing catastrophic failures [3]. In general, vibration monitoring can be classified into two types: Contact and contactless. Both approaches introduce benefits and limitations [4,5].

Contact-based vibration monitoring

Relies on sensors that are mounted on the machinery, which translate the movement into signals to be processed by specialised software for analysis [6]. It has several advantages, as it provides high accuracy in

measurements, a wide frequency range of measurement that covers a wide variety of applications, efficiency in operating in harsh environments, and is considered a well-established technology in the realm of vibration monitoring [7]. On the other hand, it suffers from multiple disadvantages, including complexity in appropriate installation for accurate measurements, subject to wear and tear, difficulty in analysing the collected data, and the requirement of highly skilled staff [8].

Contactless vibration monitoring

Recent technological advancements are in pursuit of different approaches to contact-based monitoring, which introduced various contactless methods using Laser Doppler Vibrometer (LDV), Proximity probes, and Vision-based systems [9]. They offer a range of advantages, including the fact that they are non-intrusive and have no direct contact with the machines, which improves safety and reachability, high spatial resolution that allows for measuring multiple points simultaneously, and no degradation impact on the system [10].

While it overcome certain contact disadvantages, it developed disadvantages including high initial cost of the system, high dependency on surface nature, which affect quality with various types of surfaces, limited low frequency sensitivity, and high sensitivity to surrounding environment conditions whether it is surrounding vibration of, dust or some, which affects measurements quality [11].

This paper explores the use of a modular, human-centric approach to overcome barriers to the deployment of advanced technologies in maintenance activities. Drawing on rich data from industrial partners in the pharmaceutical, manufacturing, and environmental services sectors, the paper highlights the use of non-intrusive vibration monitoring equipment as a means of engaging the workforce in advanced maintenance techniques. The findings underscore the importance of collaboration, support, and a targeted approach in leveraging advanced technology to address maintenance challenges.

2 Case study background

Energy generation from waste company, a multinational utilities corporation specialising in sustainable wastewater and waste management solutions, operates six sites in the United Kingdom. Each site has distinct objectives, ranging from water network management to landfill remediation. This diversity in operations presents challenges in implementing a unified and efficient maintenance strategy. An informal discussion with a Suez employee revealed a persistent maintenance issue at one of the company's sites, which combines waste management with energy generation. The site utilizes landfill waste to fuel a furnace, which in turn powers a steam turbine for electricity generation.

The specific issue involved a motor and pump unit, which was part of a parallel system of three motor/pump combinations responsible for pumping water from a nearby river into a manifold. The manifold distributes water to the steam turbine system, where it is converted into steam to drive the turbine and generate power. Despite multiple interventions, the reliability of one of the motor/pump units remained problematic, with excessive vibration leading to frequent breakdowns. Various maintenance strategies, including corrective and preventive measures, were employed, but the issue persisted. Condition monitoring was also implemented, with an accelerometer attached to the motor housing to trigger an alarm and pre-emptive shutdown at a displacement of 3mm. However, frequent alarms and shutdowns led to frustration and necessitated further investigation.

3 Methodology

This study focuses on a single site with a specific maintenance issue, aiming to test the hypothesis that increased engagement in predictive maintenance techniques can be achieved through accessible, visual data using a Vision-Based System. The research also explores alternative techniques when location-specific methods fail to yield results. Given that a condition monitoring sensor was already installed on the problematic unit, an alternative approach was considered using proprietary technology. This approach involved a high-performance streaming camera with a pixel array of 1920 x 1200 and a resolution exceeding 2.3 megapixels. The camera, capable of capturing up to 1300 frames per second, was paired with acquisition software licensed to a laptop,

making the system mobile and non-invasive.
The technology enables the identification of deflection, displacement, movement, and vibration, with the ability to refine or amplify images for immediate visual engagement. Once synthesized, the data can be presented in an accessible format to facilitate rapid analysis and decision-making. The system allows for the application of a Region of Interest (ROI) to any area of the captured image, enabling detailed waveform, orbit, and spectra analysis. This flexibility allows for the investigation of additional areas of interest post-data capture.
Units of measurement on graphical annotations are:

- Frequency – Hertz (Hz)
- vibration displacement - micrometres; (1 μm = 10^{-6} m)
- vibration velocity - millimetres per second; (mm/s)
- vibration acceleration - metres per square second; (mm/s)

4 Results

An initial on-site review of the amplified images revealed that all three pumps exhibited movement and vibration. Figure 1 provides a front view of the coolant motor/pump system, highlighting the three motor/pump units connected to the pipework transporting river water to the cooling system. The scaled motion vector indicates the direction of vibration for pump 1 and manifold 2.



Figure 1: A front view of the coolant motor/pump system

Figure 2 presents the near side of pumps 2 and 3, with manifold 3 in view. The ROI for pump 2 and manifold 3 has been annotated, revealing that while pump 2 displays an RMS velocity of 2.2 mm/s, manifold 3 exhibits a higher RMS velocity of 3.2 mm/s. This finding was unexpected, as previous maintenance efforts had focused solely on pump 2.

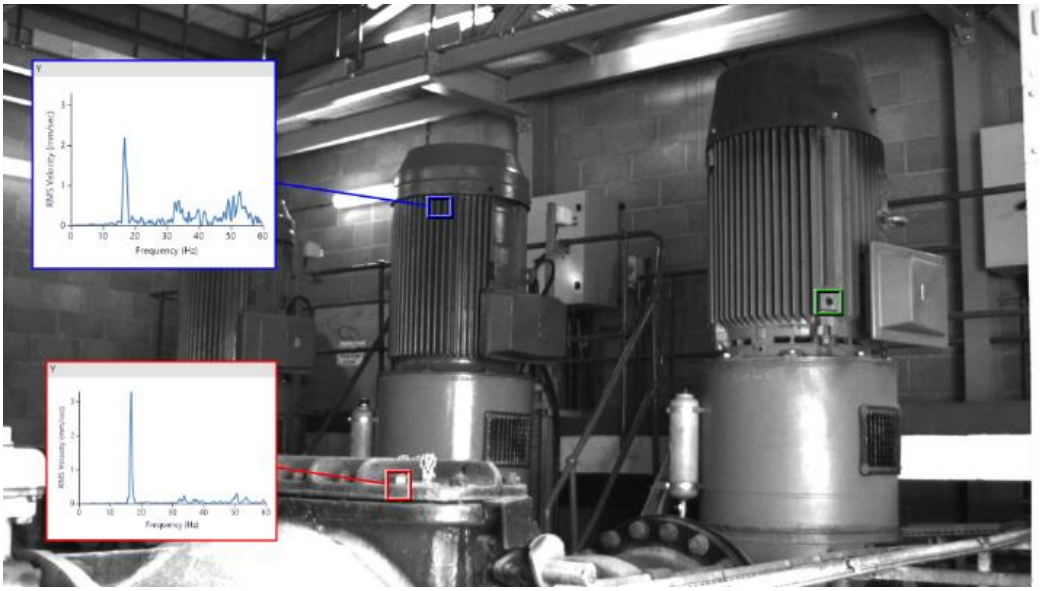


Figure 2: Annotations representing the RMS

Figures 2, 3, and 4 demonstrate an issue that was previously undetected. Manifold 3 displays vibration characteristics that are in excess of pump 2, where the original maintenance issues were focused. The displacement scale shown in Figure 3 displays a cyclic displacement of approximately 40 μm. This was previously beyond the scope of other maintenance investigations. Figure 3 also has the displacement/time graph for manifold 3 with an extended scale for clarity.


Figure 3: Pump 2, 3, and manifold 3 with displacement vector annotations

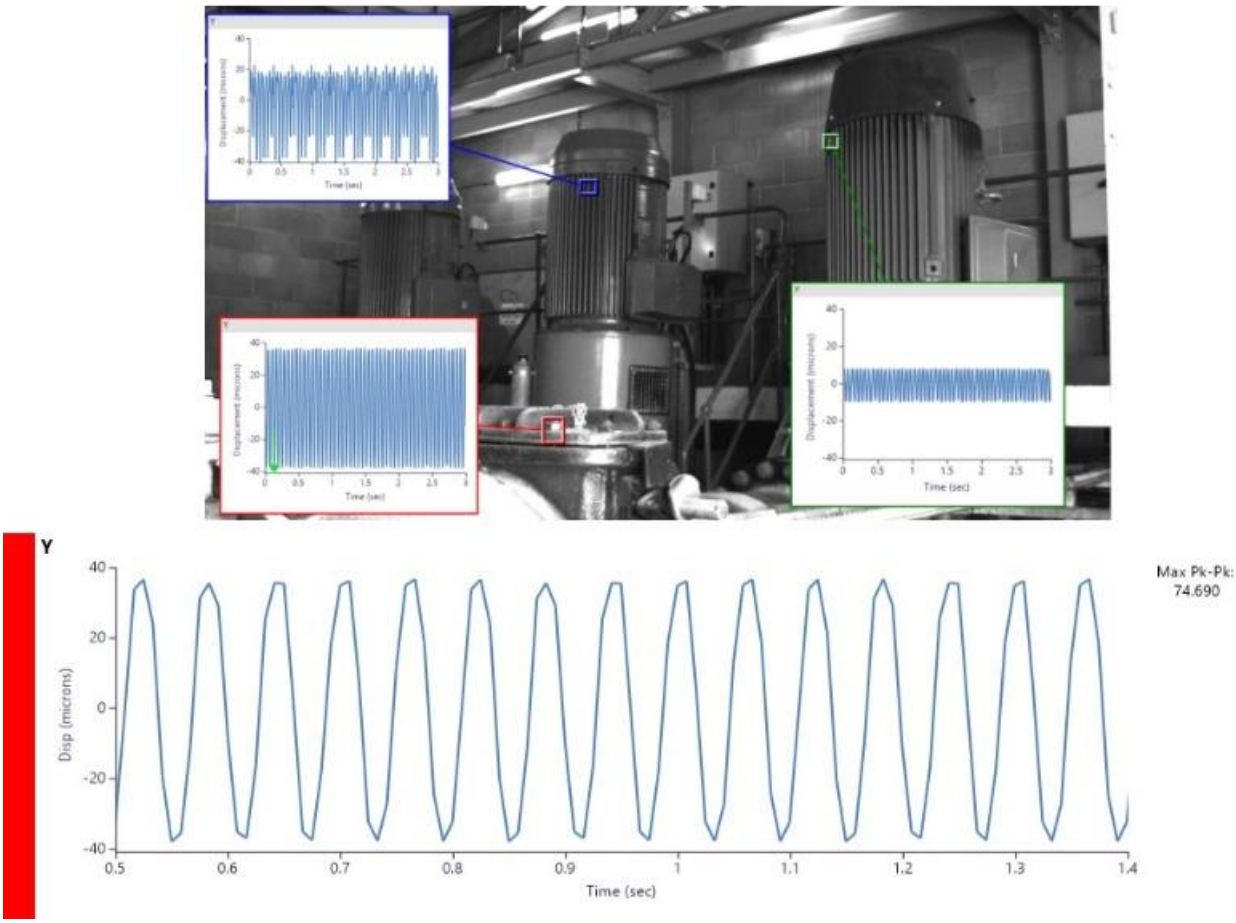


Figure 4: Pump 2, 3, and manifold 3 with displacement time axis annotations.

Due to the portable nature of the technology, several different perspectives of the asset were captured. This can be noted in Figures 4 and 5, and this proved particularly interesting. Figure 4 shows a support bracket and fixed point for the pipe system with the ROI centred on the bracket. The enlarged graph shows a pk–pk displacement in the Y axis of 278 microns. This was the greatest displacement captured across the entire investigation and can be compared to the relatively large displacement of manifold 3.

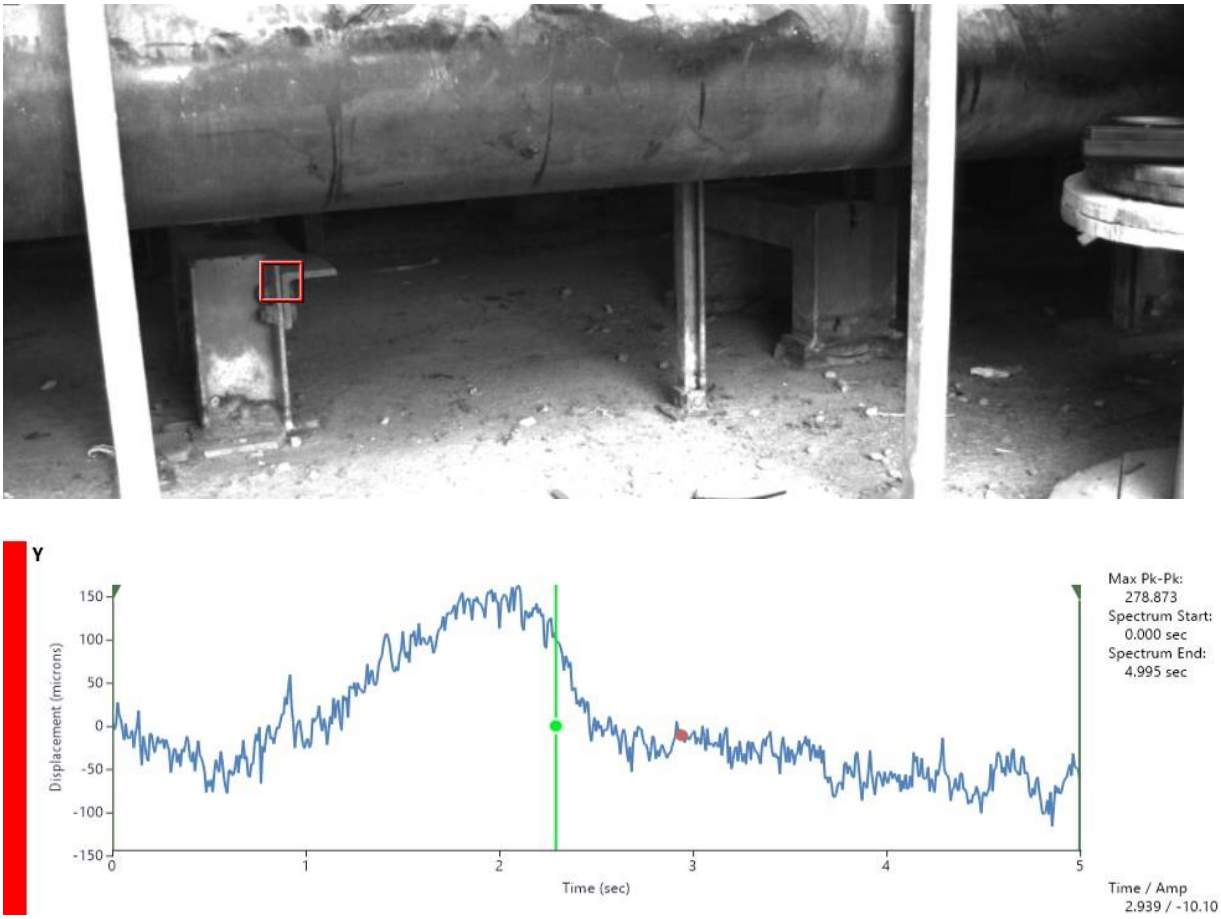


Figure 5: Support bracket and foundation of pipeline 3.

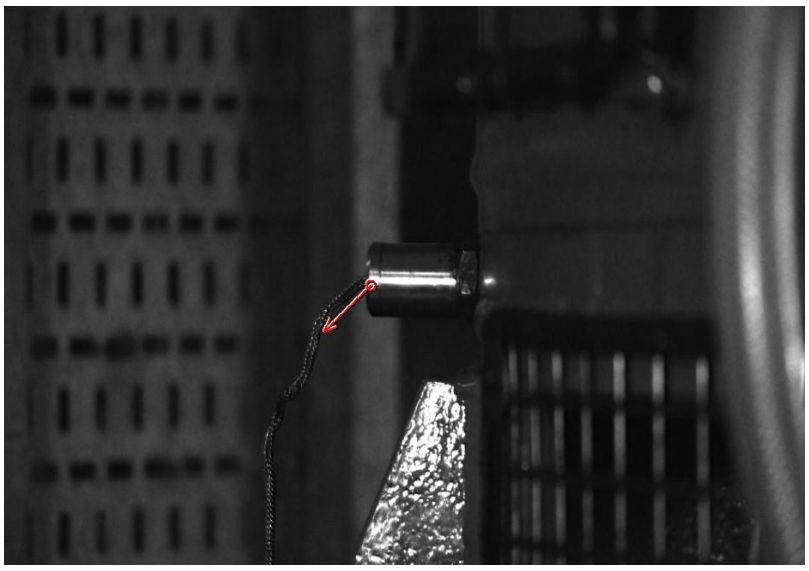


Figure 6: Accelerometer located on pump housing 2.

5 Discussion

The findings from this case study highlight the potential of motion amplification technology as a non-intrusive and sustainable alternative to traditional vibration analysis methods. Unlike constant monitoring techniques, this approach allows for targeted investigations, reducing the need for continuous sensor deployment. However, the technology is limited to lower frequency vibration faults and may be subject to mechanical excitation. Additionally, interpreting the results can be challenging, as standards for this technology are still evolving. Another limitation noted is the inability to inspect internal failures as the system needs to have a visual representation of the movement, which shows high efficiency in structural damage, connections, and pipes work, and any visible parts of the equipment in sight.

Despite these limitations, motion amplification cameras offer a valuable complementary technique to more accurate sensor-based measuring devices, particularly in identifying excitation issues that may extend to the broader asset infrastructure. Ease of measurement analysis is considered on the strength points of the system as it does not require extensive experience in carrying out the measurement of the analysis.

6 Conclusion

This paper underscores the importance of incorporating a bottom-up, human-centric approach when introducing advanced technologies to traditional engineering maintenance sectors. The case study demonstrates the effectiveness of motion amplification cameras in engaging the workforce and providing actionable insights into

complex vibration issues. While technology should be viewed as a complementary tool rather than a replacement for traditional methods, its ability to capture full-field vibration data offers significant advantages in diagnosing and addressing maintenance challenges. Future research should focus on addressing the current limitations of this technology and exploring its potential applications in other industrial contexts.

References

[1] Liu, Y. and Xu, X., 2017. Industry 4.0 and cloud manufacturing: A comparative analysis. *Journal of Manufacturing Science and Engineering*, *139*(3), p.034701.

[2] Lee, J., Bagheri, B. and Kao, H.A., 2015. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing letters*, *3*, pp.18-23.

[3] Randall, R.B., 2021. *Vibration-based condition monitoring: industrial, automotive and aerospace applications*. John Wiley & Sons.

[4] Bagri, I., Tahiry, K., Hraiba, A., Touil, A. and Mousrij, A., 2024. Vibration Signal Analysis for Intelligent Rotating Machinery Diagnosis and Prognosis: A Comprehensive Systematic Literature Review. *Vibration*, *7*(4), pp.1013-1062.

[5] Yang, Y., Dorn, C., Mancini, T., Talken, Z., Kenyon, G., Farrar, C. and Mascareñas, D., 2017. Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification. *Mechanical Systems and Signal Processing*, *85*, pp.567-590.

[6] Mohd Ghazali, M.H. and Rahiman, W., 2021. Vibration analysis for machine monitoring and diagnosis: A systematic review. *Shock and Vibration*, *2021*(1), p.9469318.

Nahavandi, S., 2019. Industry 5.0—A human-centric solution. *Sustainability*, *11*(16), p.4371.

[7] Romanssini, M., de Aguirre, P.C.C., Compassi-Severo, L. and Girardi, A.G., 2023. A review on vibration monitoring techniques for predictive maintenance of rotating machinery. *Eng*, *4*(3), pp.1797-1817.

[8] Ahmed, H. and Nandi, A.K., 2020. *Condition monitoring with vibration signals: Compressive sampling and learning algorithms for rotating machines*. John Wiley & Sons.

[9] Feng, Z., Liang, M. and Chu, F., 2013. Recent advances in time–frequency analysis methods for machinery

fault diagnosis: A review with application examples. *Mechanical systems and signal Processing*, *38*(1), pp.165-205.
[10] Ghobakhloo, M., 2018. The future of manufacturing industry: a strategic roadmap toward Industry 4.0. *Journal of manufacturing technology management*, *29*(6), pp.910-936.
[11] Goyal, D. and Pabla, B.S., 2016. The vibration monitoring methods and signal processing techniques for structural health monitoring: a review. *Archives of Computational Methods in Engineering*, *23*, pp.585-594.
[12] Ye, X.W., Dong, C.Z. and Liu, T., 2016. A review of machine vision-based structural health monitoring: methodologies and applications. *Journal of Sensors*, *2016*(1), p.7103039.

Chapter 2

A Comprehensive Digital Twin-Based Maintenance Model for Enhanced Asset Reliability

Contribution #19
Cite as:
M. A. Amiruddin, Kassandra A. Papadopoulou, Jyoti K. Sinha. A Comprehensive Digital Twin-Based Maintenance Model for Enhanced Asset Reliability. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-19,
  title ={A Comprehensive Digital Twin-Based
     Maintenance Model for Enhanced Asset Reliability},
  author={M. A. Amiruddin and  Kassandra A.
     Papadopoulou and  Jyoti K. Sinha},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

A Comprehensive Digital Twin-Based Maintenance Model for Enhanced Asset Reliability

M. A. Amiruddin[1,*], Kassandra A. Papadopoulou[2], Jyoti K. Sinha[1]

[1] School of Engineering, The University of Manchester, M13 9PL, UK
[2] Alliance Manchester Business School (AMBS), The University of Manchester, M15 6PB
*Corresponding author
mohamadafiqamiruddin.parnon@postgrad.manchester.ac.uk

Abstract

Digital Twin-based maintenance has emerged as a promising solution to address the challenges posed by traditional reactive maintenance approaches, frequently resulting in increased downtime and elevated costs. This study introduces a comprehensive maintenance framework that integrates asset criticality assessments, leveraging expert insights and quantitative techniques. It also includes Condition-Based Monitoring, which fuses real-time sensor data with historical reliability records, alongside Digital Twin calibration conducted through simulations and parameter estimation. The framework employs advanced signal processing for data pre-processing, statistical analysis to establish baselines, and "what-if" simulations for predictive maintenance, all validated through pilot studies. While this work is conceptual, the paper argues that the unified model can significantly enhance asset reliability and minimise maintenance interventions through continuous calibration and threshold limit detection. This research contributes to the state-of-the-art by offering a scalable, integrated approach that bridges theoretical Digital Twin models with practical maintenance applications, setting the stage for further exploration and real-world implementation. The proposed model is particularly relevant for industries seeking to improve asset management, especially in small and medium-sized enterprises, where data availability is often limited and can obstruct effective Digital Twin execution.

Keywords: Digital Twin; Maintenance; Condition-Based Monitoring; Predictive Maintenance; Asset Reliability; Criticality Assessment; Field Validation

1 Introduction

Implementing Digital Twin (DT) in maintenance applications marks a significant advancement that has enhanced productivity. The emergence of Industry 4.0 in the late 2010s has further accelerated progress in the maintenance sector by introducing modern technologies, including advanced computing and Artificial Intelligence (AI), which facilitate communication and production through the Internet. Integrating these technologies into maintenance systems can substantially improve machine reliability and operational performance while mitigating the risk of sudden breakdowns and catastrophic failures in industrial settings. Nevertheless, research [1], [2], [3] has highlighted considerable gaps and challenges that impede the development of comprehensive and practical solutions for incorporating these technological tools into existing maintenance models. This limitation has resulted in a gradual progression in maintenance innovations. This paper seeks to address these gaps by investigating strategies to effectively integrate DT, capitalising on advancements from Industry 4.0 to establish a more resilient and reliable maintenance model.

Before integrating DT into maintenance practices, traditional methods primarily focused on detecting, diagnosing, and predicting asset health and Remaining Useful Life (RUL). These conventional approaches heavily relied on historical data and manual interventions, often lacking the

benefits of real-time monitoring, which can yield critical insights for fault diagnosis and prognostics [4]. As a result, these methods often produce less reliable findings [5], [6]. To overcome these limitations, researchers have advanced the implementation of virtual models in maintenance models. Their adaptation in the framework aims to access and analyse real-time and historical data seamlessly, providing actionable insights for maintenance activities [7]. Traditional methods also face challenges, such as data availability, along with system interoperability, scalability, and transferability [8]. These obstacles can impede the smooth integration of data from diverse sources and hinder the progress of comprehensive health monitoring strategies. Adopting this technology mitigates these shortcomings by offering advanced capabilities such as data modelling, connection modelling, and virtual construction within maintenance frameworks. This facilitates the development of more robust and adaptable health management solutions in maintenance strategies [7], [8]. Moreover, traditional maintenance approaches depend exclusively on physical detection, diagnostics, and prognostics to assess RUL and the health status of equipment, often resulting in elevated testing costs. Minor changes in the system, such as increased bearing speed, necessitate immediate monitoring, further driving up expenses. In contrast, digital replicas of assets offer a virtual environment that accurately replicates physical objects and their material properties, facilitating the simulation of various scenarios without impacting the operation of the physical assets. This capability can reduce costs while improving the accuracy and efficiency of maintenance strategies. These indicate that integrating virtual modelling into the maintenance paradigm significantly enhances performance by overcoming the limitations of conventional methods.

DTs have emerged as a transformative technology with the potential to revolutionise maintenance practices across various industrial sectors. However, several challenges and gaps hinder their seamless integration and effectiveness in these fields. One significant limitation is the lack of comprehensive research on integrating key components within the predictive maintenance domain. While some scholars have proposed studies of hybrid frameworks that connect virtual models with prognostics, their efforts have predominantly concentrated on areas such as fault monitoring, classification, anomaly detection, and fault diagnostics rather than on prognostics itself [1], [3], [9]. The link between digital models and maintenance strategies remains tenuous, primarily due to the absence of a structured framework, which limits the effective utilisation of their benefits in maintenance strategies. Xiao *et al.* [7] noted the lack of standardisation in DT-driven maintenance, which hampers research advancement. Additionally, various researchers, including Liu *et al.* [1], Rasheed *et al.* [2], and van Dinter *et al.* [3], have called for the establishment of standard methods, models, and frameworks for developing these systems, especially within the context of maintenance across diverse industries and regions. The lack of standardisation may lead to inconsistent tools and methods utilised across different domains [9], resulting in varied implementations by various companies.

Another significant challenge of the DT-based maintenance model is its predominant emphasis on production planning, often at the expense of maintenance optimisation. This oversight has hindered the advancement of maintenance within asset management. For example, the study by Feng *et al.* [10] developed a predictive maintenance framework for the offshore oil and gas sector to minimise maintenance costs while fulfilling production requirements. Although their model offers significant insight into amalgamating predictive maintenance into production environments, it does not fully capture the effects of disturbances or delayed preventive maintenance tasks on assets. This observation suggests the significant potential of DTs to improve maintenance practices—one that the model proposed in this paper seeks to address. Nevertheless, their integration into these areas is still in its early stages, and they encounter various challenges.

The increasing interest in advanced maintenance frameworks has remained steady from 2021 to 2023, aiming to address existing gaps and challenges within this field [11]. This heightened focus is driven by various factors, notably advancements in technology. A significant approach to overcoming the gaps involves embracing Industry 5.0, where digital models facilitate human-centred and sustainable manufacturing practices [12]. In their bibliometric review of DTs within the supply chain, Lam *et al.* [12] and Warke *et al.* [13] note that from 2014 to 2020, DT technology became increasingly integrated with Industry 4.0. Researchers have combined it with big data and machine learning,

introducing new technologies like edge computing, cloud-based systems, and blockchain. The 2020 and 2021 lockdowns have further intensified interest in virtual models as tools for fostering resilience and sustainability, prompting a shift in research focus on Industry 4.0, particularly smart manufacturing. Conversely, Gao *et al.* [14] observe that while many research trends are directed at integrating critical elements such as systems, simulations, big data, and AI, their application remains predominantly confined to laboratory environments. This disconnection between theoretical advancements and practical implementation causes a challenge for adaptation in real-world industrial settings.

While current research seems promising, a proper framework for maintenance has yet to be established, leading to the purpose of this study. Addressing these gaps is important to advancing the maintenance field, and its implications could adversely affect the professionals who use this technology. One of its implications is that inconsistent tools and methods used by different organisations could impede the communication and sharing of information [9], potentially resulting in increased downtime, higher costs, or even failure to prevent critical issues in time. Another consequence of this unanswered gap could lead to the risk of detecting faults without understanding their causes [7], leading to ineffective maintenance actions that cause increased mean time to repair and time to restore the system. Therefore, this paper proposes a comprehensive framework that integrates a standardised approach to optimising maintenance tasks, balancing computational efficiency with accuracy, and evaluating the effectiveness of maintenance plans in dynamic environments. Consequently, this study aims to develop a unified methodology that identifies faults, provides an in-depth understanding of their nature, and adapts to changes to enhance overall system performance, including reliability and efficiency.

This paper is structured into five sections. Section 1 introduces the challenges in traditional maintenance and the potential of DT. Section 2 presents the proposed conceptual maintenance model, integrating asset ranking assessment, reliability data analysis, Condition-Based Monitoring (CBM), DT, and field validation. Section 3 concludes the study.

2 Proposed Conceptual Maintenance Model

Several theoretical frameworks related to the DT system have been developed to address various issues, such as minimising downtime and total maintenance costs in production systems [15], [16], [17], [18]. These frameworks, moreover, have advanced DT development in specific areas, including production lines, healthcare systems, construction, and energy plants. Nevertheless, there is a lack of connection between this system and maintenance components to achieve unified outcomes. Therefore, this section dives into developing proposed digital-based maintenance models to bridge these research gaps and enhance the effectiveness of their implementation in the maintenance domain.

2.1 Overview of the Conceptual Model

The initial theoretical framework is grounded in the three-dimensional DT framework proposed by Grieves [15], [19]. This framework identifies three essential components: a physical asset, an asset's data—whether historical or real-time—and a connection that links the two. However, it overlooks maintenance aspects associated with operations. As a result, users often tackle issues without fully understanding the underlying problems, leading to suboptimal operational outcomes that can increase both system downtime and operational costs. To address this limitation, this study proposes a unified maintenance model that seamlessly integrates critical elements of digital models and maintenance. This model encompasses asset criticality assessments, reliability data analysis, CBM, digital integration, and field validation. Figure 1 and Figure 2 illustrate the proposed maintenance model's step-by-step process.

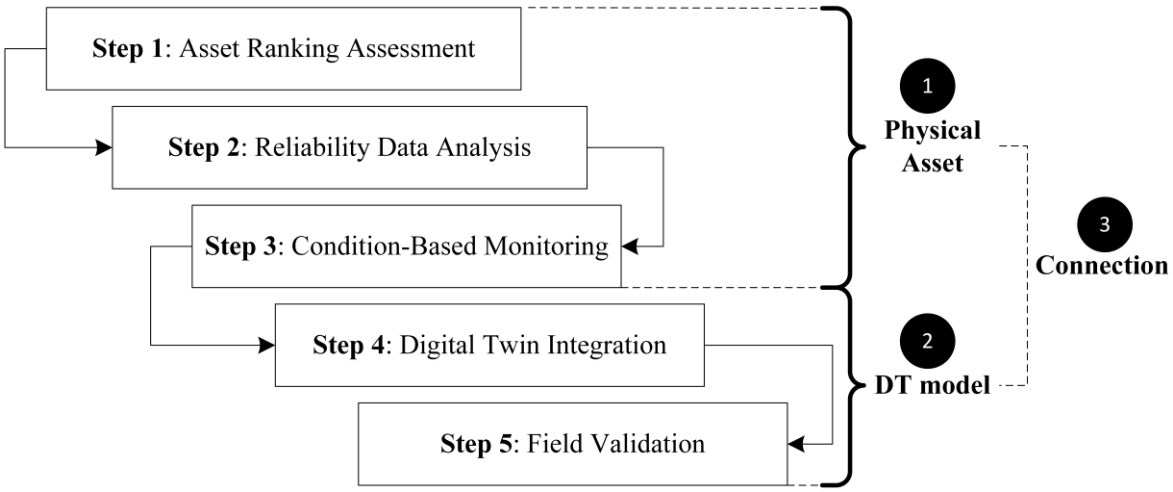


Figure 1: DT-based maintenance model process stages with three DT components.

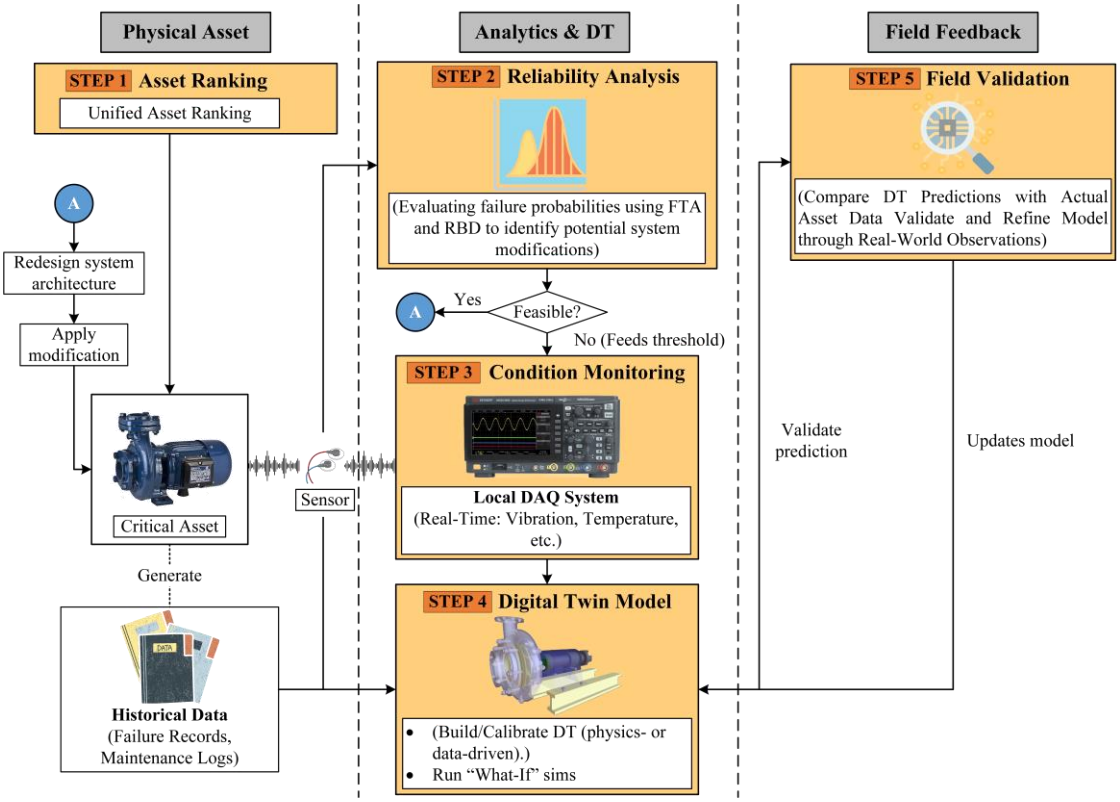


Figure 2: Conceptual Diagram of The Proposed Maintenance Model.

2.1.1 Asset Ranking Assessment– Step 1

The asset ranking assessment prioritises assets by evaluating their importance, potential failure impact, and associated risks. This process ensures that the proposed maintenance model concentrates

exclusively on critical machines for further analysis, promoting cost-effective asset management, optimising maintenance overhead, reducing downtime, and enhancing overall plant reliability and safety. To accomplish this, the maintenance model proposed in this study utilises the asset ranking assessment that adopts a multi-method approach that integrates both qualitative and quantitative analyses. Figure 3 illustrates the overall picture of the unified asset ranking framework.

It begins with the Delphi interview method, which gathers tacit knowledge from industry experts. In these interviews, carefully structured questionnaires are provided to the experts to define the criteria and sub-criteria that help identify the potential criticality of various assets.

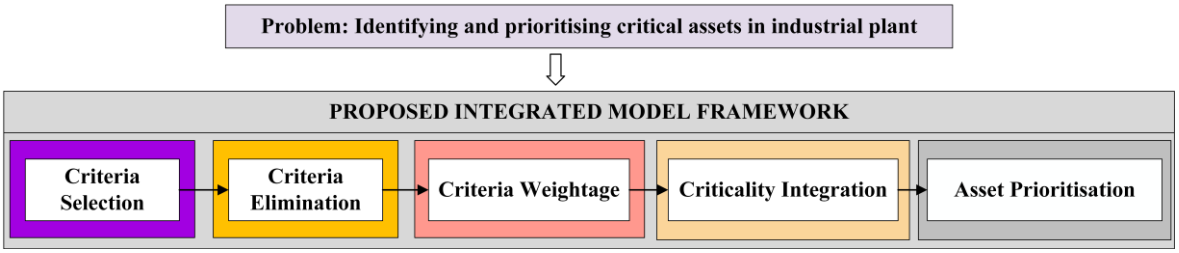


Figure 3: Integrated Criticality Model Framework

Following the Delphi study, the criticality assessment employs the Analytical Hierarchy Process (AHP) for quantitative data analysis. Experts' responses gathered during the Delphi study are processed through a pairwise comparison matrix in AHP, allowing for determining the Priority Vector (PV) for each criterion and sub-criterion. The PV value reflects the relative importance of each factor, effectively addressing the common issue of treating all criteria as equal—a limitation frequently encountered in traditional single-method approaches. Once the PV is established, the criticality framework further enhances the analysis by incorporating the PV value and additional criticality elements, such as severity and occurrence. This integration culminates in calculating a Risk Priority Index (RPI), which serves as a composite metric for ranking and prioritising assets, thereby identifying those requiring more immediate or intensive maintenance interventions. To further prove the concept, Amiruddin, Papadopoulou and Sinha [20] utilise a case study involving wind turbines that include both direct-driven and gear-driven types to illustrate their effectiveness.

This unified asset ranking analysis is pivotal in guiding subsequent maintenance activities and data monitoring efforts. By capturing the critical assets in the plant, the proposed maintenance model can proceed to the following stage, where the CBM activities can be set up to monitor acceptable limits.

2.1.2 Reliability Data Analysis – Step 2

The proposed maintenance model leverages reliability analysis to quantify failure probabilities, enhancing asset mechanisms and improving the system's overall reliability. This phase generates outputs by analysing historical performance records, failure logs, and sensor data collected throughout the asset's operational lifespan. Techniques such as Fault Tree Analysis (FTA) and Reliability Block Diagrams (RBD) are utilised in this evaluation. The process begins by accessing historical records of the critical asset identified in Step 1. The FTA then identifies the top event (system failure) and decomposes it into basic events corresponding to individual component failures. After structuring the hierarchy, failure probabilities for each basic event are assigned based on historical failure rates, such as Rate of Occurrence of Failure (ROCOF). Alternatively, a reliability assessment can be conducted by constructing an RBD model. This involves creating interconnected blocks representing the components and assigning reliability values derived from historical data. Utilising either FTA or RBD techniques, the overall system reliability is calculated based on these assigned values, enabling the identification of the system's weakest link. FTA and RBD analyses are conducted per BS EN IEC 61025 [21] and BS EN 61078 [22] standards.

Drawing from the outcomes of FTA or RBD, the proposed model effectively identifies critical failure points and assesses whether system architecture modifications can enhance reliability. If feasible modifications—such as redesigning components, optimising operational parameters, or upgrading materials—are identified, they are implemented to reduce the likelihood of asset failures. However, when modifications are deemed impractical due to technical or cost limitations, the process would advance to the CBM stage in Step 3. This flow to CBM is primarily driven by the inherent failure modes associated with the asset. These inherent failures can restrict the effectiveness of modifications and redesigns, as they may not fully resolve the underlying failure mode. Therefore, the most effective strategy is to enhance monitoring and maintenance practices. Implementing the CBM strategy can detect signs of faults better and respond before absolute failure occurs.

To dynamically update failure probabilities, live sensor data from CBM is incorporated into reliability models. Techniques such as Bayesian updating and weighted moving averages adjust historical failure probabilities based on real-time information. For example, when a sensor detects abnormal behaviour in a machine, the algorithms may assign greater significance to that observation, effectively increasing the failure probability for that component within the FTA or RBD model. This revised failure probability is then reintegrated into the reliability model to reflect the changes accurately. For instance, if the FTA or RBD model previously estimated the system's top event—overall reliability—to have a 5% chance of failure based on older probabilities, the updated data might indicate an increase to 7% due to the heightened failure probability of a specific component. This revised measure is subsequently used by the DT to predict more accurate, near-real-time failure risks and to inform proactive maintenance strategies.

2.1.3 Condition-Based Monitoring – Step 3

CBM employs condition monitoring to assess the performance and health of assets continuously. It utilises sensors and a data acquisition system to collect operational data, which is then compared with a baseline established from historical data. This approach provides immediate insights into an asset's condition. Unlike traditional time-based maintenance, which follows a fixed schedule, CBM prompts interventions only when indicators signal potential performance degradation or anomalies. Figure 4 illustrates the process sequence of the strategy in the proposed maintenance model in this study.

In the proposed maintenance model, CBM is employed for machines designated as critical during the asset ranking evaluation in Step 1, and this system is not amenable to enhancement via reliability assessment. The process begins with collecting historical data, encompassing past performance records, failure logs, and sensor readings gathered throughout the asset's operational life. This data is then processed using a computational algorithm that effectively filters out noise, normalises values, and segments continuous data streams into more manageable intervals. Comprehensive statistical analyses are then performed to establish a performance baseline of metrics such as average vibration levels, temperature ranges, or pressure values under normal operating conditions. These metrics define what is considered normal and facilitate the identification of deviations that could indicate potential failure concerns. In addition to historical data, real-time sensor data is collected by installing and configuring sensors on these critical assets. This involves setting up sensors, such as accelerometers, to monitor operational data, including vibration, temperature, and pressure. It is essential to ensure these sensors are securely mounted and properly connected to a local Data Acquisition (DAQ) system.

Once completed, the processed data from historical records is passed to the limit assessment stage, where live sensor readings are continuously monitored to determine whether they fall within an acceptable limit. If the measured data exceeds the threshold limit, the system triggers alarms to alert users, prompting them to conduct further investigations. Once the root cause of any asset faults is identified, maintenance actions are scheduled to ensure timely interventions when necessary.

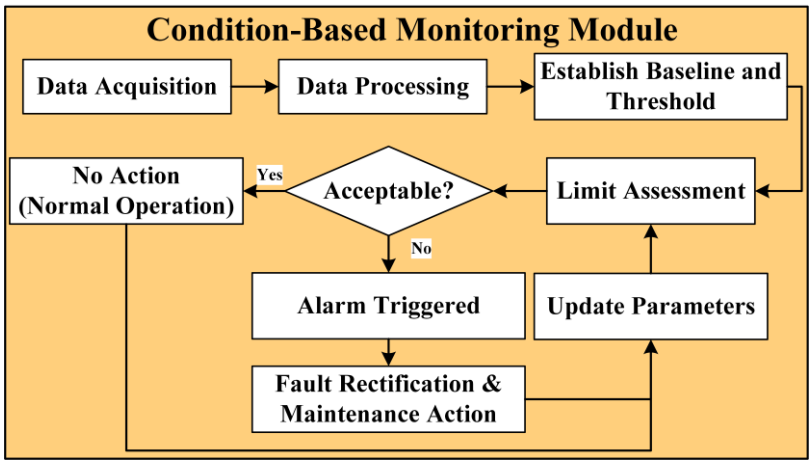


Figure 4: Condition-Based Monitoring Module

Additionally, the results of maintenance activities, along with updated sensor data, establish a feedback loop. This loop integrates the data into the system, allowing continuous refinement of the reliability models and thresholds. This dynamic updating process ensures that the CBM strategy remains precise and relevant over time.

2.1.4 Digital Twin Integration – Step 4

The insights derived from asset ranking, reliability analysis, and CBM converge into a cohesive platform known as the DT. Traditional maintenance approaches often necessitate halting plant operations for essential upkeep, which can undermine reliability. In contrast, the DT model facilitates interventions that occur only when necessary [23]. The industry typically oversees numerous assets and conducts annual shutdowns to maintain optimal performance. However, this approach can disrupt operations and negatively impact profit margins. To overcome these challenges, the proposed model offers a unified strategy that determines maintenance interventions based on the insights generated by the DT. This model effectively integrates reliability analysis and threshold limit assessment from CBM, resulting in a more efficient and minimally disruptive maintenance process.

The DT module initiates the creation of a DT model. Depending on the specific challenges at hand, the DT model can be configured as either a physics-based model or a data-driven model. A physics-based model is appropriate when the behaviour of the asset is predominantly governed by well-established physical laws. On the other hand, a data-driven approach is more effective in scenarios marked by complexity and uncertainty. Building on the failure modes identified in Step 2, this method aims to pinpoint the key parameters in the DT that influence these failures. These parameters are tuned to ensure that the DT model simulation aligns with historical failure data in the calibration process. The guidelines for identifying key parameters when assessing asset faults are detailed in Annexes A, B, and C of the ISO 17359 standard [24]. From Step 2, updated failure probabilities for critical failure modes in reliability models are incorporated into the DT model.

Next, the proposed maintenance model calibrates the DT model using the updated reliability target. During this phase, the DT model—whether physics-based or data-driven—conducts simulations over time to generate a parameter profile, such as a temperature profile. This parameter profile is then used to fit a failure distribution, such as the Weibull distribution, and determine the associated distribution parameters. Once these parameters are established, the simulated failure probability is calculated and compared to the reliability target identified in Step 2. To facilitate this comparison, an objective function can calculate the error between the simulated probability and the target or compute the Root

Mean Square Error (RMSE) when considering multiple scenarios. This error value acts as a metric minimised through iterative adjustments of the key model parameters identified in Step 4. These parameters can be adjusted through various approaches, including manual tuning, optimisation algorithms, or Bayesian updating. After the necessary adjustments are made, the DT model reruns the simulation. This iterative process continues until the simulated failure probability closely aligns with the reliability target, minimising the error to nearly 0%. When this alignment is achieved, the calibration of the DT model is deemed successful.

Once the DT model has been fine-tuned, it establishes a real-time synchronisation connection to a live data feed. At this point, data captured from CBM sensors in Step 3 continuously streams into the DT model. This enables the virtual model to dynamically reflect the current state of the physical asset and recalibrate the failure probability, ensuring that trespassing beyond the threshold limit is swiftly detected within the DT. When an alarm is triggered due to a threshold limit, appropriate maintenance strategies can be implemented promptly. For example, when the DT model detects an unacceptable limit and signals a potential fault, the Mean Time to Repair (MTTR) must account for the duration to repair or replace a failing component once a fault is detected. This information allows for scheduling interventions in a way that minimises overall downtime.

Besides detecting threshold acceptable limits, the DT model conducts "what-if" analyses to oversee the simulated operational scenarios—such as increased load conditions, environmental changes, or simulated component degradation—to assess how these factors might impact asset performance. This allows maintenance engineers to test different strategies without interfering with the asset. Once the fine-tuning DT model is successfully trained in the small-scale environment, it is validated in the actual field.

2.1.5 Field Validation – Step 5

Field validation within the model framework entails implementing the DT model in practical, real-world settings. This model can be validated through pilot studies. These small-scale tests deploy the proposed DT model on limited assets to assess its accuracy and effectiveness before full-scale deployment. For instance, in a power plant, a pilot study might focus on applying the DT-based maintenance model to a single centrifugal pump, allowing for the monitoring of its real-time performance. This approach facilitates adjustments to the model based on insights gathered from actual operations.

The validation process commences with deploying the fine-tuned DT model alongside the physical assets in the field, enabling the model to collect real-time operational data. At this stage, discrepancies between the parameters of the DT model and the sensors' parameters installed on the physical asset are identified. Users must then investigate the source of these differences. If the discrepancies are attributed to flaws in the DT model, the process should revert to calibration. On the other hand, if the issue arises from gradual faults within the asset itself, prompt maintenance actions should be scheduled.

This ongoing feedback loop between the DT and the physical asset is vital for maintaining the model's validity and effectiveness in dynamic environments. By conducting thorough field validation, organisations can trust their DT models for informed decision-making and optimised asset management.

3 Conclusion

The rapid evolution of industrial operations and the persistent challenges of reactive maintenance have underscored the need for a more proactive approach. Traditional methods, often reliant on scheduled maintenance and manual interventions, are prone to inefficiencies, increased downtime, and higher costs. These limitations have motivated this research to explore the integration of DT technology to enhance asset reliability and optimise maintenance strategies. By conceptualising a model that

leverages up-to-date sensor data, historical reliability records, and advanced simulation techniques, this study aims to address the shortcomings inherent in conventional maintenance frameworks.

To tackle these challenges, a comprehensive, DT-based conceptual maintenance model has been developed. The proposed framework combines criticality assessment using expert input and quantitative methods, such as AHP and FMECA, with CBM, incorporating real-time sensor data and historical data for baseline calibration. Advanced signal processing techniques are applied to clean and normalise data, while the DT model is calibrated to ensure it aligns with physical assets. The model employs simulation and "what-if" analyses to predict potential operational deviations and proactively guide maintenance interventions.

This conceptual model offers several advantages. It provides a unified and systematic approach to maintenance by integrating disparate data sources and employing robust analytical methods within a single computing environment. The model is designed to improve predictive maintenance through continuous calibration and iterative refinement, potentially reducing downtime and maintenance costs.

References

[1]He Liu, Min Xia, Darren Williams, Jianzhong Sun, and Hongsheng Yan. Digital Twin-Driven Machine Condition Monitoring: A Literature Review. *Journal of Sensors*, 2022: 1–13, 2022.

[2] Adil Rasheed, Omer San, and Trond Kvamsdal. Digital Twin: Values, Challenges and Enablers From a Modeling Perspective. *IEEE Access*, 8: 21980–22012, 2020.

[3] Raymond van Dinter, Bedir Tekinerdogan, and Cagatay Catal. Predictive maintenance using digital twins: A systematic literature review. *Information and Software Technology*, 151, 2022.

[4] Terrance L. Speicher and Joanna F. DeFranco. Industry 4.0 and Digital Twins. *Computer*, 56 (6): 84–88, 2023.

[5] Xianpeng Qiao, Veronica Lestari Jauw, Lim Chin Seong, and Tiyamike Banda. Advances and limitations in machine learning approaches applied to remaining useful life predictions: a critical review. *The International Journal of Advanced Manufacturing Technology*, 133(9–10): 4059–4076, 2024.

[6] Marek Molęda, Bożena Małysiak-Mrozek, Weiping Ding, Vaidy Sunderam, and Dariusz Mrozek. From Corrective to Predictive Maintenance—A Review of Maintenance Approaches for the Power Industry. *Sensors*, 23(13): 5970, 2023.

[7] Bin Xiao, Jingshu Zhong, Xiangyu Bao, Liang Chen, Jinsong Bao, and Yu Zheng. Digital twin-driven prognostics and health management for industrial assets. *Scientific Reports*, 14(1): 13443, 2024.

[8] Himanshu Gupta and Pradeep Kundu. Digital Twin Development for Feed Drive Systems Condition Monitoring and Maintenance Planning. *PHM Society European Conference*, 8(1): 4, 2024.

[9] Guillaume Pronost, Frédérique Mayer, Mauricio Camargo, and Laurent Dupont. Digital Twins along the product lifecycle: A systematic literature review of applications in manufacturing. *Digital Twin*, 3: 3, 2023.

[10] Qiang Feng, Yue Zhang, Bo Sun, Xing Guo, Donming Fan, Yi Ren, Yanjie Song and Zili Wang. Multi-level predictive maintenance of smart manufacturing systems driven by digital twin: A matheuristics approach. *Journal of Manufacturing Systems,* 68: 443–454, 2023.

[11] Miruna-Elena Iliuţă, Mihnea-Alexandru Moisescu, Eugen Pop, Anca-Daniela Ionita, Simona-Iuliana Caramihai, and Traian-Costin Mitulescu. Digital Twin—A Review of the Evolution from Concept to Technology and Its Analytical Perspectives on Applications in Various Fields. *Applied Sciences*, 14(13): 5454, 2024.

[12] Weng Siew Lam, Weng Hoe Lam, and Pei Fun Lee. A Bibliometric Analysis of Digital Twin in

the Supply Chain. *Mathematics*, 11(15): 3350, 2023.

[13] Vivek Warke, Satish Kumar, Arunkumar Bongale, and Ketan Kotecha. Sustainable Development of Smart Manufacturing Driven by the Digital Twin Framework: A Statistical Analysis. *Sustainability*, 13(18): 10139, 2021.

[14] Lu Gao, Shugang Li, and Huimin Guo. Visualization and Analysis of Research Areas and Hot Topics in the Field of Digital Twins. *IEEE Access*, 12: 26897–26904, 2024.

[15] Michael Grieves. Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1(2014):1–7, 2014.

[16] Fei Tao, Meng Zhang, Yushan Liu, and Andrew Nee. Digital twin driven prognostics and health management for complex equipment. *CIRP Annals*, 67(1): 169–172, 2018.

[17] Azad Madni, Carla Madni, and Scott Lucero. Leveraging Digital Twin Technology in Model-Based Systems Engineering. *Systems*, 7(1): 7, 2019.

[18] Yuqian Lu, Chao Liu, Kevin I-Kai Wang, Huiyue Huang, and Xun Xu. Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and Computer-Integrated Manufacturing*, 61: 101837, 2020.

[19] Micheal Grieves and John Vickers. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. *Transdisciplinary Perspectives on Complex Systems*, Cham: Springer International Publishing, 85–113, 2017.

[20] M.A.Amiruddin Parnon, Kassandra A. Papadopoulou, and Jyoti K. Sinha. A unified framework for asset ranking based on their criticality through case studies. *International Journal of Quality & Reliability Management*, Vol. ahead-of-print No. ahead-of-print.

[21] International Electrotechnical Commission. IEC 61025: Fault tree analysis (FTA). 2006, International Electrotechnical Commission.

[22] British Standard Institute. BS EN 61078: Reliability block diagrams. 2016, British Standard Institute.

[23] Ashtarout Ammar, Francesca Maier, Rachel Catchings, Hala Nassereddine, and Gabriel Dadi. Departments of Transportation Efforts to Digitize Ancillary Transportation Asset Data: A Step Toward Digital Twins. *Transportation Research Record: Journal of the Transportation Research Board*, 2677(11): 428–445, 2023.

[24] International Organization for Standardization. ISO 17359: Condition monitoring and diagnostics of machines — General guidelines. 2018, International Organization for Standardization.

Chapter 3

Optimizing asset management through reliability data analysis

Contribution #20
Cite as:
José Sobral. Optimizing asset management through reliability data analysis. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-20,
  title ={Optimizing asset management through
     reliability data analysis},
  author={José Sobral},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Optimizing asset management through reliability data analysis

José Sobral[1,2,3,*]

[1] Mechanical Engineering Department, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Lisboa, Portugal

[2] Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

[3] Research Centre in Asset Management and System Engineering (RCM2+), Instituto Superior de Engenharia de Coimbra, Coimbra, Portugal

*Corresponding author
jose.sobral@isel.pt

Abstract

Asset management has increasingly become a crucial issue for most organizations, particularly over the past decade. Key elements of asset management include aligning asset management goals with organizational objectives, adopting a strategic approach to manage assets throughout their lifecycle, and utilizing an asset management system to improve performance continually and derive value from assets. The problem addressed and motivation are based on the optimization of asset management practices. This paper presents a method that demonstrates how leveraging data-driven insights in a reliability data analysis enhances asset performance, efficiency, and longevity. Reliability data analysis provides valuable information that can help predict asset failures, optimize maintenance schedules, and improve decision-making for more effective asset management. This work aims to offer a framework and guidelines on how to use different types of data for reliability analysis, emphasizing its significance and potential for optimizing asset management. A demonstrative example of an industrial pump from a manufacturing facility is included to illustrate this importance. Based on the results achieved in the study, its relevance is related to the knowledge obtained about the behavior of the asset regarding the occurrence of failures over time, and mainly on how such information leads to decisions that optimize the lifecycle of the referred asset.

Keywords: reliability; data analysis; asset management; optimization

1 Introduction

Asset management has become a reality for most organizations, with a special incidence in the last decade. Today, it presents a way to reduce costs, decrease risk, and achieve the best performance of the assets. The importance of asset management became more visible after the publication of PAS 55-1 [1] and PAS 55-2 [2] by the Institute of Asset Management (IAM) in 2003. Later, with the international standard ISO 55000 publication, this issue was globally understood because it addresses the need for a systematic approach to managing an organization's assets, which is crucial in today's infrastructure-intensive environment. ISO 55000 [3] provides an overview, basic principles, and terminology, which are further developed and implemented through ISO 55001 [4] and ISO 55002 [5]. ISO 55001 specifies the requirements for an asset management system, and ISO 55002 gives guidelines for applying ISO 55001.

The key elements of asset management include aligning asset management objectives with organizational objectives, a strategic approach to managing assets throughout their lifecycle, and using an asset management system to continually improve performance and realize value from assets. Good and adequate asset

management ensures that organizations can achieve their strategic goals through the effective and efficient management of their assets. It gives organizations the necessary tools to optimize asset usage, enhance financial performance, manage risk, and ensure sustainability. This provides a structured approach that helps organizations improve overall efficiency.

One of the chapters of the standard ISO 55001 [4] refers to performance evaluation that includes the reviews and inspections that the organization carries out as part of the process for monitoring and establishing the health of both the assets and the organization's support systems. For example, periodic testing of equipment, condition assessments, investigations, and reliability analysis can be used to evaluate assets' performance.

Optimizing asset management through reliability data analysis involves leveraging data-driven insights to enhance assets' performance, efficiency, and longevity. Reliability data analysis provides valuable information that can help predict asset failures, optimize maintenance schedules, and improve decision-making for managing assets effectively. Usually, it involves the treatment of failure data, its adequacy to a statistical distribution, and a prediction of the probability of failure, reliability, or failure rate, for a given time of operation.

The objective of the work is to present a framework and describe how to perform a reliability data analysis and how this analysis can be used to optimize asset management. The paper describes some benefits of using reliability data analysis to gain insights for an efficient asset management system. The document is divided into five sections. The first section refers to an introduction to asset management and points out the objectives of the work. The second section is related to reliability data analysis and how to optimize asset management based on such analysis. Section three describes the different types of data that can be used in reliability data analysis, and section four presents a demonstrative example. Finally, section five presents the main results and future work that can be done on asset management using reliability data analysis.

2 Reliability data analysis

Traditional reliability studies rely on the analysis of failure data. The question is how such studies can be used in asset performance evaluation within an asset management system.

Rafati et al. [6] stated that asset reliability analysis becomes increasingly vital for achieving energy efficiency, meeting sustainability goals, and ensuring the resilience of urban infrastructures. Traditional reliability and asset management methods sometimes fail to incorporate the new measurements and varying asset conditions. The authors presented a study that uses data-driven techniques for reliability analysis and assessment of specific urban infrastructures, and explores asset management strategies with an emphasis on data-driven reliability-based asset management.

Wu et al. [7] referred that by incorporating information about asset condition from a monitoring system, engineers can utilize asset management models to manage maintenance activities, giving the example of wind turbine blades. It can lower operating and maintenance costs and increase the life of the blades. The asset management model presented relies on the monitoring system as a source of information. The proposed model can be used to support asset management decisions when the monitoring system performance degrades.

Hartini et al. [8] referred to a study that focuses on the structure, system, and components of a multipurpose reactor, emphasizing the integration of reliability improvement within asset management optimization. The study aimed to derive Maintenance Priority Index (MPI) values, contributing to system reliability ratings, crucial for assessing component functionality and estimating Remaining Useful Life (RUL), improving asset management, and offering practical insights for critical safety system component maintenance planning.

Another study developed by Mirshekali et al. [9] stated that aging, environmental factors, and operational conditions contribute to a higher likelihood of faults, which reduces reliability. The work was assessed using a real Danish distribution grid, and the findings indicate that the proposed method demonstrates superior performance in terms of reliability metrics and power outage cost.

It is common sense that reliability and availability analysis is an appropriate technique to support

proper physical asset management. However, this task can become very complex in the face of a dynamic demand during the operational course of the analyzed assets. In some cases, advanced tools for reliability and availability analysis, such as Petri nets, are expected to be applied. Melani et al. [10] presented a Petri net-based reliability and availability analysis in a dynamic demand scenario to support maintenance decision-making for asset management.

Based on the literature review and case studies presented, it has been proven that organizations can reduce downtime, improve operational efficiency, extend asset lifespan, and lower maintenance costs by optimizing asset management through reliability data analysis. The key is to collect and analyze relevant data, implement predictive maintenance, and continuously improve processes based on data insights. By focusing on the most critical assets and addressing failure risks proactively, businesses can achieve more reliable operations and maximize the value derived from their assets.

Follows a framework representation (Figure 1) for the whole process related to data processing and decision-making for critical assets.

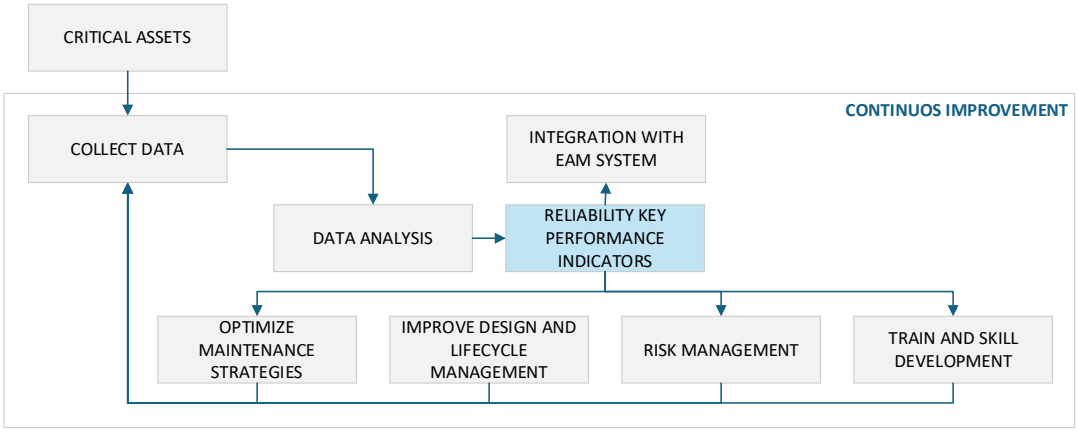


Figure 1 – Data processing, reliability analysis, and decisions

After identifying all critical assets in the plant, the first step is to collect data from those assets. These data can be from asset performance data (historical performance data such as failure rates, maintenance logs, repair records, and downtime), from condition monitoring (data from sensors or predictive technologies to monitor real-time asset health, such as temperature, pressure, and vibration), from inspections and tests (data related to additional data on asset condition), and even from environmental data that could influence asset reliability (information about working environment, such as humidity, temperature, and dust).

After that, a data analysis must be done, using predictive analytics and modelling. It is common to use reliability models (e.g., Weibull distribution, failure rate analysis) to identify common failure modes and predict when an asset is likely to fail. The results, assumed as reliability key performance indicators, such as the Mean Time Between Failures (MTBF), that measures the expected time to failure or time between failures, the Mean Time to Repair (MTTR), that measures how long it takes to restore an asset to operational status after a failure, and the availability and reliability, that analyze asset availability (uptime) and reliability (ability to perform its required function without failure) to ensure assets are functioning efficiently. The Remaining Useful Life (RUL) can be predicted by implementing machine learning or statistical models to predict the remaining useful life of assets, allowing proactive interventions before failures occur.

These reliability indicators must be integrated into the Enterprise Asset Management (EAM) system or Computerized Maintenance Management System (CMMS), providing valuable information. It should correspond to real-time data integration to ensure seamless communication and data-driven decision-making. For this reason, it is common to use dashboards to visualize key reliability metrics, asset health

status, and performance trends, providing a clear picture for decision-makers.

This reliability data analysis can be used to optimize maintenance strategies, establishing Predictive Maintenance (PdM) when historical reliability data is used to predict when equipment will fail or when performance will degrade, so that maintenance can be scheduled accordingly. This reduces unnecessary downtime and extends asset life. It can also be redirected to categorize assets based on their criticality and failure modes, and tailor maintenance strategies accordingly (e.g., run-to-failure, time-based, or predictive maintenance), meaning a Reliability-Centered Maintenance (RCM) strategy. The results of reliability data analysis can also be used to investigate the root causes of asset failures, and based on this Root Cause Analysis (RCA), modify maintenance strategies and processes, reducing recurrence. It is also possible to use predictive analytics to shift from scheduled maintenance to condition-based maintenance (CBM), where maintenance is done when the asset's performance drops below a predefined threshold. A Risk-Based Maintenance (RBM) strategy can also be derived from the results of reliability data, used to assess the risk associated with asset failure and prioritize maintenance activities based on potential consequences and failure probabilities. With predictive models, one can proactively manage risk by addressing potential issues before they lead to catastrophic failures.

The reliability data analysis can also be applied to improve asset design using failure data to inform the design and improvement of new assets, ensuring that reliability is considered during the design phase to reduce future failures. This is called Design for Reliability (DfR). The same reliability data analysis serves to inform decisions on asset replacement and upgrades. Assets with high failure rates or maintenance costs may be due for replacement sooner than initially planned.

Last, but not least, the results of a reliability data analysis can lead to the conclusion of the necessity of training and skill development to ensure that the asset management team is trained in interpreting reliability data and analytics results. This helps in making informed decisions and implementing improvements. It will also foster collaboration between engineers, operations, and maintenance teams, allowing insights from reliability data to be shared across departments for better decision-making. Finally, all this must be faced in a continuous improvement model, with feedback loops that continuously collect and analyze reliability data to improve future maintenance strategies and optimize asset management. Analyzing post-maintenance performance helps refine the process. This is a data-driven decision-making model, fostering a culture of using data to drive asset management decisions, leading to more informed strategies for optimization.

3 Types of data

To optimize asset management through a reliability data analysis, a wide range of data types can be used to improve asset performance, predict failures, and enhance maintenance strategies. Wang et al. [11] referred that Bernoulli data (pass/fail), lifetime data, and degradation data are commonly encountered in product reliability assessment. These data are often collected from different sources (such as field use, accelerated tests, history, …), and it is desirable to utilize these heterogeneous data within one computational framework to provide a comprehensive evaluation of product reliability. Thus, it is common to consider several data sources when trying to develop a reliability analysis. Follows a breakdown of the key types of data that are valuable in asset management.

3.1 Historical performance data

Historical performance data relies on failure records, referring to information on past failures, including when and how assets failed, distinct failure modes, and their causes. This helps identify patterns and predict future failures. Maintenance history is a valuable source of information because it contains records of all maintenance activities performed on assets, including preventive maintenance (PM), corrective maintenance (CM), and any repairs or replacements. Downtime data is very important information, referring to how long assets were out of service during maintenance or failure events,

helping to understand downtime trends and optimize schedules. Data related to work orders helps track completed and upcoming maintenance tasks, as well as the status and resources required for maintenance. Data on the resources (e.g., parts, labor, tools) used during maintenance activities helps optimize inventory and resource allocation. Financial data on the costs to maintain each asset helps with cost-benefit analysis for repairs versus replacement. Supply chain and spare parts data, related to i) spare parts inventory (data on the availability, usage rates, and lead times for spare parts) can help ensure that the right parts are available for maintenance when needed, ii) lead times for procurement (tracking how long it takes to acquire critical replacement parts) helps manage downtime during repairs and improve maintenance planning, and iii) supplier performance data (data on supplier reliability and delivery times) ensures that the supply chain can meet the demands of asset maintenance.

3.2 Operational data

The operational data also plays a very important role in reliability analysis. This data refers to operating hours, meaning the total time the asset has been operating, often used to track wear and tear or calculate maintenance intervals. Load conditions must also be considered because they contain data on the operational load or capacity at which the asset works, as higher loads can lead to faster wear or premature failures. The operational data includes cycle data, meaning the number of operating cycles or processes an asset has completed, especially for assets like pumps, motors, or engines that may wear with each cycle. Also, data about operational speed or efficiency of an asset must be considered, which can indicate suboptimal performance or impending failure. Operational data may also include asset lifecycle data as i) age of asset (indicating the likelihood of failure based on historical trends of similar assets), ii) remaining useful life (RUL), iii) total cost of ownership (TCO) (referred to financial data that tracks the total cost associated with maintaining and operating an asset, which is crucial for assessing its long-term value and determining the best time to replace it), and iv) warranty information (helping track the warranty status of assets, which can influence decisions regarding repairs or replacement).

3.3 Condition monitoring data

Data from condition monitoring systems point out some parameters that reflect the asset's health. This data can be about vibration levels, which are critical for detecting mechanical faults in rotating equipment such as motors, pumps, and fans, monitoring temperature trends of assets (e.g., motors, bearings, engines), and helping to detect overheating, which is a common sign of failure. Also, pressure data for assets like compressors, turbines, and hydraulic systems, where pressure measurements can indicate potential problems like leaks or blockages. Another example is the fluid quality data, for assets that use lubricants or hydraulic fluids, measurements such as viscosity, contamination levels, or acidity can indicate issues with fluid quality and, by extension, asset condition. Acoustic data, related to sound emissions, such as unusual noises or patterns, can also be used to identify wear or damage in machinery. It is also possible to have advanced sensors (temperature, pressure, humidity, etc.) connected to the Internet of Things (IoT), providing real-time data on asset health and performance. This data can be used to feed predictive models (e.g., machine learning or statistical analysis), predicting potential asset failures based on patterns identified from historical data and real-time data.

3.4 Environmental data

Environmental data is sometimes a very important issue to be considered in a reliability analysis framework. It refers to ambient temperature and humidity, as external conditions can affect asset performance and reliability, particularly in industries like manufacturing or energy generation, where equipment is sensitive to environmental changes. Data on the presence of dust, chemicals, or other contaminants in the environment that may impact the asset's condition, especially for high-precision or sensitive equipment, must also be considered. For assets exposed to outdoor or harsh environments, data

from corrosion levels can help predict the remaining life or the need for protective measures.

3.5 Design and configuration data

The model must consider original design asset specifications, including expected service life, tolerances, and performance characteristics, to help provide a baseline for evaluating asset performance. All modifications, upgrades, or retrofits, referring to changes made to the asset, should be tracked as they can influence asset reliability and performance. It is also common to include manufacturer data on recommended maintenance practices, operational limits, and expected performance standards, which can help inform maintenance schedules and operational decisions.

3.6 Other analysis data

Data from other types of analysis can also be incorporated into the model. It can be related to failure modes, identifying and documenting all potential failure modes of assets (e.g., wear, corrosion, fatigue, electrical failures), helping to predict which parts of the asset are most likely to fail and when. It can also be related to a criticality and risk assessment, ranking assets based on their criticality and impact on operations, helping to prioritize which assets need the most attention, and reducing the risk of catastrophic failures. Also, information gathered from root cause analysis (RCA) can help to determine underlying causes of failures, so corrective actions can be applied to prevent recurrence. Figure 2 graphically shows all the referred data types that can be considered in the model of a reliability analysis.

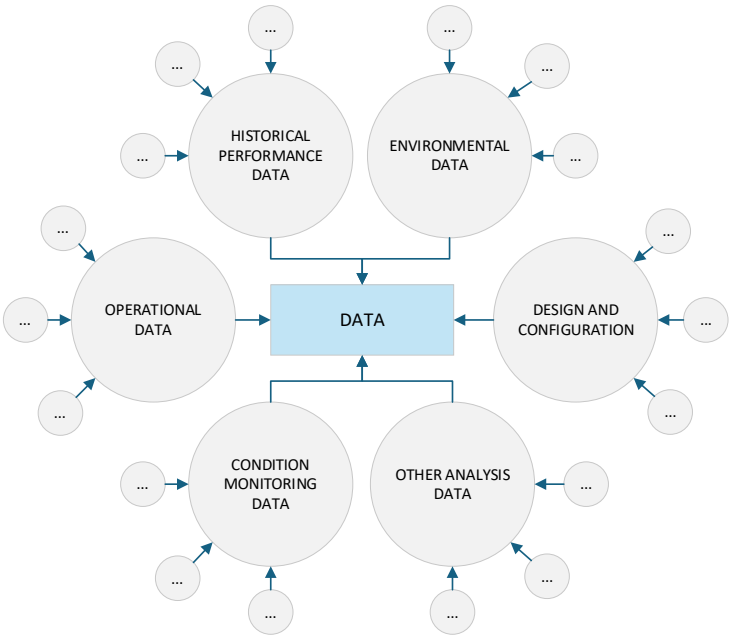


Figure 2 – Data for reliability analysis

By gathering and analyzing a combination of these data types, organizations can gain a comprehensive understanding of their assets' reliability, performance, and failure patterns. This enables more proactive, data-driven decision-making to extend asset life, reduce downtime, and optimize maintenance strategies. Leveraging the right mix of data is key to unlocking the full potential of asset management.

4 Demonstrative example

This section presents a detailed example of asset failure data, inherent reliability analysis, and the impact of the results on asset management decisions. In the present work, it will be used as a demonstrative example, an industrial pump from a manufacturing facility's fluid handling system.

4.1 Asset description

The asset is a common centrifugal pump, whose function is to transfer liquid (chemicals) from one location to another. This pump is vital to the continuous operation of the manufacturing facility. Failures would lead to significant downtime, halting production.

4.2 Failure data

The facility's maintenance team has been tracking the pump failures over the last three years. The data gathered includes the failure frequency, failure modes, failure causes, and downtime. Table 1 shows the data records.

Table 1 – Data for reliability analysis

Failure mode	**Qt.**	**Causes**	**Downtime [h]**
Mechanical seal failure	4	Resulted from misalignment during installation and poor lubrication	72
Bearing failure	2	Caused by overheating due to a lack of regular maintenance (e.g., lack of lubrication)	
Motor overload	1	Triggered by operating the pump at a higher-than-recommended pressure, causing excessive motor load	
Cavitation	1	Due to low NPSH (Net Positive Suction Head) caused by fluctuating suction pressure	

4.3 Simplified reliability analysis

The reliability analysis presented refers to a simplified manner to produce outputs for the subsequent decision-making process. It was used to understand how frequent failures occur, their causes, and their impact. This allows for informed decisions regarding maintenance, replacement, and reliability improvements.

For the gathered data presented in Table 1, a first reliability indicator can be simply defined, that is, the Mean Time Between Failures (MTBF). Taking into account a total operating time of 15000 hours (working 5,000 hours/year) over the past three years and a total of 8 failures, the MTBF was mathematically defined as 1875 hours, meaning that, on average, the pump fails every 1875 hours of operation. If we are dealing with an asset in its useful life, it corresponds to a failure rate (λ) of 0.000533 failures per hour, which can be used to predict failures over the next 1000 operating hours, for instance.

A more detailed analysis shows that mechanical seal failures and bearing failures have the most significant impact on operations due to the downtime associated with their repairs. A mechanical seal failure can cause leaks, leading to fluid loss, and potentially a hazardous situation because it is a chemical pump. It also results in increased maintenance costs and labor hours for replacing seals. A bearing failure results in complete pump stoppage and typically requires more time to disassemble and inspect the pump components for further damage. The cavitation failure mode, although less frequent, can cause significant damage to the pump impeller and other internal components, requiring more expensive repairs or even

replacement.

Looking at the data related to the time between failures, and applying a statistical approach to the data (applying statistical tests like the Kolmogorov-Smirnov test, a normalized correlation coefficient or the likelihood value) [12], it could be seen that the most suitable reliability distribution (between the most common distributions, as Weibull, Normal, Lognormal, Exponential) to model pump failure characteristics was the Weibull distribution (biparametric), with a shape parameter (β) of 1.65 (suggesting a wear-out failure mode) meaning that the likelihood of failure increases as the pump ages, and a scale parameter (η) of 1960 hours. The correlation coefficient 0.9967 is considered a relatively good value for that parameter. Figure 3 shows the graphical representation of the Weibull probability plotting paper (failures), the reliability function, the failure rate function (increasing), and the probability density function.

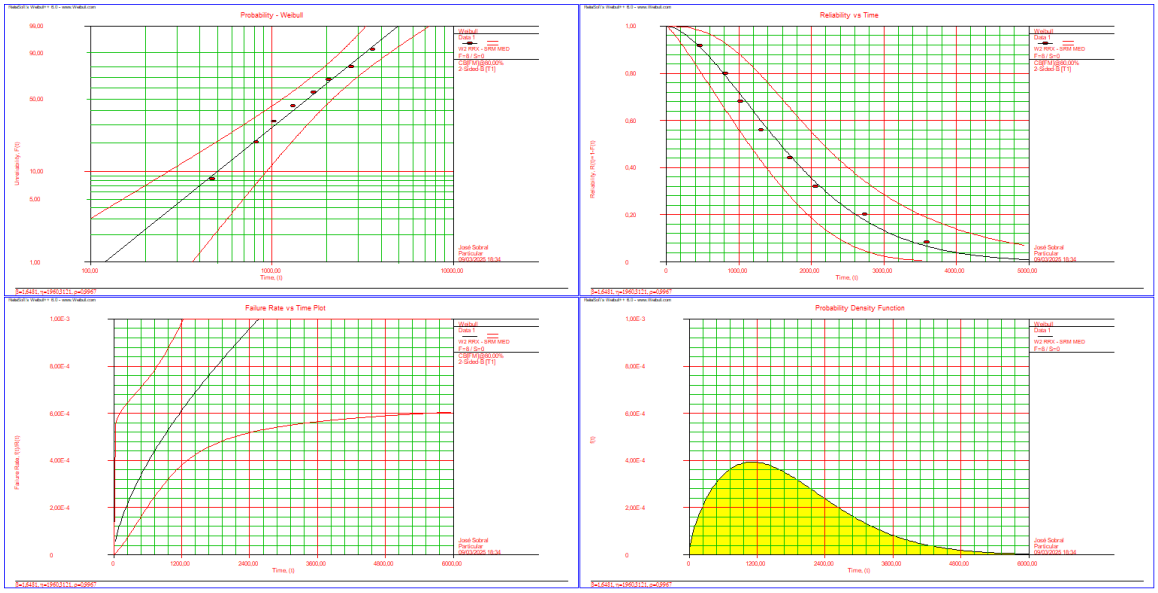


Figure 3 – Graphical representation of reliability functions

The results achieved in the analysis can be used as useful information to increase the knowledge about the behavior of the asset regarding the occurrence of failures over time, and such information can lead to decisions that optimize the lifecycle of the asset under study.

4.4 Impact of reliability results on asset management decisions

As the objective is to optimize asset management through reliability data analysis, based on the simplified reliability analysis developed in the example, the following decisions can be made:

- **Improve preventive maintenance program** – Regarding the failure modes identified as critical (mechanical seal and bearing failures) are due to poor maintenance practices, such as misalignment and insufficient lubrication, the decision was to revise the preventive maintenance (PM) schedule to include more frequent checks on mechanical seals and bearings, lubrication, and alignment procedures. It was then established to develop regular lubrication intervals, the addition of seal and bearing inspections every 500 operating hours, and the implementation of a vibration analysis program to detect early signs of bearing degradation;

- **Implement predictive maintenance (PdM) strategies** – Regarding that the motor overload failures and cavitation are both related to operational conditions (pressure levels and suction conditions), it was decided to use sensors to monitor the pump's operational parameters such as pressure, flow, and motor load, and integrate this data into a predictive maintenance system. This leads to the installation of pressure and flow sensors to monitor NPSH and prevent cavitation, and the use of motor load sensors to track overload conditions and trigger alarms before a failure occurs;

- **Spare parts management** – As bearing and seal failures occur more frequently, leading to stock-outs and longer repair times, it was established to increase spare parts inventory for critical components to ensure rapid replacement and reduce downtime. Thus, it was decided to increase the stock of critical spare parts (mechanical seals, bearings), and consider using a just-in-case inventory strategy for high-use items that need to be readily available during unscheduled maintenance;

- **Invest in asset replacement or overhaul** – The relatively high failure rate of the pump (being the common occurrence of mechanical and bearing failures), with a short MTBF, suggests that it is reaching the end of its useful life, and frequent failures might continue or increase. Based on this, it was decided to evaluate the possibility of replacing the pump with a more reliable model or overhauling the current pump. Thus, a cost-benefit analysis must be done, comparing the cost of continued maintenance vs. replacing the pump and, if maintenance costs are significantly higher than replacement, initiating the procurement of a new, more reliable pump model with an improved MTBF;

- **Risk mitigation strategy** - Some failure modes (like cavitation and bearing failure) have higher consequences, including production stoppage and safety risks. Regarding this evidence, a focus on a risk-based approach to prioritize interventions on the most critical failure modes must be accomplished. It was then decided to develop a Risk-Based Maintenance (RBM) strategy to focus efforts on the most critical failure modes (bearing failures and cavitation) and perform a Root Cause Analysis (RCA) to identify how cavitation can be better controlled by adjusting suction pressure or implementing better pressure regulation.

All these decisions can have a long-term impact on asset management as a cost reduction, by proactively maintaining the pump, scheduling replacement of worn-out components (seals and bearings), and introducing predictive monitoring, reducing unplanned downtime and maintenance costs. It can also have an impact on asset lifespan because a proper alignment, lubrication, and operational adjustments (e.g., avoiding cavitation) will extend the pump's useful life. The actions will also increase operational efficiency because by maintaining operational conditions within optimal parameters, the pump will operate more efficiently, reducing energy consumption and improving throughput. In the end, all these decisions will undoubtedly improve reliability metrics over time. For example, MTBF will improve as a result of better maintenance practices and component monitoring.

5 Conclusions and future work

The present work deals with reliability analysis and how such analysis can be an important tool to make decisions and optimize asset management. Several types of data that can be collected from the perspective of reliability studies were described, and a demonstrative example was presented. In this demonstrative example, decisions were made, and actions were established, in a way to optimize many aspects of the management of an asset.

It can be concluded that by analyzing asset failure data (such as failure modes, failure causes, and downtime), performing inherent reliability analysis, and implementing the insights gained into asset management decisions, organizations can optimize their maintenance strategy, reduce unplanned downtime, and extend the life of critical assets. These improvements lead to more reliable operations, reduced costs,

and better risk management, ultimately enhancing overall asset performance.

Some limitations can be addressed in this type of study, such as the lack of data, poor data, or even the fear of the mathematical part of the analysis by the professionals.

As an actual tendency, the use of incorporated sensors to collect relevant data and the development of machine learning algorithms to predict failure events can lead to setting up a system for real-time monitoring and predictive maintenance, reducing resources and spending less time. Combining effective data collection, machine learning models, and real-time alerts, organizations can enhance asset management, reduce downtime, and optimize maintenance strategies.

References

[1] BSI, PAS 55-1 - Asset Management Part 1: Specification for the optimized management of physical assets, British Standards Institution, London, United Kingdom, 2008

[2] BSI, PAS 55-2 - Asset Management Part 2: Guidelines for the application of PAS 55-1, British Standards Institution, London, United Kingdom, 2008

[3] ISO, ISO 55000 - Asset management — Vocabulary, overview and principles, International Organization for Standardization, Geneva, Switzerland, 2024

[4] ISO, ISO 55001 - Asset management - Asset management system - Requirements, International Organization for Standardization, Geneva, Switzerland, 2024

[5] ISO, ISO 55002 - Asset management - Asset management system - Requirements, International Organization for Standardization, Geneva, Switzerland, 2018

[6] Amir Rafati, Maryamsadat Tahavori and Hamid Reza Shaker. Data-driven reliability analysis of district heating systems for asset management applications: A review. *Sustainable Cities and Society*, Volume 118, 106052, ISSN 2210-6707, https://doi.org/10.1016/j.scs.2024.106052, 2025

[7] Wen Wu, Darren Prescott, Rasa Remenyte-Prescott, Ali Saleh and Manuel Chiachio Ruano. An asset management modelling framework for wind turbine blades considering monitoring system reliability. *Reliability Engineering & System Safety*, Volume 252, 110478, ISSN 0951-8320, https://doi.org/10.1016/j.ress.2024.110478, 2024

[8] Entin Hartini, Sigit Santoso, Deswandri Deswandri, Sriyono, Veronica Indriati Sri Wardhani, Endiah Puji Hastuti, Djati Hoesen Salimy, Damianus Toersiwi Sony Tjahyani, Ignatius Djoko Irianto, Sanda and Farisy Yogatama Sulistyo. Optimizing asset management for Structure System Components of RSG-GAS: A reliability-centric approach. *Nuclear Engineering and Technology*, Volume 56, Issue 11, Pages 4905-4913, ISSN 1738-5733, https://doi.org/10.1016/j.net.2024.06.055, 2024

[9] Hamid Mirshekali, Lasse Kappel Mortensen and Hamid Reza Shaker. Reliability-aware multi-objective approach for predictive asset management: A Danish distribution grid case study. *Applied Energy*, Volume 358, 122556, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2023.122556, 2024

[10] Arthur Henrique de Andrade Melani, Miguel Angelo de Carvalho Michalski, Renan Favarão da Silva, Adherbal Caminada Netto and Gilberto Franciso Martha de Souza. Petri net-based reliability and availability analysis to support asset management: a CODOG propulsion system case study. *IFAC-PapersOnLine*, Volume 55, Issue 19, Pages 19-24, ISSN 2405-8963, https://doi.org/10.1016/j.ifacol.2022.09.178, 2022

[11] Lizhi Wang, Rong Pan, Xiaohong Wang, Wenhui Fan and Jinquan Xuan. A Bayesian reliability evaluation method with different types of data from multiple sources. *Reliability Engineering & System Safety*, Volume 167, Pages 128-135, ISSN 0951-8320, https://doi.org/10.1016/j.ress.2017.05.039, 2017

[12] Dimitri Kececioglu. Reliability and Life Testing Handbook - Vol. 1. DEStech Publications, Inc., Lancaster, USA, 1993

Chapter 4

Medical Equipment Classification and Normative Methodologies: A Complete Approach to Standards and Metrology

Contribution #21
Cite as:
Inês Sousa, José Torres Farinha, Hugo Raposo. Medical Equipment Classification and Normative Methodologies: A Complete Approach to Standards and Metrology. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-21,
  title ={Medical Equipment Classification and
     Normative Methodologies: A Complete Approach to
     Standards and Metrology},
  author={Inês Sousa and  José Torres Farinha and Hugo
     Raposo},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Medical Equipment Classification and Normative Methodologies: A Complete Approach to Standards and Metrology

Sousa, Inês[1,2*], Farinha, José Torres [1], Raposo, Hugo [1]

[1] Instituto Superior de Engenharia de Coimbra, Polytechnic University of Coimbra, RCM[2+] Research Centre in Asset Management and System Engineering, 3030-199 Coimbra, Portugal.
[2] CISE - Electromechatronic Systems Research Centre, University of Beira Interior, Calçada Fonte do Lameiro, 6201-001 Covilhã, Portugal.

ines.guedes.sousa@ubi.pt; tfarinha@isec.pt; hugo.raposo@isec.pt

Abstract

The classification of medical equipment and the implementation of normative methodologies are crucial to ensuring the safety, efficacy and accuracy of devices used in the healthcare context. This study addresses the complexity involved in the compliance of medical equipment with rigorous international standards, as well as the application of appropriate metrological practices for their evaluation. The motivation for this research resides in the necessity to improve certification processes, ensuring that devices meet the highest quality and performance criteria, to guarantee the protection of both patients and healthcare professionals. The methods adopted include a thorough analysis of the technical standards in force, such as ISO 13485 and IEC 60601, together with an investigation of the metrological practices applied to the calibration and verification of equipment performance. The main results show that harmonization between normative and metrological methodologies promotes substantial improvement in certification processes and contributes to continuous innovation in the area. The relevance of this work is indisputable, as it broadens the understanding of the state of the art, providing robust and applicable guidelines for the classification and validation of medical equipment, with a direct impact on safety and quality in the health sector.

Keywords: Medical Equipment, Physical Assets, Maintenance Standards, Asset Management Standards

1 Introduction

The growing technological evolution in the medical equipment sector poses significant challenges in terms of safety, efficacy and regulatory compliance. The classification and certification of these devices are essential to ensure that they meet the highest international standards, protecting patients and healthcare professionals [1].

This study aims to address the importance of the normative and metrological methodologies applied in the classification of medical equipment, with the aim of providing a comprehensive understanding of the normative and calibration processes used to ensure the quality and reliability of these devices [2].

The motivation for this work lies in the need for a more in-depth understanding of the regulatory and metrological practices governing medical equipment, promoting advances in the development and certification of these devices [3].

The literature on the classification of medical equipment and regulatory standards is vast and complex. Several studies analyze the international technical standards, such as ISO 13485 and IEC 60601, which govern the manufacture and performance of medical devices [4].

The application of metrology to the calibration and verification of devices is also widely discussed,

with an emphasis on the accuracy of measurements and ensuring that devices meet the requirements demanded by regulatory bodies [5]. However, there are significant gaps regarding the efficient integration of normative methodologies with metrological practices, especially in contexts of accelerated technological innovation [6].

To address the proposed problem, a qualitative research methodology was adopted, combining a literature review and document analysis. The review included international standards such as ISO 13485 and IEC 60601, as well as academic articles and technical reports that discuss metrology practices applied to medical equipment. In addition, a case analysis of specific medical devices was carried out, focusing on the application of normative and calibration methodologies in certification processes [7] [8].

2 Classification of Medical Equipment

According to the World Health Organization (WHO), medical equipment (or medical device) is defined as:

"Any instrument, apparatus, implement, machine, implantable device, reagent for in vitro use, software, material or other similar or related article intended by the manufacturer to be used, alone or in combination, in human beings for the purpose of:

- diagnosis, prevention, monitoring, treatment or relief of a disease;
- diagnosis, monitoring, treatment, relief or compensation of an injury;
- research, replacement, modification or support of anatomy or a physiological process;
- support or maintenance of life;
- control of conception;
- disinfection of medical devices;
- providing information by in vitro means on samples from the human body [3]."

Figure 1 shows some examples of medical devices according to their respective purpose.



Figure 1: Medical device examples according to their respective purpose [1]

According to the World Health Organization (WHO), medical equipment is classified primarily on the basis of the level of risk it poses to the patient and/or user. This approach is similar to that adopted by

other international regulatory bodies, such as the European Union and the US FDA, promoting a global harmonization of the regulation of medical devices.
The WHO divides medical equipment into four main classes, according to the risk associated with their use [9]:

Class I: Low risk (e.g. gauze, wheelchairs, stethoscope)
Class II-a: Moderate risk (e.g. syringes, X-ray equipment)
Class II-b: High moderate risk (e.g. monitors, ventilators)
Class III: High risk (e.g. defibrillators, implants)

Table 1 show how different countries use different classifications:

Table 1: Medical Device Classification in Different Countries

Countries				
Europe	Class I	Class II-a Class II-b	Class III	
USA	Class I	Class II	Class III	
Canada	Class I	Class II	Class III	Class IV
Australia	Class I	Class II-a Class II-b	Class III	
India	Class A	Class B	Class C	Class D

The current device-related regulations to Regulation (EU) 2017/745 on medical devices (MDR) and Regulation (EU) 2017/746 on in vitro diagnostic medical devices (IVDR, in force since May 26, 2022), recognizes medical device nomenclature and define medical devices as:

- "non-invasive" devices;
- "active devices" utilizing the source of energy for their use, software is also included in this category;
- "implantable devices" that intend to enhance the function of certain body parts or to replace damaged tissue;
- "invasive devices" the function of which lies inside the human body.

The performance of the medical device as intended by its manufacturer requires the classification of the medical device and other evaluation steps. In addition, Table 2 show that there are classification rules that must be considered when evaluating the medical device [10]:

Table 2: Classification rules for medical equipment

NON-IN-VASIVE DEVICES	This equipment is classified as class I, unless: • non-invasive devices intended to channel or store blood, body fluids, cells or tissues, liquids or gases for the purpose of infusion, administration or eventual introduction into the body are classified in class IIa: • if they can be connected to an active device of class IIa, class IIb or class III; • if they are intended to be used to channel or store blood or other body fluids or to store organs, parts of organs or body cells and tissues of organs or body cells and tissues, except for blood bags; blood bags are classified in Class IIb; All non-invasive devices intended to modify the biological or chemical composition of human tissues or cells, blood, other body fluids or other fluids intended to be implanted or administered in the body are classified as class IIb, unless the treatment for which the device is used consists of filtration, centrifugation or gas or heat exchange, in which case they are gas and heat exchanges, in which case they are classified in class IIa; All Non-invasive devices consisting of a substance or mixture of substances intended to be used in vitro in direct contact with human cells, tissues or organs taken from the human body or used in vitro with human embryos prior to their implantation or administration in the human body are classified in class III;

	All non-invasive devices that come into contact with injured skin or mucous membranes are classified as: class I, if they are intended to be used as a mechanical barrier, for compression or for exudate absorption; class IIb, if they are intended to be used mainly on skin lesions that have broken through the dermis or mucosa and can only heal by second intention; class IIa, if they are intended primarily to manage the microenvironment of injured skin or mucous membranes mucous membrane; class IIa in all other cases; This rule also applies to invasive devices that come into contact with the injured mucosa.
INVASIVE DEVICES	All body orifice invasive devices, with the exception of invasive devices of a surgical nature, which are not intended for connection to an active device or which are intended to be connected to a class I active device are classified as: • class I, if they are intended for transient use; • class IIa, if they are intended for short-term use, except if they are used in the oral cavity up to the pharynx in an ear canal up to the eardrum or in the nasal cavity, in which case they are classified as class I; • class IIb, if they are intended for long-term use, except if they are used in the oral cavity up to the pharynx, mucous membrane, in which case they are classified as class IIa. All body orifice invasive devices, except for invasive devices of a surgical nature, intended to be connected to an active device of class IIa, class IIb or class III, are classified in class IIa; All surgically invasive devices intended for transient use are classified in class IIa, except: • are specifically intended to control, diagnose, monitor or correct a defect of the heart or central circulatory system through direct contact with those parts of the body, in which case they are classified as class III; • are reusable surgical instruments, in which case they are classified as class I; • are specifically intended to be used in direct contact with the heart or central circulatory system or the central nervous system, in which case they are classified in class III; • are intended to deliver energy in the form of ionizing radiation, in which case they are classified in class IIb; • have a biological effect or are totally or mainly absorbed, in which case they are classified in class IIb; • are intended to administer medicinal products via a delivery system, if such administration of a medicinal product is carried out in a manner which is potentially hazardous having regard to the mode of application, in which case they are classified in class IIb. All surgically invasive devices intended for short-term use are classified in class IIa, unless • they are specifically intended to control, diagnose, monitor or correct a defect of the heart or central circulatory system through direct contact with those parts of the body, in which case they are classified as class III; • they are specifically intended to be used in direct contact with the heart or central circulatory system or the central nervous system, in which case they are classified as class III; • if they are intended to deliver energy in the form of ionizing radiation, they are classified in class IIb; • they have a biological effect or are totally or mainly absorbed, they are classified in class III; • if they are intended to undergo chemical changes in the body, they are classified in class IIb, except if the devices are placed in the teeth; • if they are intended to administer medicines, they are classified in class IIb. All implantable devices and long-term invasive devices of a surgical nature are classified in class IIb, unless: • they are intended to be placed in teeth, they are classified as class IIa; • if they are intended to be used in direct contact with the heart, central circulatory system or central nervous system, they are classified in class III; • they have a biological effect or are totally or mainly absorbed, they are classified in class

	III; • they are intended to undergo chemical changes in the body, they are classified in class III, except if the devices are placed in the teeth; • they are intended to administer drugs, they are classified in class III; • they are active implantable devices or their accessories, are classified in class III • they are breast implants or surgical meshes, are classified in class III; • they are total or partial joint replacements, are classified in class III, except for auxiliary components such as screws, wedges, plates and instruments; • they are vertebral disc replacement implants or implantable devices that come into contact with the spine, are classified as class III, except for components such as screws, wedges, plates and instruments.
ACTIVE DEVICES	All active therapeutic devices intended to administer or exchange energy are classified as class IIa unless their characteristics are such that they may administer energy to or exchange energy with the human body in a potentially hazardous way, taking account of the nature, the density and site of application of the energy, in which case they are classified as class IIb. All active devices intended to control or monitor the performance of active therapeutic class IIb devices, or intended directly to influence the performance of such devices are classified as class IIb. All active devices intended to emit ionizing radiation for therapeutic purposes, including devices which control or monitor such devices, or which directly influence their performance, are classified as class IIb; All active devices that are intended for controlling, monitoring or directly influencing the performance of active implantable devices are classified as class III. Active devices intended for diagnosis and monitoring are classified as class IIa: • if they are intended to supply energy which will be absorbed by the human body, except for devices intended to illuminate the patient's body, in the visible spectrum, in which case they are classified as class I; • if they are intended to image in vivo distribution of radiopharmaceuticals; • if they are intended to allow direct diagnosis or monitoring of vital physiological processes, unless they are specifically intended for monitoring of vital physiological parameters and the nature of variations of those parameters is such that it could result in immediate danger to the patient, for instance variations in cardiac performance, respiration, activity of the central nervous system, or they are intended for diagnosis in clinical situations where the patient is in immediate danger, in which cases they are classified as class IIb. Active devices intended to emit ionizing radiation and intended for diagnostic or therapeutic radiology, including interventional radiology devices and devices which control or monitor such devices, or which directly influence their performance, are classified as class IIb. Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause: • death or an irreversible deterioration of a person's state of health, in which case it is in class III; • a serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb. Software intended to monitor physiological processes is classified as class IIa, except if it is intended for monitoring of vital physiological parameters, where the nature of variations of those parameters is such that it could result in immediate danger to the patient, in which case it is classified as class IIb. All other software is classified as class I. All active devices intended to administer and/or remove medicinal products, body liquids or other substances to or from the body are classified as class IIa, unless this is done in a manner that is potentially hazardous, taking account of the nature of the substances involved, of the part of the

	body concerned and of the mode of application in which case they are classified as class IIb. All other active devices are classified as class I.
SPECIAL RULES	All devices incorporating, as an integral part, a substance which, if used separately, can be considered to be a medicinal product, as defined in point 2 of Article 1 of Directive 2001/83/EC, including a medicinal product derived from human blood or human plasma, as defined in point 10 of Article 1 of that Directive, and that has an action ancillary to that of the devices, are classified as class III. All devices used for contraception or prevention of the transmission of sexually transmitted diseases are classified as class IIb, unless they are implantable or long term invasive devices, in which case they are classified as class III. All devices intended specifically to be used for disinfecting, cleaning, rinsing or, where appropriate, hydrating contact lenses are classified as class IIb. All devices intended specifically to be used for disinfecting or sterilising medical devices are classified as class IIa, unless they are disinfecting solutions or washer-disinfectors intended specifically to be used for disinfecting invasive devices, as the end point of processing, in which case they are classified as class IIb. This rule does not apply to devices that are intended to clean devices other than contact lenses by means of physical action only. Devices specifically intended for recording of diagnostic images generated by X-ray radiation are classified as class IIa. All devices manufactured utilising tissues or cells of human or animal origin, or their derivatives, which are non-viable or rendered non-viable, are classified as class III, unless such devices are manufactured utilising tissues or cells of animal origin, or their derivatives, which are non-viable or rendered non-viable and are devices intended to come into contact with intact skin only. All devices incorporating or consisting of nanomaterial are classified as: • class III if they present a high or medium potential for internal exposure; • class IIb if they present a low potential for internal exposure; and • class IIa if they present a negligible potential for internal exposure. All invasive devices with respect to body orifices, other than surgically invasive devices, which are intended to administer medicinal products by inhalation are classified as class IIa, unless their mode of action has an essential impact on the efficacy and safety of the administered medicinal product or they are intended to treat life-threatening conditions, in which case they are classified as class IIb. Devices that are composed of substances or of combinations of substances that are intended to be introduced into the human body via a body orifice or applied to the skin and that are absorbed by or locally dispersed in the human body are classified as: • class III if they, or their products of metabolism, are systemically absorbed by the human body in order to achieve the intended purpose; • class III if they achieve their intended purpose in the stomach or lower gastrointestinal tract and they, or their products of metabolism, are systemically absorbed by the human body; • class IIa if they are applied to the skin or if they are applied in the nasal or oral cavity as far as the pharynx, and achieve their intended purpose on those cavities; • class IIb in all other cases. Active therapeutic devices with an integrated or incorporated diagnostic function which significantly determines the patient management by the device, such as closed loop systems or automated external defibrillators, are classified as class III.

The accurate classification is the manufacturer's responsibility besides following the General Safety and Performance Requirements (GSPRs) [11, 12]. The GSPRs, detailed in Annex I of the MDR, set the essential criteria that medical devices must meet to ensure safety and performance before market entry. The GSPRs encompass general safety principles, specific design and manufacturing standards, performance criteria, information requirements, and conditions for certain device types.

Furthermore, the manufacturer of the device is subjected to closely monitor the device's performance and present it to the notified bodies during device assessment, if needed. If device certification requires it, NBs are there to check and confirm the manufacturer's proposed classification, advise the certification route, and estimate the associated costs for the device's CE-marking [13].
If the medical device is used in combination with medicines, with ancillary medicinal substances, or with substances that are systemically absorbed, or the device is of high risk, the device will be classified as class IIa, IIb, or III, based on point 5.2 of Annex IX (Rule 14) [10]. If this is the case, European Medicines Agency (EMA) might participate in the regulatory process and could be consulted by notified bodies during the assessment procedures [14].
CE marking is a legal requirement for DMs to be placed on the European market. The initials "CE" are an abbreviation of the French designation Conformité Européene, which means European Conformity.

The CE marking is represented by the symbol CE the affixing of which must follow certain rules. It is a key indicator of a product's compliance with EU legislation ("New Approach" EU directives) and enables the free movement of products within the European Economic Area (EEA). By affixing the CE marking to a product, the manufacturer declares the conformity of the product with all the relevant ERs of the applicable directives.

3. International Maintenance and Asset Management Standards

In recent years, the concept of sustainability has emerged as a critical priority across various domains, including the healthcare sector. Within the context of medical devices, however, the significance of maintenance transcends environmental considerations, representing a fundamental pillar for ensuring patient safety and the consistent operational integrity of clinical equipment. Device malfunctions may result in adverse clinical outcomes, thereby underscoring the imperative for healthcare institutions to implement robust maintenance strategies. Effective asset management throughout the device life cycle is essential for preserving functional reliability, optimizing performance, and mitigating the risk of failure. This chapter provides a comprehensive analysis of the European regulatory framework concerning the maintenance and life cycle asset management of medical devices, emphasizing the contribution of these standards to sustainable practices, enhanced patient safety, and the overall efficacy of medical technologies.

3.1 European Maintenance Standards

The maintenance of medical equipment is a multifaceted activity that requires adherence to strict standards to ensure safety, functionality, and regulatory compliance. European standards assist professionals in adopting consistent practices aligned with the healthcare sector's requirements. The identified standards and their respective summaries are presented below (Table 1).

Table 1: European Maintenance Standards

Standards	Description
EN 13306:2017- Maintenance Terminology	Provides a common language to describe maintenance concepts and processes.
EN 13460:2009 - Maintenance Documentation	Requirements for technical documentation, essential for medical equipment, enabling traceability in critical environments.
EN 15628:2014 - Maintenance Personnel Qualifications	Specifies competency requirements for specialists in medical equipment maintenance.
EN 17007:2017 - Maintenance and Maintenance Engineering	Addresses integrated maintenance management with a focus on planning and optimization in healthcare environments.
EN 15341:2019 - Maintenance Key Performance Indicators (KPIs)	Offers applicable KPIs to monitor performance and maintenance effectiveness in hospitals and medical centres.
ISO 41000: 2018 – Standards in Facility Management	International standard for facilities management (plan, implement, monitor).

IEC 60601-1 - General Requirements for the Safety of Medical Electrical Equipment	A central standard for electrical safety, essential for the maintenance and inspection of medical electrical equipment.
IEC 62353:2014 - Testing of Medical Electrical Equipment	Guidelines for periodic and post-repair testing of medical electrical devices, ensuring continued safety
ISO 14971:2019 - Risk Management for Medical Devices	Requirements for identifying, assessing, and controlling risks in medical equipment, also applicable to maintenance.
EN ISO 13485:2016 - Quality Management Systems for Medical Devices	Includes requirements related to maintenance and inspection of equipment to ensure compliance with quality management systems.
ISO/TR 24971:2020 - Guidelines for Risk Management Application in Medical Devices	Complements ISO 14971, with practical examples for the sector.
EN 45552-45559 - Durability and Maintenance Standards for Electronic Products	Targeted at the maintenance and repairability of medical devices with electronic components.

These standards form a comprehensive framework of guidelines for the safe, efficient, and regulated maintenance of medical equipment, promoting a safer healthcare environment.

3.2 European Standards Related to Life Cycle Assessment

In the medical device sector, sustainability is a growing concern. From production to disposal, medical equipment can have significant environmental impacts. Applying standards related to Life Cycle Assessment (LCA) allows organizations to quantify and minimize these impacts, fostering practices aligned with the Sustainable Development Goals (SDGs). The identified standards and their respective summaries are presented below (Table 2).

Table 2: European Standards Related to LCA

EN ISO 14040:2006 - Principles and Framework for LCA	Establishes the fundamentals for conducting life cycle assessments (LCA), applicable to medical devices, ensuring environmental impacts are considered from design to end-of-life.
EN ISO 14044:2006 - Requirements and Guidelines for LCA	Details methodologies for performing LCA, ensuring rigor and consistency in results, applicable to the environmental assessment of medical devices throughout their life cycle.
ISO/TS 14072:2014 - LCA Applied to Organizations	Methodology for evaluating the environmental impact of organizations that manufacture medical equipment, considering all processes and products involved in the production of medical devices, with a focus on reducing impacts throughout the life cycle.
ISO 55000:2014 - Asset Management - Overview, Principles, and Terminology	This standard provides the fundamental principles and terminology for asset management, which is essential for the life cycle management of medical devices, as it defines asset management as the process of making strategic decisions about the maintenance, operation, and disposal of assets to maximize the value of medical devices.
ISO 55010:2024 – Asset Management – Guidance on the Alignment of Financial and Non-financial Functions in Asset Management	Alignment between the various areas of an organization, with the aim of maximizing the value obtained from assets, promoting an integrated vision in technical, operational and financial aspects.

ISO 55011:2014 – Asset Management – Guidance for the Development of Public Policy to Enable Asset Management	It aims to ensure that public policies support the adoption of asset management strategies that maximize returns for society, minimize risks and promote efficiency in the use of public resources.
ISO 55012:2024 – Asset Management Guidance on People Involvement and Competence	This standard applies to organizations of all types and sizes. Its principles can also extend beyond formalized asset management systems to other organizational contexts where asset management practices are employed
ISO 55013: 2024 – Asset Management – Guidance on the Management of Data Assets	This standard is applicable across all sectors and types of organization, serving as a strategic tool for integrating data management into asset management practices. This standard also complements existing ISO 55000 series standards by addressing the growing importance of digital data assets in asset-intensive sectors.
ISO 55014: - Guidance for Asset Management Decision Making	This standard provides guidance on developing, implementing, and maintaining a framework for asset management decision-making. It aims to help organizations create consistent and effective processes for making decisions about asset investments, operations, and lifecycle management.
EN ISO 14067:2018 - Carbon Footprint of Products	Provides guidelines for calculating and communicating the carbon footprint of products, especially relevant for disposable or high-energy medical devices such as diagnostic equipment.
EN 15804:2012+A2:2019 - Environmental Product Declarations (EPD)	Specific to the medical device industry, this standard enables the assessment and communication of environmental impacts of devices, creating a transparent platform for companies and consumers.
EN ISO 14025:2010 - Type III Environmental Declarations	Establishes requirements for creating environmental declarations based on LCA, aimed at informing about the environmental impacts throughout the life cycle of medical devices.
ISO 50001:2018 - Energy Management	Focused on continuous improvement of energy performance, this standard is also applicable in the design and operation of medical equipment, promoting energy efficiency throughout the life cycle of devices.
ISO 20400:2017 - Sustainable Procurement	Encourages the selection of medical devices with lower environmental impacts, from acquisition to disposal, supporting the use of sustainability criteria in equipment procurement.
ISO 21929-2:2015 - Sustainability Indicators for Products and Construction	Provides guidelines for evaluating sustainability in materials, especially useful for advanced medical devices that require detailed analysis of environmental impacts.
ISO 18523-2:2021 - Sustainable Life Cycle Management of Products	Focused on integrating sustainability into the design of medical devices, this standard promotes practices that ensure effective life cycle management of medical products.
EN 45557:2020 - Requirements for Reusable Content in Electronic Products	Applicable to medical devices, promoting sustainable design and reuse of electronic components, crucial for reducing waste and environmental footprint.
IEC 62353:2014 - In-service Inspection, Testing, and Maintenance of Electrical Medical Equipment	This standard focuses on the maintenance of medical devices, especially those of an electrical nature, ensuring safety and proper functioning of devices throughout their

	life cycle, with specific inspection and testing procedures.
ISO 14971:2019 - Application of Risk Management to Medical Devices	Though focused on risk management, this standard is crucial for the life cycle of medical devices, as it addresses risk identification, assessment, and mitigation throughout the entire life cycle of the device, including continuous maintenance and operation.
ISO 17025:2017 - General Requirements for the Competence of Testing and Calibration Laboratories	Although not specific to maintenance, this standard is relevant in contexts where calibration and continuous testing of medical equipment are part of the maintenance strategy to ensure that devices remain in optimal operational condition.

4 Metrology

Medical equipment metrology is an essential area of biomedical engineering, as it ensures that devices used in the diagnosis, monitoring and treatment of patients make accurate, safe and reliable measurements. This area ensures the quality of clinical results, protects patient safety and contributes to the efficiency of health services [15].

Metrology is the science of measurement, and in the medical context, it is involved in the verification, calibration and validation of instruments such as: vital signs monitors (blood pressure, ECG, oximetry); infusion pumps; ventilators; thermometers; imaging equipment (ultrasound, MRI, etc.).

This equipment must provide reliable results, as clinical decisions often depend on exact numerical values. Metrology is important because [5]:

- Patient safety: Measurement errors can lead to incorrect diagnoses or ineffective treatments;
- Regulatory compliance: Standards such as ISO/IEC 17025 require equipment to be calibrated by competent laboratories and traceable to international standards;
- Traceability: Each measurement must be traceable to a recognized standard (such as those of the BIPM - Bureau International des Poids et Mesures);
- Quality of health services: Correctly calibrated equipment guarantees consistency in procedures and confidence in clinical results.

Medical metrology involves [5]:

- Periodic calibration of devices by accredited laboratories;
- Functional verification before clinical use;
- Preventive and corrective maintenance based on manufacturers' manuals and standards;
- Evaluation of measurement uncertainties, which is important in critical clinical contexts.

Medical device owners and users are generally medical doctors who are not very familiar with metrology and calibrations. Hospital managers do not know the importance of calibration and think that it has a high cost [16]. Non-metrological calibrations/measurements, and thus, production of improper certificates, reports and labels, are still done by secondary calibration laboratories for economic reasons.

Regulations are missing, and it remains an open question in many countries whether a medical device must be covered by legal or scientific metrology [17].

5 Conclusion

This study confirms that the correct classification of medical equipment, the rigorous application of international standards and the implementation of efficient metrological practices are crucial to guaranteeing the safety and effectiveness of medical devices.

Classification, through a risk-based system, provides a solid starting point for regulatory compliance, while standards such as ISO 13485 and IEC 60601 provide the basis for production and quality control.

Metrology, with its ability to guarantee the accuracy and traceability of measurements, complements these processes and ensures the reliability of devices. However, the constant evolution of technology requires continuous adaptation of these methodologies.

Collaboration between regulators, manufacturers and health professionals will be essential to ensure that medical devices meet international standards and are effective in promoting public health and safety.

The integration of these components will form the basis for the continuous improvement of the safety and quality of medical devices, having a direct impact on the state of the art and the advancement of regulatory practices.

Conflicts of interest: The authors have no competing interests to report.

References

[1] Amaral, C., Paiva, M., Rodrigues, A. R., Veiga, F., & Bell, V. (2024). Global Regulatory Challenges for Medical Devices: Impact on innovation and market access. *Applied Sciences*, *14*(20), 9304. https://doi.org/10.3390/app14209304

[2] Dharavath, D., & Maddi, R. (2022). ISO Standards of medical Devices. *World Journal of Current Medical and Pharmaceutical Research*, 33–39. https://doi.org/10.37022/wjcmpr.v4i3.213

[3] Cheng, M., & Organization, W. H. (2003). *Medical device regulations: Global Overview and Guiding Principles*. World Health Organization.

[4] Tevetia, N., Bhatt, S., Pathak, A., Prakash, S., Bhardwaj, A., Tyagi, M., & Tomar, R. (2022). global prospective of medical devices and their regulations. *International Journal of Health Sciences*, 2764–2778. https://doi.org/10.53730/ijhs.v6ns6.9957

[5] Ferreira, M. D. C. (2011). The role of metrology in the field of medical devices. *International Journal of Metrology and Quality Engineering*, *2*(2), 135–140. https://doi.org/10.1051/ijmqe/2011101

[6] Mojab, Y., Pacifici, E. H., Graham, T. F., Kim, R. E., & Chen, S. W. (2025). Integration of teaching of Digital Health-Driven Medical Devices in Pharmacy Education. *Pharmacy*, *13*(2), 35. https://doi.org/10.3390/pharmacy13020035

[7] Salisbury, J. P. (2021). Using medical device standards for design and risk management of immersive virtual reality for At-Home therapy and remote patient monitoring. *JMIR Biomedical Engineering*, *6*(2), e26942. https://doi.org/10.2196/26942

[8] Branaghan, R. J., Hildebrand, E. A., & Foster, L. B. (2020). Designing for medical device safety. In *Elsevier eBooks* (pp. 3–29). https://doi.org/10.1016/b978-0-12-816427-3.00001-4

[9] Guerra-Bretaña, R. M., & Flórez-Rendón, A. L. (2018). Impact of regulations on innovation in the field of medical devices. *Research on Biomedical Engineering*, *34*(4), 356–367. https://doi.org/10.1590/2446-4740.180054

[10] The European Parliament And The Council Of The European Union. Regulation (EU) 2017/745 of the European Parliament and of the council of 5 April 2017 on medical devices, amending directive 2001/83/EC, regulation (EC) No 178/2002 and regulation (EC) No 1223/2009 and repealing council directives 90/385/EEC and 93/42/EE. Official Journal of European Union; 2017.

[11] *View of Getting your devices ready for MDR compliance – a clinical approach and orthopaedic device manufacturers' perspective*. (n.d.). https://journals.aboutscience.eu/index.php/aboutopen/article/view/294/262

[12] Fennema, P., & Achakri, H. (2019). Getting your devices ready for MDR compliance – a clinical approach and orthopaedic device manufacturers' perspective. *AboutOpen*, *5*(1), 70–75. https://doi.org/10.33393/abtpn.2019.294

[13] CPD with Regulatory Rapporteur, 2019, www.topra.org/CPDsupplements, https://eur-lex.europa.eu/.

[14] European Medicines Agency - Human Regulatory. https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices.

[15] Squara, P., Imhoff, M., & Cecconi, M. (2014). Metrology in Medicine. *Anesthesia & Analgesia*, *120*(1), 66–75. https://doi.org/10.1213/ane.0000000000000477

[16] Turgunbaev, A., Usmanova, H., & Abdurakhmanov, O. (2020). METROLOGICAL CERTIFICATION AND TEST CONDITIONS FOR VERIFICATION EQUIPMENT. *Technical Science and Innovation*, *2020*(4), 124–139. https://doi.org/10.51346/tstu-01.20.4-77-0094

[17] Karaboce, B. (2020). Challenges for medical metrology. *IEEE Instrumentation & Measurement Magazine*, *23*(4), 48–55. https://doi.org/10.1109/mim.2020.9126071

Chapter 5

Fault Analysis of a biomass generator

Contribution #25
Cite as:
André Luis , José Torres Farinha, Mateus Mendes. Fault Analysis of a biomass generator. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-25,
  title ={Fault Analysis of a biomass generator},
  author={André Luis  and  José Torres Farinha and
     Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Fault Analysis of a biomass generator

André Luis[1], José Torres Farinha[1,2], Mateus Mendes[1,2,3]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal.
a2019136057,tfarinha,mmendes@isec.pt
[2] RCM^{2+} Research Centre for Asset Management and Systems Engineering, Rua Pedro Nunes, Coimbra, 3030-199, Portugal.
[3] Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
{a2019136057,tfarinha,mmendes}@isec.pt

Abstract

The efficiency and reliability of electrical generators are essential for power generation systems. Condition monitoring is essential to check system health to deal with unexpected faults that can cause degradation of overall performance over time. Predictive maintenance in electrical generators uses data collected from sensors that need to be processed to distinguish between normal and anomalous functioning states. This paper explores one practical implementation to check biomass generator system condition through a given data set, enabling the distinction between normal and anomalous functioning states using efficient machine learning methods, PCA to reduce data set dimensionality, keeping most of the information about the data set, and DBSCAN to cluster the new variables obtained by PCA.

Keywords: Condition Monitoring, PCA, DBSCAN, Electrical Generator, Biomass Generator

1 Introduction

Electrical generators are often faced with unexpected faults that can result in degradation of the overall performance of the system and affect production in several ways, or even completely stop it. In order to prevent this problem from happening, equipment monitoring and predictive maintenance are used.

Sensor technologies and computational methods have evolved in order to enable increasingly more accurate and timely fault detection. McKinsey predicts that digital maintenance in the industry can increase asset availability by 5 to 15 percent and reduce maintenance costs by 18 to 25 percent [1].

The aim of the present work is to study the operation of a biomass electrical generator and some possible occurrences of faults, applying machine learning methods to detect patterns in data, allowing to distinguish between normal operating and abnormal operating states.

The generator operating condition is described in a data set that contains multiple variables relative to sensors of different parameters, such as temperature, rotation velocity, or voltage, among others. The data were submitted to different analytics methods. Principal Component Analysis (PCA) was used to reduce dimensionality and facilitate visualization. That was followed

by the application of DBSCAN to identify new clusters with the new data points. That resulted in different subsets that reflected each functioning state of the generator.

Despite the outstanding importance of other types of energy, such as hydroelectric energy, biomass energy represents an important source in many countries, especially in developing countries. In 2023, biomass represented 5% of the total primary energy consumed in the United States[1]. The other developed countries in general follow the same trend, using biomass as a type of complementary energy.

The first model of electrical equipment maintenance focused on emergency maintenance, to repair equipment if a failure occurred in normal functioning or was detected in periodic inspections, performed to check the overall state of the production line.

With technological development, Condition-Based Maintenance (CBM) became possible. In this way, limited maintenance interventions can be made only on machines that need it, when they need it. Input data are needed, to describe in a reliable and objective way the actual condition of the equipment through several installed sensors. Those sensors are related to electrical, mechanical or temperature variables [2]. Such capacity for early problem detection can be crucial to achieve a higher performance, efficiency, reliability, and safety, especially for processes like power plants.

In terms of the structure of the manuscript, it is as follows. Section 2 presents a review of the state of the art. Section 3 describes the materials and methods used during the research. Section 4 presents all results obtained in the study, including graphs, tables and explanations. Section 5 presents the discussion about the contributions of the study. Section 6 presents the conclusions about the study.

2 State of the art

2.1 Maintenance practices in Hydro/Turbine Generators

The architectures and maintenance practices used for Hydro/Turbine generators are similar to those applied in Biomass generators. Therefore, the maintenance approaches discussed in the referenced article on Hydro/Turbine generators, specifically regarding predictive maintenance can also be effectively applied to Biomass generators in this context.

According to Wang *et al.* [3], the current maintenance practice applied to Hydro/Turbines Generators (HTG) is mainly preventive maintenance. Preventive maintenance is a more efficient approach to HTG, serving the objective of carrying out the appropriate maintenance work to maximize the equipment's service life. Predictive Maintenance (PM), however, is increasingly more popular, as sensors, computing power and algorithms advance. Predictive maintenance actions can be divided into two types: Diagnostic actions and Prognostic actions. Diagnostic deals with fault detection, isolation and identification when they occur. Prognostic deals with fault prediction before the faults occur. From the perspective of predictive maintenance, it consists in three key steps: Data Acquisition, Data Processing and Data Mining.

2.2 Applications of PCA and DBSCAN

PCA and DBSCAN are the two machine learning algorithms used to identify different state faults in the current paper. Some articles studied implementations that can serve as base to implement these algorithms in the current article.

[1]Biomass explained by U.S. Energy Information Administration (EIA), https://www.eia.gov/energyexplained/biomass/ (last access on 10 March 2025).

In Shetty *et al.* [4], the goal is to find the best hyperparameters input to feed DBSCAN for environmental assessment on air quality and mitigation strategy. Predicting air pollution is a complex task due to various challenges associated with the dynamic and multifaceted nature of the atmosphere, such as meteorological variability, local emissions and sources, data quality and availability, and emerging pollutants. Extensive experiments prove that the search tree approach to find Eps, a hyperparameter for DBSCAN, is quicker and more efficient in comparison to the widely used KNN algorithm. The time reduction to find Eps makes a significant impact as the dataset size increases. The input parameters determine in part the DBSCAN algorithm's clustering results.

Hao *et al.* [5] propose a new algorithm based on the analysis of wavelet decomposition of pulse signals recorded during Partial Discharge experiments. Principal Component Analysis is applied to reduce the number of energy values derived from each decomposition level. The DBSCAN algorithm has been utilized to discover clusters of arbitrary shapes in an n-dimensional space.

Li *et al.* [6] use a method of combining principal component analysis and DBSCAN, adopted to establish the aggregated model of a permanent magnet synchronous generator offshore wind farm. MATLAB/Simulink is employed in order to perform the simulation verification. The result shows that the proposed model can accurately represent the dynamic response of a wind farm.

Muskatin *et al.* [7] propose a comparison study about the application of DBSCAN and PCA-DBSCAN algorithm to group an earthquake area. Clustering in this study was conducted through two experiments. The first experiment used DBSCAN without dimensionality reduction and the second experiment used DBSCAN clustering with dimensional reduction using PCA. The highest Silhouette Index was obtained using DBSCAN and PCA.

Savvas *et al.* [8] propose a study about the application of PCA and DBSCAN to maximize the quality and perfomance in detection of anomalies. Decreasing the dimensionality of the data points, the quality was also decreasing. However, the quality of the solutions produced by applying PCA to the corresponding data sets, until dimension 4, is above 80%, which is acceptable in most cases. The results obtained with the application of DBSCAN to PCA data points were very promising, since DBSCAN produces clusters of high quality even when the dimension is reduced to 3.

Ni and Jinhang [9] use a clustering algorithm framework based on principal component analysis. A classification simulation for iris data set was performed using DBSCAN algorithm to cluster the iris data set. Comparing the results of classification with the real results, the accuracy obtained was up to 80% by the DBSCAN algorithm. Another experiment involved the application of PCA to the iris raw data set with four rescaled principal components obtained. The cumulative variance contribution rate of Component 1 and Component 2 is up to 91.541%, the addition of Component 3 led to 98.789% of the variance. A new iris data set was constructed by the main components C1 and C2 and clustered by DBSCAN. The clustering accuracy can be up to 78%.

Yiu [10] propose a self-adaptive incremental PCA-Based DBSCAN of acoustic features for anomalous sound detection. To resolve issues related to high volume of the training dataset and the imbalance between normal and anomalous data, the algorithms used for Anomalous Sound Detection (ASD) like deep convolutional neural networks and generative adversary networks require one or more graphic processing units to process and generate good predictive results. To resolve those issues in practice, the proposed algorithm integrates PCA with unsupervised DBSCAN.

Bajwa *et al.* [11] applied PCA and DBSCAN in wavelet energy distribution for partial discharge pulse extraction. Partial discharge pulse detection and extraction is the first step in condition monitoring. The research is focused on assessing the capability of wavelet energy distribution for pulse extraction with the help of PCA and DBSCAN. Partial Discharges, if left

untreated, can result in complete breakdown of insulation over long periods. Principal component analysis was used to reduce the dimension of the wavelet energy distribution, so that it could be plotted in a 3-D space. Wavelet energy with PCA and DBSCAN proved to be a very effective method when it came to extracting PD pulses.

Wibisono *et al.* [12] used the DBSCAN algorithm to separate normal and anomalous weather data, considering multiple weather variables. Then, PCA is used to visualize the clusters. The experimental result demonstrated that DBSCAN is capable of identifying peculiar data points that are deviating from the normal data distribution.

2.3 Relations between IT fields

The current article presents a exploratory case where occurs relations between IT fields, namely between machine learning and Internet of Things given that the data representative of generator states comes from generator installed sensors that need to be processed using machine learning algorithms to reveal the generator behavior.

Li *et al.* [13] present a review about the literature and patents in four closely related fields: Artificial Intelligence, Big data, Internet of Things and Blockchain. They aim is to provide a holistic view of how they are related and their integrability in relation to smart energy management strategies.

3 Methodology

The present article contains a data analysis of a biomass electrical generator and proposes to detect abnormal operations that are evident from the data. Machine learning models are used to study the data in order to implement the data mining process to extract useful information from the dataset, its patterns and useful information that will be necessary to diagnose possible abnormal states.

The research is supported by the utilization of some methods to serve the purpose of distinguishing the normal generator functioning from the abnormal functioning. One of the methods to use is Principal Component Analysis, which consists of a method used to reduce the dimensionality of large data sets by transforming a large set of features into a smaller one that still contains most of the information of the large data set. The method transforms the data set into a simpler version where some variance may be lost, in exchange of simplicity that is beneficial to explore and visualize, because a smaller data set is easier to process for machine learning algorithms. The new data set consists of new variables (principal components), sorted by the amount of variance that they explain. Therefore, the first components are the most important for analysis, while the last ones are the least relevant for analysis.

After the conversion of the data set variables into new ones using PCA, a clustering technique was used to facilitate visualization and further processing. Clustering is effective in facilitating the identification and removal of noise in a data set. It is also useful for data cleaning and outlier detection. The method used was DBSCAN.

DBSCAN works by partitioning data into dense regions of points that are separated by less dense or empty areas. It defines similarity-based clusters. The points in a cluster are close to each other. Points that are far from any cluster are considered outliers or noise. Other techniques to partition observations into clusters include K-Means algorithm and K-Medoids. However, DBSCAN stands out for being more efficient than other algorithms, according to a study conducted by C. Kondal Raj [14].

4 Results

4.1 Dataset used

The data set used consisted in a total of 17,569 entries from 35 features of different sensor types related to temperature, pressure, power, current, frequency, or vibration. The 17,569 entries provided were collected between 2024-08-03 and 2024-10-03. The samples were collected with a period of 5 minutes. Table 1 describes each data set feature.

Table 1: Features Descriptions and Values for TG4 Parameters. Min is the minimum, Max the maximum, STD the Standard Deviation

Feature	Description	Mean	Min	Max	STD
TG4 - Steam Flow AP	Steam flow in power plants controls the fuel flow. When homogeneous, gives very predictable performance and stable operation	347.78 t/h	0.0 t/h	471.39 t/h	181.78 t/h
TG4 - Steam Temperature AP	Warm air temperature related to the steam flow	404.7 ºC	0.0 ºC	484.94 ºC	147.65 ºC
TG4 - Active Power	Active power is the electrical power that converts electrical energy into other forms of energy	49.7921 MW	-0.1669 MW	66.6450 MW	25.89 MW
TG4 - Power Factor	Measures how efficiently a machine uses its energy. Ideally, would be 1	0.78229	0.0	0.99976	0.38
TG4 - Reactive Power	Exchange of electric and magnetic fields in the circuit	7.9297 MVAr	-4.5368 MVAr	22.336 MVAr	9.27 MVAr
TG4 - Frequency	Directly related to rotation velocity. Ideally 50 Hz	49.989 Hz	0.0 Hz	50.0803 Hz	0.75 Hz
TG4 - Turbine Speed CH0	Turbine rotations per minute. Ideal 3000 rpm for 50 Hz frequency	2424.1570 1/min	0.0 1/min	3005.4990 1/min	1171.32 1/min
TG4 - Turbine Speed CH1	Turbine rotations per minute. Ideal 3000 rpm for 50 Hz frequency	2424.0477 1/min	0.0 1/min	3005.0129 1/min	1171.26 1/min
TG4 - Temperature Generator Bearing, front	Temperature from the front bearing	77.628 ºC	0.0 ºC	89.40 ºC	21.49 ºC
TG4 - Temperature Generator Bearing, rear	Temperature of the rear bearing	76.089 ºC	0.0 ºC	87.4926 ºC	21.23 ºC
Shaft Vibration Gen. Journal Bearing, front	Two transducers, 90 º to each other, measure vibration in the same plane at each bearing	21.2670 um	−1.10 um	32.4289 um	10.29 um
TG4 Current L1	Current at phase L1	1493.6919 A	0.0 A	2066.414 A	775.68 A
TG4 Current L2	Second phase generator current value.	1475.9438 A	0.0 A	2652.66 A	766.72 A
TG4 Current L3	Third phase generator current value.	1428.0716 A	0.0 A	2534.68 A	741.17 A
Generator Warm Air Temperature	Warm air temperature at the generator output	56.7 ºC	0.0 ºC	67.6 ºC	14.02 ºC
Shaft Vibration Gen. Journal Bearing, front	First sensor relative to Shaft Vibration Generator Journal Bearing of the front.	21.2670 um	−1.10 um	32.4289 um	10.28 um
Shaft Vibration Gen. Journal Bearing, front	Second sensor relative to Shaft Vibration Generator Journal Bearing of the front.	16.4974 um	−1.62 um	25.227 um	7.98 um
Shaft Vibration Gen. Journal Bearing, rear	First sensor relative to Shaft Vibration Generator Journal Bearing of the rear.	18.921 um	0.0 um	26.029 um	9.03 um
Shaft Vibration Gen. Journal Bearing, rear	Second sensor relative to Shaft Vibration Generator Journal Bearing of the rear.	14.548 um	0.0 um	18.7746 um	5.16 um

4.2 Correlation Matrix

A correlation matrix was created, to explore the correlation coefficients between the variables, as a first step to search for patterns between those variables. The correlation indicates how strong

the relationship between two variables is. This can be important to better understand which variables are the most important in the model, which follow the same trends, and which follow opposite trends.

Figure 1 shows a heat map of the correlation matrix. In the figure, it is possible to see a high correlation between a vast majority of the features. Some of the features are highly correlated, namely the Steam Flow, Colector Pressure, Temperature Steam, Active Power, Power Factor, Current L1, Current L2, Current L3, Turbine Speed CH0, Turbine Speed CH1, Temperature Bearing Front, Stator Slot Temperature Phase U, Temperature Phase V, Temperature Phase W, Warm Air Temperature, Temperature Generator Bearing Rear, Shaft Vibration Generator Journal Bearing Front and Vibration Generator Journal Bearing Rear. On the other hand, the features Reactive Power, Bus Voltage and Frequency present a low correlation to the other dataset features. Table 2 describes the abbreviations used in Figure 1.

Table 2: Feature Abbreviations for Correlation Matrix

Feature	**Abbrev.**
Steam Flow	SF
Collector Pressure	CP
Steam Temperature	ST
Active Power	AP
Power Factor	PF
Current L1	CL1
Reactive Power	RP
Bus Voltage	BV
Frequency	FQ
Turbine Speed CH0	TSCH0
Cold Air Temperature 1	CAT1_1

Feature	**Abbrev.**
Temperature Bearing front	TBF
Warm Air Temperature	WAT
Stator Slot Temperature Phase U	SSTPU_1
Stator Slot Temperature Phase V	SSTPV_1
Stator Slot Temperature Phase W	SSTPW_1
Cold Air Temperature 2	CAT2_1
Temperature Generator Bearing rear	TGBR
Shaft Vibration Generator Journal Bearing front	SVGJBF_1
Shaft Vibration Generator Journal Bearing front	SVGJBF_2
Shaft Vibration Generator Journal Bearing rear	SVGJBR_1
Shaft Vibration Generator Journal Bearing rear	SVGJBR_2

4.3 Features behavior

Through graphic representation of each feature, it is possible to see an irregular behavior throughout the period of the data set. That will be analyzed in order to identify the potential causes of this unexpected behavior.

Graphs were plotted using the matplotlib pyplot library. Analysing the behavior of the majority of the features it is possible to see three major periods of abnormal states, where some anomalies in the normal behavior of the features are visible.

The first verified anomaly is depicted in Figure 2a. It occured in the interval from 2024-08-21 to 2024-08-23. It is possible to verify a small anomaly that could be caused by missing data given that it occurs in a 10 minutes interval where no data was registered. The features Current L1, Current L2, Current L3, Power Factor, Bus Voltage and Frequency registered this behavior during this interval occured between 17:50 and 18:00 on the day 2024-08-21. The other features continued their normal behavior. A possible explanation behind this anomaly can be due to a sudden disconnect from the power source, as there is no driving source to maintain it. This is expected in a event of total loss of supply. Another explanation to this phenomenon can rely on an activation of the system protection mechanism in result of a fault detection which forced a load disconnect. Another possible explanation can be a sensor fault.

During the following day, 2024-08-22, another anomaly occured around 13:35 and the periods between 14:10 and 14:30, as shown in Figure 2a. The majority of the features not affected by the anomaly occurred on 2024-08-21 showed unusual patterns. During these intervals, the features Steam Flow, Collector Pressure, Steam Temperature, Turbine Speed CH0, Turbine Speed CH1,

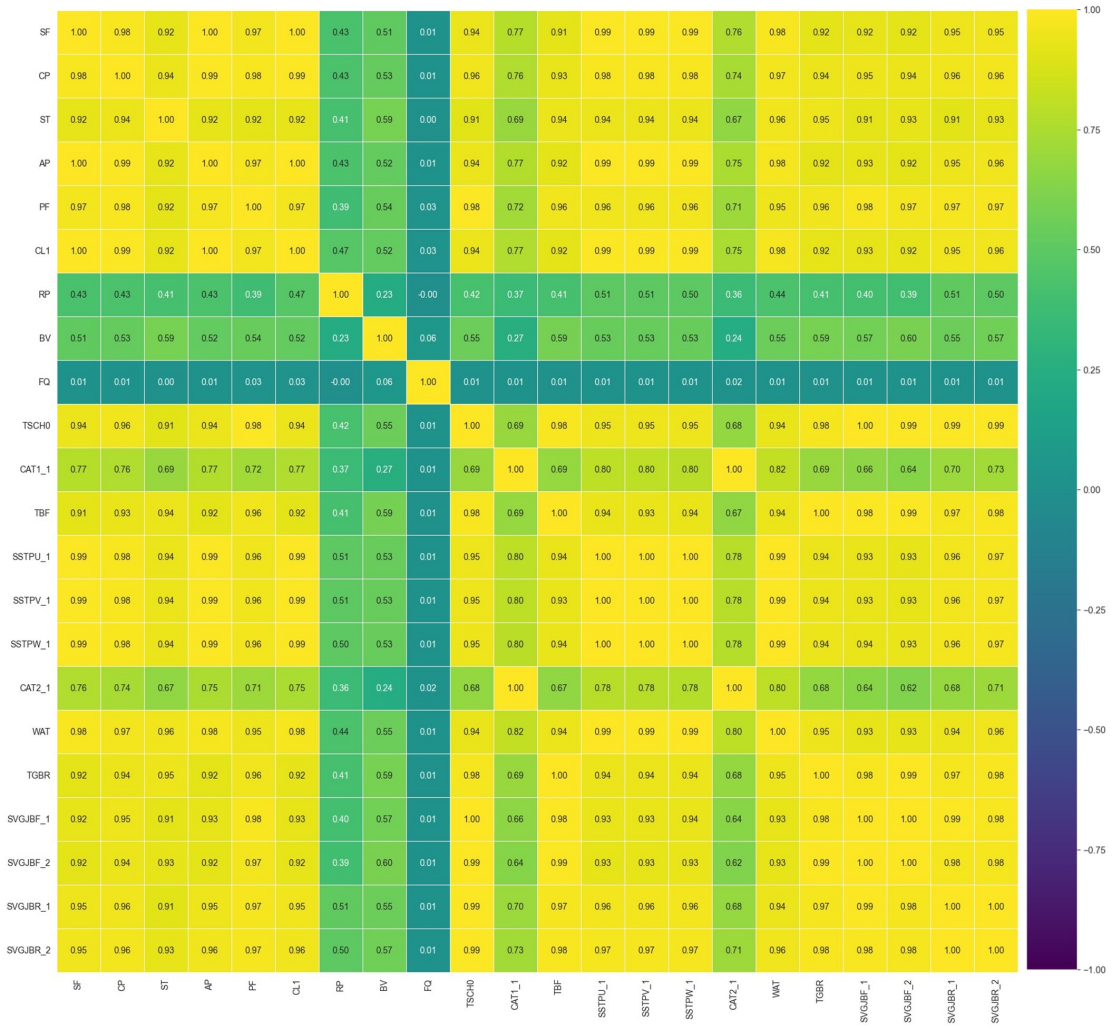


Figure 1: Dataset Correlation Matrix

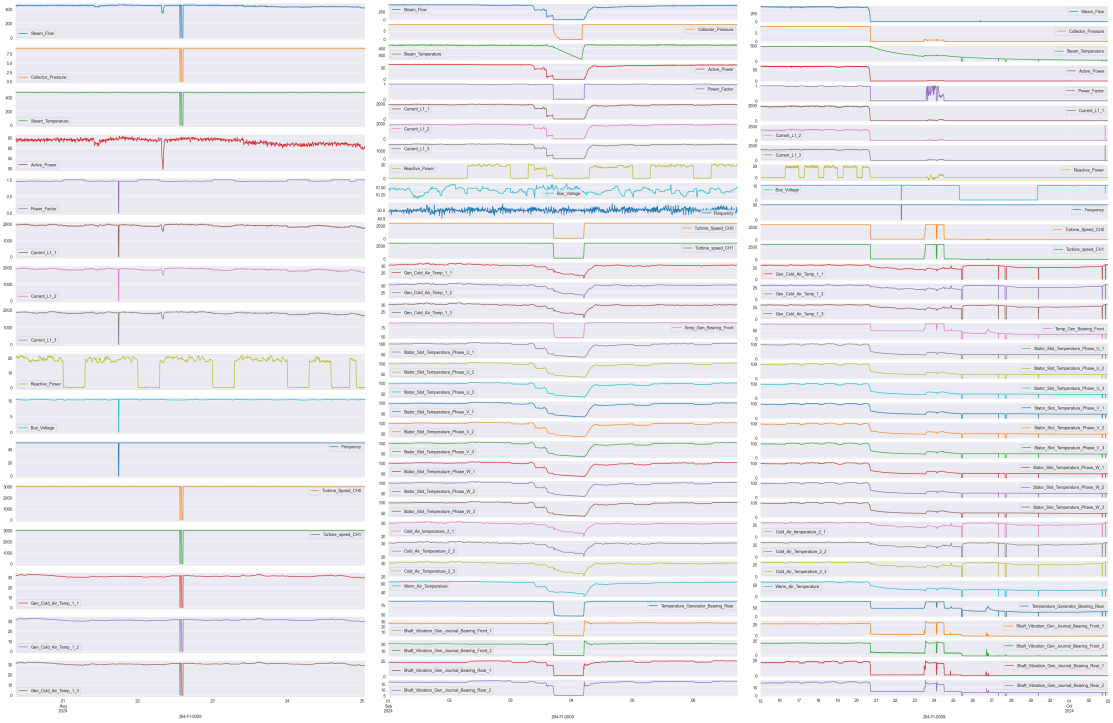
Generator Cold Air Temperature 1, Generator Cold Air Temperature 2, Generator Cold Air Temperature 3, Temperature Generator Bearing Front, Temperature Generator Bearing Rear, Stator Slot Temperature Phase U, Stator Slot Temperature Phase V, Stator Slot Temperature Phase W, Cold Air Temperature, Warm Air Temperature, Shaft Vibration Generator Journal Bearing Front and Shaft Vibration Generator Journal Bearing Rear registered a normal behavior before these periods and suddenly dropped to 0 value. This indicates some possible faults in data collection of these features. The components involved could register a normal behavior in the generator at the time, but for some reason the corresponding sensors could not detect or register the values, which can be a plausible explanation since the values relative to temperatures registered an abrupt drop instead of a progressive one, as characteristic of these types of variables.

Figure 2b presents the reported behavior of the second identified abnormal state. Analyzing the graphs that describe the general feature behavior, a second identified abnormal state affecting the majority of the features is possible to visualize between the periods 2024-09-03 16:45 and 2024-09-04 05:10. A progressive drop on the normal values of some features occurs. It is possible to visualize a sharp reduction on the turbine speed related feature, where the values decay from around 3000 rotations per minute to 100 rotations per minute in a period of 25 minutes, although the frequency variable values remain constant. This reduction of turbine speed can be justified by a previous gradual drop on steam flow and active power registered since 2024-09-03 13:20, which caused a consequent reduction of produced L1, L2 and L3 currents. Temperatures related to generator cold and warm air, generator stator slot temperature values also started to decrease since 2024-09-03 13:20, which affected the generator turbine speed, generator bearing front and rear and shaft vibration some hours later, from 2024-09-03 16:45 to 2024-09-04 05:10. In the same way, these values started to decrease at the same rate.

A third identified abnormal state is possible to visualize, starting at approximately 2024-09-20 13:30, when the values of some features started to decrease progressively, such as the steam flow, which is directly correlated to active power, reaching a 0 value by 2024-09-20 16:55. Other features started to decrease as well. Some of them only started to be affected by the reduction of the active power, the power that is consumed as energy to fuel the generator, like the features related to turbine speed that only around 2024-09-20 16:45 started to rapidly decrease as a consequence of limited active power.

The temperature-related features registered a slower and more prolonged decrease, as is characteristic for this type of variables. This behavior seems to suggest that a controlled decrease in generator production occurred. The anomaly started in the generator energy source. Some days after the general breakdown that happened across all features around 2024-09-23 11:35, some features registered a progressive increase. That seemed to be intentional, given that the power factor progressively increased. The same for the turbine speed, which reached its normal functioning from 2024-09-23 13:30 to 2024-09-24 02:40. Then another breakdown occurred, pushing turbine speed functioning to a minimum functioning followed by another progressive increase until 2024-09-24 03:45, when a normal functioning was reached. Then a new and permanent breakdown was recorded until the end of the sampling period. This behavior is similar for this period for the shaft vibration generator journal bearing features, due to a strong correlation between those features. A similar behavior is registered during this period for the power factor feature, which shows successive ascents and breakdowns. However, this scenario practically did not influence the current features values. The same for the different phases of stator slot temperature, which remained low because it was a short-lived change. The steam flow, highly correlated to active power feature, did not suffer any perturbation in this period. Therefore, it was not involved in this anomaly. Figure 2c presents the reported behavior of the third identified abnormal state.

It is not clear if the deviation from standard operation from first abnormal state is intentional without the confirmation of the operation status of the generator. Second and third identified ab-



(a) First identified abnormal state (b) Second identified abnormal state (c) Third identified abnormal state

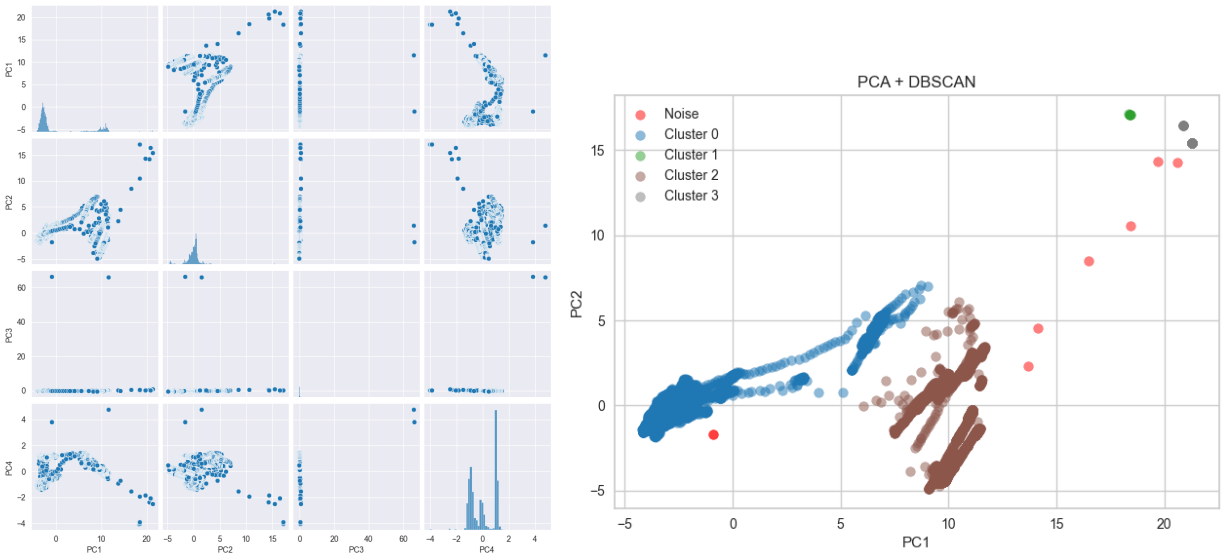
Figure 2: Identified abnormal operation sets

normal states seems to suggest a controlled decrease in generator production given that generator energy source goes into progressive decline.

4.4 PCA Analysis

Given that many of the data set features exhibit high correlation between them, it is possible to apply a Principal Component Analysis for dimensionality reduction. PCA is used mostly to reduce the number of dimensions of large data sets. The purpose of PCA is to reduce the number of variables loosing the least possible amount of information. The objective of PCA is to form a new coordinate system with maximum variance to group data based on similarity. In general, the steps of PCA are as follows:

1. Data standardization
2. Calculate the covariance matrix
3. Calculate eigen vector and eigen value
4. Select the component and feature vector
5. Obtain new data set



(a) PCA graph

(b) PCA + DBSCAN

Figure 3: Visualization of the dataset using PCA components and clusters formed using DBSCAN algorithm.

In the present work, the number of principal components used was adjusted along the experiments. The amount of variance explained by the first four components was, respectively: 0.839, 0.07, 0.029, and 0.023, for Principal Component 1, Principal Component 2, Principal Component 3 and Principal Component 4. Given that PC3 and PC4 represent very little significant variance values, clustering was performed with only PC1 and PC2.

Figure 3a shows the graphical representation of the distribution of the data points relying on PCA applied to the current data set. Among these graphs stands out the one that shows PC2 (y) in function of PC1 (x), for those components explain most of the variance. The graph shows a high correlation between both components. Two clusters and some noise are also clearly visible in this visualization.

4.5 Clustering

A pratical implementation of the DBSCAN clustering with dimensional reduction using PCA results was performed in the data set.

The python DBSCAN function has two parameters, eps and min samples. Eps represents the maximum distance between two samples for one to be considered as in the neighborhood of the other. The higher the value of Eps the less clusters will be formed. Min samples is relative to the number of samples in a neighborhood for a point to be considered as a core point. A high value for min samples will form denser clusters, while a low value will generate sparser clusters. In the practical implementation, the values chosen for DBSCAN were 1.2 for eps and 5 for min samples. This resulted in a set of four clusters, which can be seen in the graphical representation of Figure 3b. The points relative to -1 cluster are considered areas with low point density and are called noise.

After clustering, some statistics of the clusters found were calculated, in order to better understand the results generated. Table 3 shows some statistics of the clusters, including the

Table 3: Clusters formed. Num Points is the number of points in the cluster

Cluster	Num Points	% of Total	Sillhouette score
-1	8	0.05	-0.57
0	14139	80.4	0.84
1	6	0.03	0.98
2	3369	19.18	0.77
3	47	0.27	0.89

number of points relative to each cluster. The respective percentage of the number of points in relation to the total. It is possible to see that cluster 0 aggregates 80% of the total of data points. Cluster 2 contains 19% of the data points, then cluster 3 contains approximately 0.3% of the data points and cluster 1 contains only 0.03 % of data points. Clusters 0 and 2 aggregate more than 99% of the data points. The remaining clusters contain little significant data.

Table 3 also shows the Silhoutte Score for each cluster. Silhouette is a method of interpretation and validation of consistency within data clusters. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A clustering with an average silhouette of over 0.7 is considered to be strong, a value over 0.5 is considered as reasonable and about 0.25 weak. The silhouette score is specialized for measuring cluster quality when the clusters are convex-shaped and may not perform well if the data clusters have irregular shapes or are of varying sizes.

According to Table 3, the best silhouette scores are for cluster 2 with 0.98, cluster 0 with 0.84 and cluster 1 with 0.77. The overall silhouette score excluding noise points was calculated and the result obtained was 0.83, which indicates that the data points are well matched to their own clusters.

Table 4 present clusters' statistics related to mean, minimum, maximum and standard deviation values of each cluster, along with comparison to the whole dataset. 'Cl' stands for Cluster. 'St Dev' stands for standard deviation.

Table 4: Features Data Across Clusters

Feature	Statistic	Dataset	Noise	Cl 0	Cl 1	Cl 2	Cl 3
Steam Flow	Mean	347.78	115.38	432.08	0	0.03	0.0
	Minimum	0.0	0.0	0.0	0.0	0.0	0.0
	Maximum	471.39	462.31	471.4	0.0	35.63	0.0
	St Dev	181.78	213.65	68.07	0.0	0.92	0.0
Collector Pressure	Mean	6.99	2.19	8.67	0.0	0.04	0.00
	Minimum	-0.02	-0.02	0.26	0.0	-0.02	-0.01
	Maximum	8.98	8.83	8.98	0.0	8.91	0.0
	St Dev	3.56	4.09	1.09	0.0	0.53	0.0
Steam Temperature	Mean	404.71	140.03	470.37	0.0	136.07	0.0
	Minimum	0.0	1.53	171.79	0.0	30.72	0.0
	Maximum	483.94	476.48	483.94	0.0	467.89	0.0
	St Dev	147.65	208.50	40.22	0.0	117.74	0.0
Active Power	Mean	49.79	16.03	61.83	64.68	0.0	0.0
	Minimum	-0.17	0.0	-0.16	64.22	0.0	0.0
	Maximum	66.65	64.13	66.64	65.53	0.0	0.0
	St Dev	25.89	29.69	9.38	0.45	0.0	0.0
Power Factor	Mean	0.78	0.0	0.97	0.96	0.0	0.0
	Minimum	0.0	0.0	0.0	0.96	0.0	0.0
	Maximum	0.99	0.0	0.99	0.96	0.0	0.0
	St Dev	0.39	0.0	0.07	0.0	0.0	0.0
Current L1	Mean	1493.69	0.0	1855.20	1998.14	0.0	0.0

	Minimum	0.0	0.0	0.0	1986.76	0.0	0.0
	Maximum	2066.41	0.0	2066.41	2032.13	0.0	0.0
	St Dev	775.68	0.0	276.59	17.75	0.0	0.0
	Mean	1475.94	0.0	1832.96	1979.24	0.83	0.0
Current L2	Minimum	0.0	0.0	0.0	1967.33	0.0	0.0
	Maximum	2652.66	0.0	2041.96	2011.19	2652.66	0.0
	St Dev	766.71	0.0	274.60	16.99	45.77	0.0
	Mean	1428.07	0.0	1773.51	1900.30	0.79	0.0
Current L3	Minimum	0.0	0.0	0.0	1887.23	0.0	0.0
	Maximum	2534.68	0.0	1973.73	1930.42	2534.68	0.0
	St Dev	741.17	0.0	263.36	17.19	43.74	0.0
	Mean	7.93	4.68	9.84	19.62	0.0	0.0
Reactive Power	Minimum	-4.54	0.0	-4.54	18.82	0.0	0.0
	Maximum	22.34	18.72	22.34	20.38	0.0	0.0
	St Dev	9.28	8.67	9.38	0.63	0.0	0.0
	Mean	9.70	2.59	10.39	10.28	6.90	1.77
Bus Voltage	Minimum	0.0	0.0	10.09	10.27	0.0	0.0
	Maximum	10.70	10.40	10.67	10.28	10.70	10.45
	St Dev	2.59	4.80	0.09	0.0	4.89	3.95
	Mean	49.99	37.50	50.0	49.99	49.97	50.0
Frequency	Minimum	0.0	0.0	49.90	49.97	0.0	49.98
	Maximum	50.08	50.02	50.08	49.99	50.08	50.04
	St Dev	0.75	23.14	0.02	0.01	1.22	0.01
	Mean	2424.16	756.43	3000.04	0.0	49.41	0.00
Turbine Speed CH0	Minimum	0.00	0.00	2207.06	0.0	0.00	0.00
	Maximum	3005.50	2999.87	3005.50	0.0	2176.89	0.00
	St Dev	1171.32	1384.27	8.99	0.0	140.52	0.00
	Mean	2424.05	474.71	2999.90	0.0	49.41	0.00
Turbine Speed CH1	Minimum	0.00	0.00	2194.66	0.0	0.00	0.00
	Maximum	3005.01	2999.56	3005.01	0.0	2173.78	0.00
	St Dev	1171.26	1384.20	9.08	0.0	140.52	0.00
	Mean	29.77	15.36	30.84	0.0	25.74	0.00
Gen Cold Air Temp1_1	Minimum	0.00	1.70	18.33	0.0	18.17	0.00
	Maximum	33.60	31.29	33.60	0.0	32.42	0.00
	St Dev	3.13	11.61	1.44	0.0	2.56	0.00
	Mean	29.73	15.09	30.76	0.0	25.95	0.00
Gen Cold Air Temp1_2	Minimum	0.00	1.27	18.49	0.0	18.31	0.00
	Maximum	33.59	31.20	33.59	0.0	32.90	0.00
	St Dev	3.05	11.94	1.40	0.0	2.64	0.00
	Mean	29.35	15.21	30.34	0.0	25.72	0.00
Gen Cold Air Temp1_3	Minimum	0.00	1.69	18.26	0.0	17.99	0.00
	Maximum	33.18	30.70	33.18	0.0	32.20	0.00
	St Dev	2.96	11.38	1.37	0.0	2.51	0.00
	Mean	77.63	31.14	87.92	0.0	35.78	0.00
Temp Gen Bearing Front	Minimum	0.00	1.56	75.03	0.0	24.70	0.00
	Maximum	89.4	87.60	89.40	0.0	83.57	0.00
	St Dev	21.49	35.79	0.56	0.0	10.50	0.00
	Mean	76.09	30.80	86.24	0.0	34.80	0.00
Temp Gen Bearing Rear	Minimum	0.00	2.40	71.96	0.0	23.67	0.00
	Maximum	87.49	86.19	87.49	0.0	79.62	0.00
	St Dev	21.23	35.04	0.54	0.0	10.64	0.00
	Mean	85.90	34.41	99.44	0.0	30.53	0.00
Stator Slot Temp PhaseU_1	Minimum	0.00	1.87	26.11	0.0	23.91	0.00
	Maximum	109.58	105.58	109.58	0.0	54.36	0.00
	St Dev	29.04	44.28	9.94	0.0	4.73	0.00
	Mean	85.50	34.26	98.96	0.0	30.51	0.00
Stator Slot Temp PhaseU_2	Minimum	0.00	1.84	26.12	0.0	23.80	0.00
	Maximum	109.11	105.02	109.11	0.0	54.0	0.00
	St Dev	28.84	44.04	9.85	0.0	4.64	0.00
	Mean	84.84	34.10	98.19	0.0	30.25	0.00
Stator Slot Temp PhaseU_3	Minimum	0.00	1.83	25.93	0.0	23.46	0.00
	Maximum	108.09	104.10	108.09	0.0	53.89	0.00
	St Dev	28.61	43.54	9.71	0.0	4.62	0.00
	Mean	84.35	33.83	97.51	0.0	30.56	0.00
Stator Slot Temp PhaseV_1	Minimum	0.00	1.84	25.90	0.0	23.60	0.00
	Maximum	107.32	103.28	107.32	0.0	53.49	0.00
	St Dev	28.23	43.28	9.67	0.0	4.49	0.00
	Mean	85.64	34.18	99.18	0.0	30.28	0.00
Stator Slot Temp PhaseV_2	Minimum	0.00	1.84	25.59	0.0	23.60	0.00
	Maximum	109.30	105.17	109.30	0.0	53.89	0.00
	St Dev	29.04	44.20	9.97	0.0	4.74	0.00
	Mean	86.90	34.37	100.73	0.0	30.32	0.00
Stator Slot Temp PhaseV_3							

	Minimum	0.00	1.23	25.72	0.0	23.61	0.00
	Maximum	110.84	106.82	110.84	0.0	54.77	0.00
	St Dev	29.65	45.15	10.10	0.0	4.76	0.00
Stator Slot Temp PhaseW_1	Mean	86.20	34.42	99.87	0.0	30.27	0.00
	Minimum	0.00	1.85	25.76	0.0	23.49	0.00
	Maximum	110.26	105.82	110.26	0.0	54.32	0.00
	St Dev	29.30	44.45	9.92	0.0	4.76	0.00
Stator Slot Temp PhaseW_2	Mean	85.66	34.33	99.11	0.0	30.69	0.00
	Minimum	0.00	1.85	26.13	0.0	23.97	0.00
	Maximum	109.43	105.02	109.43	0.0	54.47	0.00
	St Dev	28.83	44.00	9.84	0.0	4.55	0.00
Stator Slot Temp PhaseW_3	Mean	84.83	33.63	98.13	0.0	30.45	0.00
	Minimum	0.00	1.04	25.96	0.0	23.67	0.00
	Maximum	108.09	103.83	108.09	0.0	53.69	0.00
	St Dev	28.52	43.76	9.72	0.0	4.58	0.00
Gen Cold Air Temp2_1	Mean	29.09	15.14	30.11	0.0	25.28	0.00
	Minimum	0.00	1.67	18.03	0.0	17.68	0.00
	Maximum	32.90	30.55	32.90	0.0	30.42	0.00
	St Dev	3.04	11.33	1.39	0.0	2.65	0.00
Gen Cold Air Temp2_2	Mean	29.58	15.30	30.69	0.0	25.43	0.00
	Minimum	0.00	1.69	18.14	0.0	17.96	0.00
	Maximum	33.54	31.20	33.54	0.0	30.62	0.00
	St Dev	3.19	11.59	1.45	0.0	2.70	0.00
Gen Cold Air Temp2_3	Mean	29.26	15.16	30.32	0.0	25.27	0.00
	Minimum	0.00	1.67	17.92	0.0	17.68	0.00
	Maximum	33.20	30.75	33.20	0.0	30.50	0.00
	St Dev	3.11	11.40	1.43	0.0	2.66	0.00
Gen Warm Air Temp	Mean	56.70	10.50	63.16	0.0	30.53	0.00
	Minimum	0.00	1.61	24.96	0.0	23.80	0.00
	Maximum	67.60	20.07	67.60	0.0	47.00	0.00
	St Dev	14.02	8.02	4.64	0.0	4.46	0.00
Shaft Vibration Gen Journal Bearing Front 1	Mean	21.27	6.40	26.30	0.0	0.50	0.00
	Minimum	-1.10	-0.62	16.81	0.0	-1.10	0.00
	Maximum	32.43	25.77	32.43	0.0	15.77	0.00
	St Dev	10.29	11.94	0.63	0.0	2.13	0.00
Shaft Vibration Gen Journal Bearing Front 2	Mean	16.50	4.79	20.37	0.00	0.55	0.00
	Minimum	-1.62	-1.01	12.85	0.00	-1.62	0.00
	Maximum	25.23	19.73	25.23	0.00	15.03	0.00
	St Dev	7.98	9.25	0.67	0.0	2.81	0.00
Shaft Vibration Gen Journal Bearing Rear 1	Mean	16.50	4.79	20.37	0.0	0.55	0.00
	Minimum	-1.62	-1.01	12.85	0.0	-1.62	0.00
	Maximum	25.22	19.73	25.23	0.0	15.03	0.00
	St Dev	7.98	9.25	0.67	0.0	2.81	0.00
Shaft Vibration Gen Journal Bearing Rear 2	Mean	14.54	5.70	17.05	0.0	4.29	0.00
	Minimum	0.0	0.24	11.88	0.0	3.59	0.00
	Maximum	18.77	18.28	18.77	0.0	11.93	0.00
	St Dev	5.16	7.81	0.82	0.0	0.95	0.00

Analyzing the tables that show the mean, minimum and maximum values for each cluster as well as the dataset, it is possible to see that cluster 0 contains the higher values for each one of the parameters, which reflects that cluster 0 represents the behavior of the generator well functioning state. On the contrary, cluster 1, cluster 2 and cluster 3 metrics contain values well below average, representing error functioning states.

For each feature, a graph showing the distribution of the feature values across the dates assigned to the correspondent cluster was performed. It is possible to see that cluster 0 corresponds to the generator normal functioning, while cluster 1, cluster 2 and cluster 3 correspond to some abnormal generator functioning state. An example can be seen in Figure 4.

It is possible to visualize that the majority of the data values belong to cluster 0 in all graphs. The distribution of the values can differ between the graphs. The clusters' data values can be scattered between the graphs, not obeying an equal distribution.

Figure 4 shows the distribution of steam flow feature where each value matches a correspondent cluster. The active power feature shows an equal distribution given that these features are

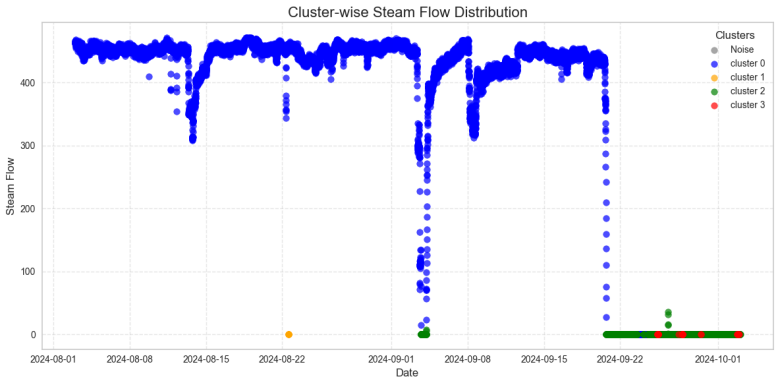


Figure 4: Steam Flow Clusters distribution

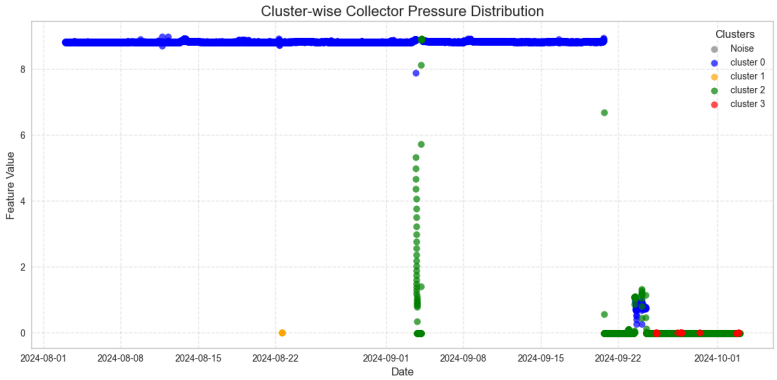


Figure 5: Collector Pressure Clusters distribution

highly correlated. Figure 5 shows the same distribution applied to collector pressure feature. Figure 6 shows the same distribution applied to steam temperature feature. Figure 7 shows the same distribution applied to current L1 feature. Figure 8 shows the same distribution applied to Shaft Vibration Generator Journal Bearing Front 1 feature.

Through graphical representation it is possible to see the high correlation between current L1 and shaft vibration generator journal bearing feature, when the current presents normal operating values, shaft vibration generator journal bearing feature matches normal functioning, otherwise it is also reflected in the behavior of this variable. This behavior reflects what would be expected from the correlation matrix analysis, where the correlation between these features lies in an interval between 0.92 and 0.96.

The following graphs describe the relation between some features whose relationship may be more relevant to be the target of analysis.

Figure 9 plots steam temperature as x-axis and current l1 as y-axis. It shows that those variables are highly correlated, which is according to the correlation matrix (0.92 correlation). When steam temperature reaches its peak functioning is when current L1 begins to register a progressive increase, until reaching its full operating value. From 2024-09-23 11:35 to 2024-09-24 02:40, both variables registered a slight increase, similar to other features. It can be noticed in the graph, as the cluster 0 isolated data points that stand out from the cluster 1 neighboring points. This change in the value of the variables was probably due to a controlled and expected

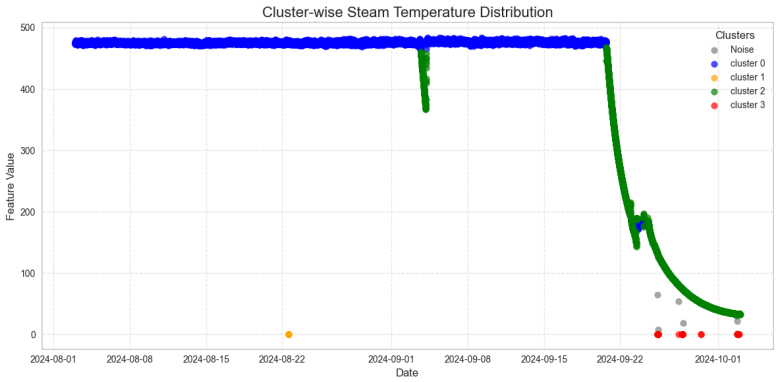


Figure 6: Steam Temperature Clusters distribution

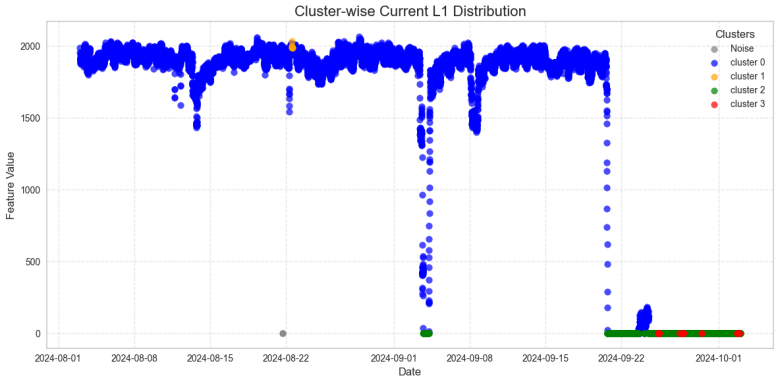


Figure 7: Current L1 Clusters distribution

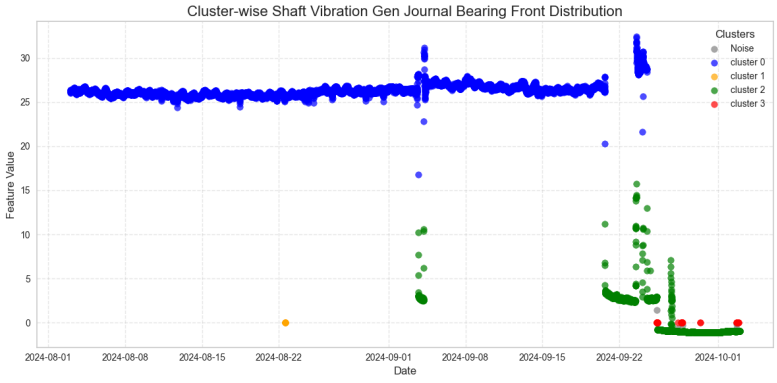


Figure 8: Shaft Vibration Generator Journal Bearing Front 1 Clusters distribution

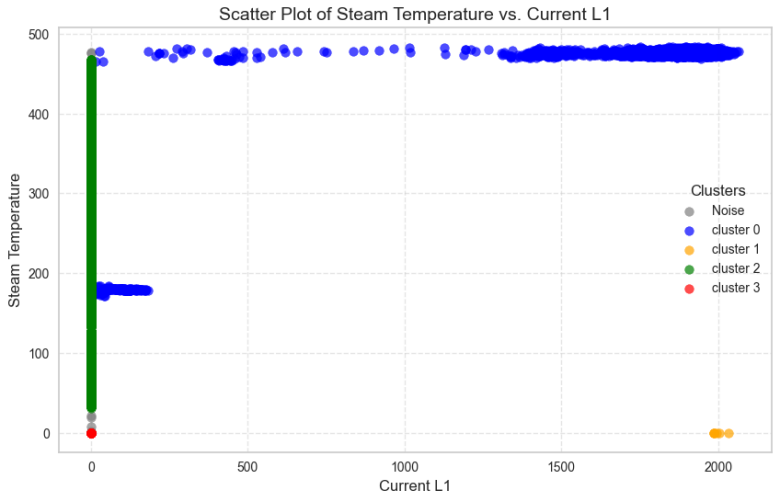


Figure 9: Steam Temperature vs Current L1 Features scatter plot distribution

load peak.

Figure 10 shows shaft vibration generator journal bearing front 1 as x-axis and current L1 as y-axis. It follows the same pattern of the current L1 and steam temperature plot features, where current L1 begins to register a progressive increase until reaching its full operating value, when shaft vibration generator journal bearing front 1 reaches a high functioning operating value. One remarkable pattern is represented by cluster 1 data points, where for current L1 a peak production value is reached, around 2000 A, contrary to shaft vibration generator journal bearing front 1, where the value for those data points is nullable. This anomaly is relative to 2024-08-22, between 14:10 and 14:30, where some features registered a sudden drop, in this case shaft vibration.

Figure 11 shows the scatter plot between the variables steam temperature and shaft vibration generator journal bearing front. It is possible to observe the existence of activity peaks of shaft vibration generator journal bearing front when steam temperature reaches values around 90 ^{o}C, 180 ^{o}C and 480 ^{o}C.

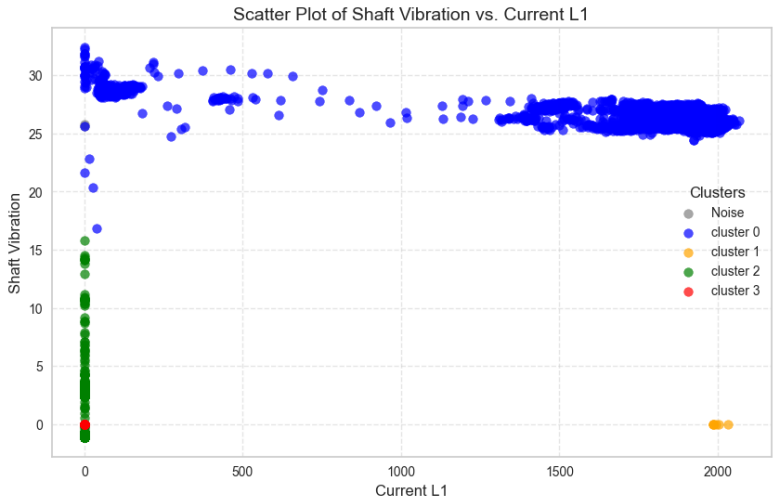


Figure 10: Shaft Vibration Generator Journal Bearing Front vs Current L1

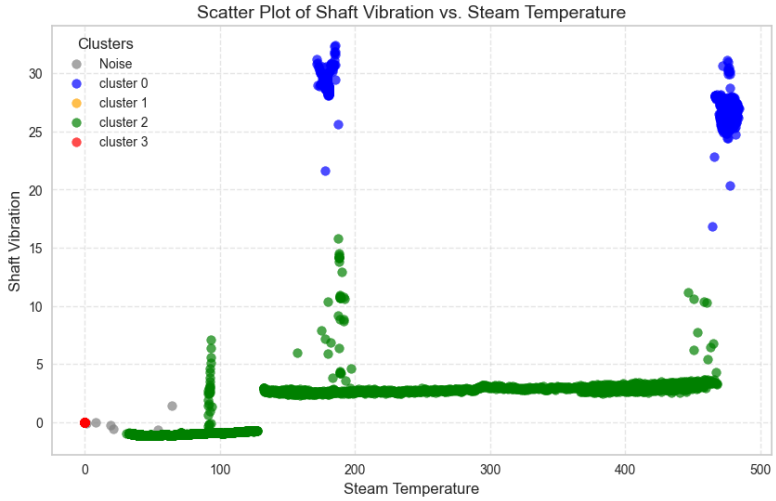


Figure 11: Steam Temperature vs Shaft Vibration generator journal bearing front

Analyzing the graphs with more detail it is possible to see that the first identified major abnormal state occurred in 2024-08-21 and 2024-08-22. A sudden anomaly visible in the behavior of the features Current L1, Current L2, Current L3, Power Factor, Bus Voltage and Frequency registered a decrease, between 17:50 and 18:00, on 2024-08-21. Those points were not assigned to any cluster, being therefore marked as noise. In the following day, 2024-08-22, another anomaly occurred around 13:35 and the periods between 14:10 and 14:30, where the majority of the features not affected by the anomaly occurred in 2024-08-21 showed unusual patterns, registering a sudden decrease, passing from recording normal functioning values to recording null values, these data points were assigned to cluster 1. After this period they return to normal operation.

A second identified abnormal state affecting the majority of the features is possible to visualize between the periods 2024-09-03 16:45 and 2024-09-04 05:10, where a progressive drop on the

normal values of some features occurred. Those values are assigned to cluster 2 or marked as noise.

A third identified abnormal state is possible to visualize since 2024-10-20 approximately at 13:30, when the values of some features started to decrease progressively, until day 2024-10-03. Since day 2024-10-20 approximately at 13:30 until day 2024-10-26, the registered data points were assigned to cluster 2. From day 2024-10-26 on, these values register a slight increase, which could indicate a return to normal values. Those values were assigned to cluster 0. Then there was a consistent decrease until day 2024-10-03, where the majority of the data points were assigned to cluster 2 too and others to cluster 3.

Cluster 1 data points represents the short and sudden values downfall occured in 2024-08-22. Cluster 2 data points correspond to progressive and longer decreases in generator functioning that indicate some type of abnormal situation. Cluster 3 indicates some sudden and short downfall in data values, during the progressive decrease registered from 2024-10-26 to 2024-10-03. And finally, data points that were not assigned to any cluster were marked as noise.

5 Discussion

The present research focused on analyzing a data set from a biomass electrical generator. PCA was used to reduce the data set dimensionality, keeping most of the data's variance. DBSCAN was used in order to determine clusters of samples corresponding to the generator's working states. PCA and DBSCAN were used to solve a practical problem and determine data points where the generator failed or worked in abnormal mode.

In the state of the art, many articles describe condition monitoring techniques in a general way or applied to a specific problem, such as the article for grouping earthquake area [7] or the approach for anomalous sound detection [10]. To the best of the authors' knowledge, this is the first time PCA and DBSCAN were applied in order to identify a biomass generator's functioning states. With this combined implementation, it was possible to see that some clustering states were distinguished, allowing to differentiate between states that can belong to normal or anomalous functioning.

6 Conclusion

Biomass electrical generators are important industrial equipment, whose failure can endanger people and property and cause enormous economic losses, or even loss of life. The data analysis method used in the present work was adequate to automatically determine the generator's working states, differentiating normal and anomalous functioning. The research offers the possibility to explore a practical scenario from real condition data where the generator functioning state can be analyzed through application of some machine learning algorithms, namely PCA and DBSCAN, which proved once more to be effective for clustering type problems.

In future work, classification methods will be used, in order to automatically determine the state of the generator.

References

[1] Marek Molęda, Bożena Małysiak-Mrozek, Weiping Ding, Vaidy Sunderam, and Dariusz Mrozek. From corrective to predictive maintenance—a review of maintenance approaches for the power industry. *Sensors*, 23(13), 2023.

[2] Denis Ilić and Mileta Žarković. Machine learning for power generator condition assessment. *Electrical Engineering*, 106(3):2691–2703, 2023.

[3] Shewei Wang, Kesheng Wang, and Zhe Li. A review on data-driven predictive maintenance approach for hydro turbines/generators, 2016/11.

[4] Chetan Shetty, S Seema, BJ Sowmya, Rajesh Nandalike, S Supreeth, Dayananda P, Rajeev Ranjan, and Venugopal Goud. A machine learning approach for environmental assessment on air quality and mitigation strategy. *Journal of Engineering*, 2024(1):2893021, 2024.

[5] L. Hao, P. L. Lewin, J. A. Hunter, D. J. Swaffield, A. Contin, C. Walton, and M. Michel. Discrimination of multiple pd sources using wavelet decomposition and principal component analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 18(5):1702–1711, 2011.

[6] Xuan Li, Ying Xu, Fengmo Li, Qiao Kang, Zhuling Li, and Cheng Long. The equivalent model for PMSG offshore wind farm based on clustering algorithm. In *2021 6th Asia Conference on Power and Electrical Engineering (ACPEE)*, pages 761–766, 2021.

[7] Mustakim, Emi Rahmi, Mediantiwi Rahmawita Mundzir, Said Thaufik Rizaldi, Okfalisa, and Idria Maita. Comparison of DBSCAN and PCA-DBSCAN algorithm for grouping earthquake area. In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, pages 1–5, 2021.

[8] I. K. Savvas, A. V. Chernov, M. A. Butakova, and C. Chaikalis. Increasing the quality and performance of n-dimensional point anomaly detection in traffic using PCA and DBSCAN. In *2018 26th Telecommunications Forum (TELFOR)*, pages 1–4, 2018.

[9] Liu Ni and Si Jinhang. The analysis and research of clustering algorithm based on pca. In *2017 13th IEEE International Conference on Electronic Measurement Instruments (ICEMI)*, pages 361–365, 2017.

[10] Yiu S.M Tan X. Self-adaptive incremental pca-based DBSCAN of acoustic features for anomalous sound detection. *SN COMPUT. SCI.*, pages 5,542, 2024.

[11] Abdullah Akram Bajwa, Salman Habib, and Muhammad Kamran. Wavelet energy distribution with PCA DBSCAN for partial discharge pulse extraction. In *17th IEEE International Multi Topic Conference 2014*, pages 422–427, 2014.

[12] A Supriyanto S Wibisono, M T Anwar and I H A Amin. Multivariate weather anomaly detection using DBSCAN clustering algorithm. *Journal of Physics: Conference Series*, 1869, 2021.

[13] Joey Li, Munur Sacit Herdem, Jatin Nathwani, and John Z. Wen. Methods and applications for artificial intelligence, big data, internet of things, and blockchain in smart energy management. *Energy and AI*, 11:100208, 2023.

[14] C. K. Raj. Comparison of K Means, K Medoids, DBSCAN algorithms using DNA microarray dataset. *International Journal of Computational and Applied Mathematics*, 12:344–355, 2017. Accessed: 2024-12-15.

Chapter 6

The Transformative Role of FPGA Platforms in Accelerating AI tasks

Contribution #26
Cite as:
Gulfarida Tulemissova, Shynar Yelezhanova, Almiri Karkimbayeva. The Transformative Role of FPGA Platforms in Accelerating AI tasks. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-26,
  title ={The Transformative Role of FPGA Platforms in
     Accelerating AI tasks},
  author={Gulfarida Tulemissova and  Shynar Yelezhanova
     and  Almiri Karkimbayeva},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

The Transformative Role of FPGA Platforms in Accelerating AI tasks

Gulfarida Tulemissova[1*], Shynar Yelezhanova[2], Almiri Karkimbayeva[3]

[1*]Gulfarida Tulemissova, SDU University, 1/1 Abylaikhan Street, Kaskelen, Almaty
[2] Shynar Yelezhanova, Kh. Atyrau University, 1 b Student Avenue, Atyrau
[3]Almiri Karkimbayeva, Aubakirova School, 140 Karasay batyra, Almaty
*Gulfarida Tulemissova
gulfarida.tulemissova@sdu.edu.kz

Abstract

FPGAs continue to be one of the key hardware solutions for real-time and low-power AI applications. With fine-grained parallelism control, deep pipelining, and adaptive precision, they are a great complement to GPUs and ASICs, especially in systems with strict latency and power consumption requirements. At the same time, the percentage of GPU processors worldwide is growing compared to the percentage of FPGA platforms. Therefore, the research question of the article is "Whether FPGA platforms really have a limited scope of use for processing AI applications?"

The goal of the research is to confirm the facts of the importance of using FPGA platforms for AI tasks.

This paper focuses on important platform characteristics, such as FPGA sources of parallelism due to the architecture of the hardware itself, support for AI infrastructure. Such characteristics as performance, throughput, energy efficiency, and latency are compared with the characteristics of other digital processors such as GPU, CPU, TPU, ASIC. The study revealed the fact of lower FPGA performance compared to GPU, but emphasized the need for further development of Xilinx Vitis AI (DPU), FINN & Brevitas, Intel OpenVINO for FPGA, HLS (C/C++) tools in the future.

The results of the research had been obtain through scientific analysis of the FPGA platform features, its advantages and limitations, conditions of use and their impact on the acceleration of AI tasks.

Had been confirm of the fact of the need to use FPGA platforms in the future to solve AI problems and further development of training specialists in the field of FPGA platform programming. The research results are important for designers using hardware and technologies with increased performance. in areas requiring accurate, high-speed processing and low latency operations.

Keywords: FPGA, GPU, AI infrastructure, processor accelerator, parallel technologies, hardware neural network, energy efficiency, low latency.

1 Introduction
Current state of using digital devices for AI tasks

Artificial intelligence is leading in technological innovation, accelerating the development of many areas of human activity. Graphics processing units (GPUs) and specialized tensor processing units (TPUs) traditionally has been use to implement AI algorithms, which provide high performance due to parallel computing. Accordingly, the AI infrastructure places high demands on the technical hardware of their implementations, which depend on the type of AI, its tasks and level of complexity. First, this is the computing power of GPU graphics processors, such as NVIDIA CUDA, AMD ROCm [1, 2]. In addition to GPUs, FPGA and ASIC platforms had used for AI applications for specific AI tasks require high-energy efficiency and optimization for specific algorithms. They have used to train large language models (LLM), such as GPT, which require clusters with hundreds or thousands of GPUs / TPUs, providing performance of tens of petaflops.

However, the high computational load of AI applications poses serious challenges. Traditional processors (CPUs) and graphics processing units (GPUs), despite their power, often fail to cope with these tasks effectively. This is where field-programmable gate arrays (FPGAs) come to the rescue. FPGAs are transforming AI acceleration by offering a flexible, efficient, and high-performance solution for artificial intelligence tasks. In addition, FPGAs have unique potential for AI tasks, including energy efficiency, flexibility of hardware configuration, and the ability to optimize for specific algorithms [3]. Some authors believe that despite these advantages of FPGA platforms, they has been under use in AI tasks due to a number of limitations. The complexity of project development and incomplete integration into popular AI frameworks (TensorFlow, PyTorch) are commonly mentioned [4]. Most software engineers often focus on GPU processors, completely ignoring the capabilities of FPGA platforms. There is a situation and opinion that FPGA platforms have some limitations for solving AI problems and even the percentage of trained specialists for this industry is falling [5, 6]. The incomplete use of FPGA in AI is due to both technical and organizational factors. On the one hand, developing for FPGA requires deep knowledge in the field of hardware programming, which limits the circle of specialists. On the other hand, there are no standardized tools and libraries to simplify the integration of FPGAs into AI ecosystems. This article examines the reasons for the incomplete use of FPGA and tests the hypothesis - do FPGA platforms really have a limited scope of use for processing AI applications?

Accordingly, the objectives of the study are:

• State-of-the-art analysis: Study the current approaches to using FPGAs in AI tasks, including a review of existing tools, frameworks, and libraries.

• Limitations identification: Identify key technical and organizational barriers that hinder the widespread use of FPGAs in AI, such as programming complexity, performance, and compatibility with frameworks.

• Processor benchmarking: Compare the performance of FPGAs with GPUs and TPUs on typical AI tasks (e.g., convolutional neural networks).
• Prospect assessment: Assess the potential for using FPGAs in edge computing and other energy-efficient AI scenarios
• Conclusions and development of recommendations: and suggest recommendations for further research.
This research is necessary to revive the idea of the uniqueness of FPGA platforms, to implement industrial laboratory work in the training of engineers in this area, since such projects with AI applications are in demand for aerospace, defense, telecommunications, and other areas.
The relevance of FPGAs is often due to the efficiency of programming logic blocks and connections that could be configure to perform specialized hardware tasks. This flexibility allows developers to create highly efficient solutions that has been precisely tailor to specific computing needs. It is because of the ability to reprogram the hardware structure for specific tasks that FPGA platforms are ideal for both neural networks and deep learning applications. FPGAs could been reconfigure with new logic circuits to adapt to new AI algorithms or optimize existing ones, which is critical as AI models and standards continue to evolve.
The content of the paper presents in the following parts:
• Background
• Analysis of the current state
• Existing limitations
• Comparative analysis of FPGA, GPU, TPU
• Results
• Discussion
• Conclusion
This paper does not pretend to be a detailed review of the problem due to the limited volume of material required by the conference.

Background

There is an opinion that FGPA platforms could not been use as often as GPU processors for AI tasks. Opponents mainly note two main reasons for the limited use of FGPA platforms. The first reason is that FPGA programming requires a deep understanding of the hardware, which narrows the circle of qualified specialists. The second reason is the lack of standardized tools and libraries that could facilitate the integration of FPGAs into artificial intelligence systems [7]. It is also be known that GPUs have traditionally been popular for their powerful parallel processing capabilities, which are well suited for training deep learning models. It should be note that FPGA platforms have some advantages in certain criteria for AI:
• Power efficiency
• Lower latency
• Flexibility and configurability
• Customizability for specific workloads
• Scalability in AI applications
• Excellent data flow management
• Reduced hardware redundancy
• Increased data security and privacy
• Long-term cost effectiveness
• High parallel processing for inference
• Robustness in complex environments [8].

Overview of approaches

FPGAs has been use in AI tasks due to their parallel computing capabilities, customizable architecture, and low latency. Key approaches include:
• Deep neural network (DNN) acceleration: FPGAs efficiently implement convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers [1, 2]. They provide optimization of matrix multiplication and convolution operations due to hardware tuning for specific models. For example, FPGAs can perform calculations with reduced precision (e.g., INT8 or FP16), which reduces power consumption and increases performance [3, 4].
• Real-time data processing: FPGAs has be use in computer vision, natural language processing (NLP), and autonomous systems where low latency is required. This includes video stream processing in self-driving cars or sensor data analysis in robotics [7, 8].
• Quantization and model optimization: To improve efficiency, FPGAs use quantization (translating models into a lower-precision format) and network compression (pruning). This reduces memory and computing resource requirements while maintaining accuracy [1, 9, 10, 20].
• Hybrid architectures: FPGAs are often integrate with processors (CPUs), graphics processing units (GPUs), or tensor processing units (TPUs) in heterogeneous systems. For example, embedded FPGA processors (like ARM in SoCs) allow for the combination of control and AI computations [12, 13],
• Energy-efficient solutions for edge devices (Edge AI)x[2]: FPGAs are widely used in Internet of Things (IoT) devices and edge computing, where power consumption and compactness low are important. They provide local data processing, minimizing the need for transfer to the cloud [16, 18, 19].

• Adaptive computing: FPGAs support dynamic reconfiguration, which allows the hardware architecture to adapt to changing AI tasks, for example, to update models without replacing hardware [21].

The following development tools usually use for these approaches: hardware description languages (HDL, such as VHDL and Verilog), high-level synthesis tools (HLS, for example, Vitis HLS), specialized platforms (Intel OpenVINO, Xilinx Vitis AI).

The relevance of FPGAs is often due to the efficiency of programming logic blocks and connections that can be configure to perform specialized hardware tasks. This flexibility allows developers to create highly efficient solutions that are precisely tailor to specific computing needs. It is the ability to reprogram the hardware structure for specific tasks makes FPGA platforms ideal for both neural networks and deep learning applications. FPGAs can be reconfigure with new logic circuits to adapt to new AI algorithms or optimize existing ones, which is critical as AI models and standards continue to evolve.

2 Analysis of the current state

There is an opinion that FGPA platforms be cannot use as often as GPU processors for AI tasks. The problem of underutilization of FPGA in AI is cause by both technical and organizational factors. Opponents note two main reasons for the limited use of FGPA platforms: FPGA programming requires a deep understanding of the hardware, which narrows the circle of qualified specialists and the lack of standardized tools and libraries that could facilitate the integration of FPGA into artificial intelligence systems [7]. It is also know that GPUs are traditionally popular for their powerful parallel processing capabilities, which are well suited for training deep learning models. However, FPGA platforms offer distinct advantages in certain AI scenarios:

- Power efficiency
- Lower latency
- Flexibility and re configurability
- Customizability for specific workloads
- Scalability in AI applications
- Superior data flow management
- Reduced hardware redundancy
- Enhanced data security and privacy
- Long-term cost effectiveness
- High level parallel processing
- Robustness in complex environments [8].

FPGAs has been use in AI tasks due to the following cases:

Acceleration of deep neural networks (DNN): FPGAs efficiently implement convolutional neural networks (CNN) and also recurrent neural networks (RNN), and transformers. They provide optimization of matrix multiplication and convolution operations through hardware tuning for specific models [9]. Real-time data processing: FPGAs are used in computer vision, natural language processing (NLP), and autonomous systems where low latency is required [10, 11].

Quantization and optimization of models: To improve efficiency, FPGAs employ quantization (translating models into a format with reduced precision) and network compression (pruning). This reduces memory and computational requirements while maintaining accuracy [12].

Hybrid architectures: FPGAs are often integrate with processors (CPUs), graphics processing units (GPUs), or tensor processing units (TPUs) in heterogeneous systems. For example, embedded FPGA processors (like ARM in SoCs) allow for the combination of control and AI computation.

Energy-efficient solutions for edge devices (Edge AI): FPGAs are widely used in Internet of Things (IoT) devices and edge computing, where power consumption and compactness low are important. They enable local data processing, minimizing the need to transfer data to the cloud [13].

Adaptive computing: FPGAs support dynamic reconfiguration, which allows the hardware architecture to adapt to changing AI tasks, for example, to update models without replacing hardware [14].

FPGA Programming

First, it is important to note that any field of instrumentation science requires deep knowledge. In addition, this applies to classical programming in high-level languages. The most criticism goes to the design of digital devices on FPGAs and hardware description languages HDL, which are very difficult to learn. The author's experience of programming in high-level languages and comparing them with HDL languages gives the right to disagree with this statement. Hardware description languages (HDL), such as VHDL, Verilog or System Verilog [15], have good qualities such as:

• Hardware abstraction: HDL allows you to describe complex hardware systems at a high level, simplifying the design and modeling of digital circuits without the need to work with physical components.

• Modeling and verification: HDL supports the simulation of circuit operation before their physical implementation, which allows you to identify errors at the design stage, saving time and resources. • Hardware synthesis: HDL code can be automatically converted into circuits (synthesized) for implementation on FPGA, ASIC or other hardware platforms.

• Modularity and reusability: HDLs allow for modular designs where components could been reuse across designs, simplifying development and maintenance.

• Parallelism: HDLs naturally describe parallel processes, which corresponds to the real operation of hardware systems where many operations had been perform simultaneously.
• Accuracy and detail: HDLs provide an accurate description of timing characteristics (e.g., latencies, clock signals), which is critical for the design of high-performance systems.
• Wide tool support: Many HDL- tools (ModelSim, Vivado, Quartus) support simulation, synthesis, and analysis [16, 17].
• Versatility: HDLs has been use to design a wide range of systems, from simple logic circuits to complex processors and SoCs (System on Chip).
• Documentation: HDL code also serves as documentation, which simplifies project maintenance and transfer.
• Standards support: Languages such as VHDL and Verilog are standardized (e.g. IEEE), which ensures compatibility and portability of projects between different platforms and tools. These advantages make HDL indispensable tools in the development of digital electronics.
Designing digital equipment for efficient computing on FPGA has a number of advantages that show the logical connection of digital device design with logic algebra, signal process theory, and mathematical problem statement. Such a powerful logical connection in the design of digital devices forms logical thinking in students when faced with real-life problems. Hardware description languages are perfect and you can get a circuit code that is much smaller than the desired physical circuit. It is possible to reduce cumbersome design schemes due to the Karnaugh method. Moreover, the automated design system forms a circuit for its chip without factory technologies. Therefore, training specialists to design digital devices on the platform is a large school of digital technologies, where you can see the process from step zero to obtaining the result. Such industries as aerospace, where FPGAs are indispensable for processing signals from sensors in real time for radars, lidars, inertial systems and cameras. FPGAs are base for working with gyroscopes and accelerometers in satellite or drone control systems, for satellites during encoding/decoding, modulation, signal filtering and processing satellite communication protocols, for processing images from cameras installed on satellites, drones or rovers, image compression, object detection. FPGAs implement control algorithms for engines, stabilizers, autopilots and orientation systems of spacecraft, in processing images of the Earth in remote sensing satellites. FPGAs has been use in systems that require high reliability, such as control of booster blocks, emergency shutdown systems or on-board computers [18]. These are radiation-hardened FPGAs [Xilinx Virtex-5QV, Microchip RTG4] [17], developed specifically for space. FPGAs are used to prototype new processors, controllers or SoCs (System on Chip) before they are manufactured as ASICs. FPGAs allow the creation of specialized hardware accelerators for tasks such as cryptography, machine learning or telemetry processing. Now, in addition to these specific FPGA tasks, AI processing is being added - implementing machine learning algorithms to analyze data on board a satellite. The FPGA platform is becoming a universal device! Specific project examples:
• Satellites: FPGAs control orientation systems, camera data processing and communications (e.g. in Starlink satellites).
• Mars rovers: FPGAs process sensor data and control movement (e.g. in NASA's Perseverance rovers).
• Rockets: FPGAs are used in guidance and engine control systems (e.g. in SpaceX rockets).
• UAVs: FPGAs implement image processing and autonomous control in drones [18].

Parallelism in FPGAs

I would especially like to emphasize the main characteristic of FPGAs - task parallelism, which divided into several types - Data Parallelism, SIMD-like mode, Pipeline Parallelism, Spatial Parallelism, and Task Parallelism. All these types of platform parallelism support HDL languages, which significantly increases the performance and flexibility of the designed systems. Processing multiple operations simultaneously is essential for matrix and vector calculations that underlie machine learning and neural networks. FPGAs can manage threads to reduce latency and increase throughput, which is critical for AI applications that require real-time processing, such as video analysis and autonomous driving systems. Therefore, FPGAs can perform calculations faster and more efficiently than other processors, which leads to significant energy savings. Therefore, we can say that FPGA parallelism provides high performance: parallel blocks operate independently. Example: parallel execution of multiple neural layers in AI accelerators. All these features affect the scalability of tasks.
General advantages of parallelism on FPGA:
• Hardware optimization: FPGAs provide low-level control over the implementation of parallelism, which allows achieving high energy efficiency and performance.
• Configurability: FPGAs can be adapted to a specific type of parallelism by changing the architecture without replacing the hardware.
• Low latency: Parallel computing on FPGAs minimizes delays, which is critical for real-time applications.
• Support for complex applications: A combination of different types of parallelism (for example, pipelined and spatial) allows solving complex problems such as AI processing, telecommunications, or cryptography [19].

FPGA accelerators

With processor performance being an important metric for AI, it is important to note that there are FPGA platforms called accelerators [20]. FPGA accelerators can significantly increase the speed of certain computations, such as convolutional operations in deep learning networks, by optimizing hardware paths specifically for these operations. FPGA accelerators offer several key benefits:
• Speed: They can process tasks faster than software running on general-purpose processors by performing multiple operations in parallel.

• Customization: They could been customize to accelerate specific parts of a machine-learning algorithm, improving performance where it has been need most.
• And energy efficiency.

Integration into AI workflows

FPGAs can be integrate into both the machine learning and inference stages of AI workflows. FPGAs can accelerate certain tasks during the training stage, such as forward and backward propagation, handle intensive computations more efficiently, and can process data in real time with low latency. They are especially useful in environments where decisions must made quickly and where deploying large GPU systems is impractical [21].

3 Existing Limitations

There is a problem with energy efficiency. This is when AI algorithms are not adapted to the FPGA architecture, and then a significant part of their computing resources may be idle. There is an underload for non-specific AI algorithms, when energy consumption can be higher than that of specialized ASICs or even GPUs, due to inefficient resource allocation. In addition, in cases of underutilization of memory (BRAM) or logic blocks, there may be excess energy consumption, when unused elements still consume energy in standby mode.
The next limitation is when it is difficult to support standard AI frameworks such as TensorFlow or PyTorch, which had been better optimize for GPUs and CPUs. Partial use of FPGA often requires additional efforts to integrate with these frameworks, which can lead to reduced performance or the need for additional software layers such as Vitis or OpenVINO. Without full customization for specific AI models (e.g., convolutional neural networks or transformers), FPGAs may be less efficient than dedicated solutions. Partial use of FPGAs may lead to problems with scaling AI solutions, especially for tasks that require processing large amounts of data (e.g., training large language models). Partial use of FPGAs limits the ability to scale compared to GPU clusters.
Moreover, underutilization of high-speed input/output (I/O) interfaces or networks-on-chip (NoC) reduces the ability of FPGAs efficiently process in real-time data streams.
Some FPGAs support partial reconfiguration, which allows updating individual parts of the chip without stopping others. However, if improperly configured, partial reconfiguration may lead to timing errors or performance degradation. Underutilization of this capability limits the benefits of FPGAs in dynamically changing AI applications [22].

4 Comparative analysis of FPGA and GPU

In this research, a system method had been use in descriptive and comparative aspects. The current state of hardware development had been determine by the independence of FPGA development from CPU, GPU and TPU in terms of its functioning as a whole and according to its inherent internal laws. To analyze the characteristics and understand the processes of all the above-mentioned devices, a multi-criteria comparative analysis method had been use according to certain criteria listed above. Below is a comparative table 1 by key criteria. A comparative analysis of the characteristics of digital devices had been carry out according to various criteria in order to determine the range of their use for artificial intelligence tasks. The main attention had been pay to GPU and FPGA devices. All data had been summarize in several tables. Below is comparative table for FPGA, ASIC, GPU, CPU, and TPU, focusing on measurable units where qualitative assessments for non-quantifiable criteria are.

Table 1. Comparative table of characteristics of digital devices

N	**Criterion**	**FPGA**	**ASIC**	**GPU**	**CPU**	**TPU**
1	**Core Functionality**	CNN training, Edge computing, Telecom tasks	Fixed-function hardware for specific tasks	For AI tasks, HPC, Gaming	General-purpose computing	Tensor operations for AI/ML
2	**Performance**	10-50 TFLOPS (e.g., Xilinx Versal ACAP)	100-500 TFLOPS (e.g., custom AI ASIC)	100-312 TFLOPS (e.g., NVIDIA H200)	1-5 TFLOPS (e.g., AMD EPYC 9004)	400-800 TFLOPS (e.g., Google TPU v5)
3	**Power Consumption**	50-200 W	10-100 W	300-700 W	100-400 W	50-150 W
4	**Parallel Processing**	~1,000-10,000 logic units	~10,000-100,000 equivalent cores	~10,000-100,000 CUDA cores	8-128 cores	~10,000-100,000 matrix units
5	**Power Efficiency**	0.05-0.25	1-5	0.1-0.5	0.01-0.05	2-5

		TFLOPS/W	TFLOPS/W	TFLOPS/W	TFLOPS/W	TFLOPS/W
6	**System Latency**	1-10 µs	0.1-1 µs	10-100 µs	100-1,000 µs	0.5-5 µs
7	**Flexibility and Re Configurability**	Very High (reprogrammable hardware)	Very Low (fixed post-fabrication)	Moderate (shader programming)	High (general-purpose software)	Low (AI-specific design)
8	**Customizability for Specific Workloads**	High (custom logic design)	Very High (task-specific fabrication)	Moderate (optimized for parallel tasks)	Low (general-purpose)	High (AI/ML-specific)
9	**Scalability in AI Applications**	High (modular, but redesign needed)	Very High (optimized for scale)	Very High (data center integration)	Moderate (core-limited)	Very High (cloud-optimized)
10	**Superior Data Flow Management**	High (custom data paths)	Very High (optimized pipelines)	High (parallel memory access)	Moderate (general-purpose)	Very High (tensor pipelines)
11	**Reduced Hardware Redundancy**	High (reconfigurable resources)	Very High (no unused components)	Moderate (general-purpose cores)	Low (redundant for AI tasks)	Very High (minimal overhead)
12	**Enhanced Data Security and Privacy**	High (custom security protocols)	Very High (hardware-level isolation)	Moderate (software-based security)	Moderate (software-based)	High (hardware security modules)
13	**Long-term Cost Effectiveness**	Moderate ($10K-$100K/unit, reusable)	High ($1M-$10M dev, low per-unit at scale)	Moderate ($10K-$50K/unit)	High ($500-$10K/unit, versatile)	High (cloud pricing, ~$1-$2/hour)
14	**Robustness in Complex Environments**	High (adaptable to diverse tasks)	Very High (reliable for specific tasks)	Moderate (less flexible)	High (broad compatibility)	High (reliable for AI tasks)

The peculiarity of FPGA Accelerator and GPU Accelerator architectures are different, but FPGAs are efficient in implementing CNN inference, especially for optimized models. They allow creating hardware-accelerated kernels for convolutions, minimizing latencies. While GPUs use SIMD (Single Instruction, Multiple Data) architecture and support platforms such as CUDA (NVIDIA). GPUs excel in CNN training due to their high computing power and optimized libraries (e.g., cuDNN). They are also efficient for inference, but less flexible in customization. Below is table 2 of the physical parameters of digital accelerators.

Table 2. Architecture and physical parameters of FPGA and GPU

N	Device type	Size	Memory	Architecture
1	**FPGA Accelerator**	45 x 32 mm, 52.5 x 42.5 mm, 56 x 45 mm	190 – 287 MGB	Programmable Hardware Accelerated Kernels for CNN Convolutions
2	**GPU Accelerator**	267 x 110 мм	80 GB	SIMD (Multiple Data)

Performance GPUs have high computing power (see Table 1). FPGAs also have high performance for specific tasks due to the ability to implement custom compute units. However, for large models (e.g. with billions of parameters, like GPT-3, they are inferior to GPUs due to resource limitations [26]. In terms of energy efficiency, GPUs consume significant power (up to 400 W), which makes them less energy efficient than FPGAs. Power consumption becomes an issue in large data centers. GPUs require powerful cooling systems, which increases the overall cost of operation [27]. FPGAs are much more energy efficient. FPGAs are also flexible in configurability and programmability. FPGAs are suitable for custom AI tasks such as optimized convolutions or real-time data processing. GPUs are easy to program due to mature ecosystems (CUDA, TensorFlow, PyTorch) [28].

GPUs had been well integrate with popular AI frameworks (TensorFlow, PyTorch, cuDNN). However, FPGA has limited support for them and requires additional tools, such as Xilinx Vitis or Intel OpenVINO. The cost of such devices is important, but today mid-range FPGAs are becoming more affordable for automotive, telecommunications, and IoT).

Modern GPUs are expensive, but they are available through cloud services (AWS, Google Cloud) and are widely used in AI due to the support of cloud platforms and libraries.

Below is a comparative table 3 by toolkit: frameworks, libraries for working with FPGA and GPU (available web sources such as Habr, FPGA-Systems had been use). It is important that FPGA is suitable for real-time tasks [30].

Table 3. FPGA and GPU toolkit: core, frameworks, libraries

N	**Tools**	**FPGA**	**GPU**
1	Main tools	Vivado HLS (Xilinx): High-level synthesis to simplify FPGA development. - Intel FPGA SDK for OpenCL: Enables OpenCL to program FPGAs. - Quartus Prime (Intel/Altera): FPGA design and synthesis tool. - SDAccel/SDSoC (Xilinx): Frameworks for FPGA integration in machine learning and data science tasks.	CUDA (NVIDIA): The primary tool for programming GPUs, optimized for parallel computing. - ROCm (AMD): An alternative to CUDA for AMD GPUs. - OpenCL: A cross-platform standard supported on both GPUs and FPGAs. - cuDNN: NVIDIA's library for accelerating deep learning
2	Frameworks	TensorFlow/PyTorch (with limitations): Support via OpenCL integration or specialized libraries (e.g. Xilinx DNNDK). - Vitis AI (Xilinx): Framework for implementing neural networks on FPGA. - OpenVINO (Intel): Optimization of deep learning models for FPGA.	TensorFlow: A leader in machine learning supports CUDA and cuDNN. - PyTorch: Popular for research, flexible prototyping, CUDA support. - Keras: High-level API, integrates with TensorFlow to simplify running on GPUs. - Caffe: Used for computer vision tasks, optimized for GPUs.
3	Libraries	OpenCV: Computer vision support with FPGA integration. - Algo-Logic: Low-latency library for financial applications on FPGA. - Xilinx Runtime (XRT): Manages host-FPGA communication. - ONNX: Model exchange standard, supports porting to FPGA.	cuDNN: Deep learning library optimized for NVIDIA GPUs. - OpenCV: Widely used for image processing on GPUs. - NumPy/Pandas: Support for data preprocessing for machine learning. - ONNX: Support for sharing models between frameworks.

5 Results

A comparative analysis of the characteristics of digital devices had carry out according to various criteria in order to determine the range of their use for artificial intelligence tasks. The focus had been pay to GPU and FPGA devices. Answers had been receive to the question about the level of use of FPGA platforms for solving artificial intelligence problems. All data had been summarize to several tables. To begin with, we make a comparison according to Table 2, which contains the physical and architectural parameters of the device. Table 2 shows the advantages of FPGA in terms of physical size and higher memory capacity. A comparative analysis of other FPGA characteristics showed that they are highly valued in AI applications that require flexible configuration, low latency, and energy efficiency. Their programmability allows you to optimize performance for tasks such as neural network inference and real-time data processing at the edge, which allows you to adapt to changing AI requirements and reduces the need for new equipment.

According to Table 1, we can draw the following conclusions: FPGA and TPU are focused on specific tasks, ASICs are focused on specialized tasks, GPUs are focused on HPC and Gaming, and CPUs are focused on general-purpose computing. In terms of performance, GPUs and TPUs are leaders in computing power, while FPGAs are more had been narrowly focus. CPUs lag behind in this area. In terms of energy consumption, TPUs and FPGAs are energy efficient, ASICs are leaders, while GPUs consume much more. In terms of parallel processing, TPUs and ASICs provide high parallelism (up to 100,000+ units), FPGAs — up to 10,000 logical units, GPUs — up to 128 cores, and CPUs — minimum (8-128 cores). In terms of energy efficiency, ASICs and TPUs are leaders, FPGAs and GPUs are average, and CPUs are the least efficient. System latency: ASICs and TPUs provide minimal latency, FPGAs — average, and GPUs and CPUs — the highest. Flexibility and configurability: FPGAs and ASICs excel in adaptability (hardware/software), GPUs moderate in adaptability (software), CPUs low in adaptability, TPUs narrowly focused (AI-specific). Customizability for specific tasks: FPGAs and ASICs excel in customization, GPUs and CPUs are limited to general tasks, TPUs are optimized for AI. Scalability in AI: TPUs and ASICs excel in optimization, FPGAs are high (modular), GPUs are moderate, and CPUs are limited. Data management and performance: TPUs and ASICs provide excellent management, FPGAs are high (custom paths), GPUs and CPUs are moderate. Reduction in hardware redundancy: FPGAs and ASICs minimize redundancy, TPUs are highly optimized, GPUs and CPUs are lower. Security and privacy: FPGAs and ASICs offer high security (hardware), TPUs via modules, GPUs and CPUs via software. Long-term costs: ASICs (high initial, low per unit) and TPUs (cloud, $1-2/hour) are economical to operate, FPGAs and CPUs are moderate, GPUs are high. Reliability in harsh conditions: ASICs and TPUs are leaders (specialization), FPGAs (adaptability) and CPUs (broad compatibility) are high, GPUs are moderate.

Overall, based on these criteria: The choice between FPGAs, ASICs, GPUs, CPUs, and TPUs depends on the task — FPGAs and ASICs for custom solutions, TPUs for AI, GPUs for HPC, CPUs for versatility. ASICs and TPUs are optimal in efficiency but less flexible, while FPGAs and GPUs offer a balance between flexibility and performance.

Explanations and key findings for Table 3. Характеристики из Таблицы также подтверждают сравнительные вывод по таблице 1. The choice of device depends on the task, budget and the level of expertise of the team. For rapid prototyping and standard AI tasks, GPU is preferable, for custom solutions - FPGA.

The main conclusion from all comparisons is that GPU has higher performance, but FPGAs have significantly better energy efficiency, and a comprehensive assessment of nine parameters in Fig. 1 [31] shows that GPU leads in the score 5:4 in its favor, but this score could been interpret in different ways.

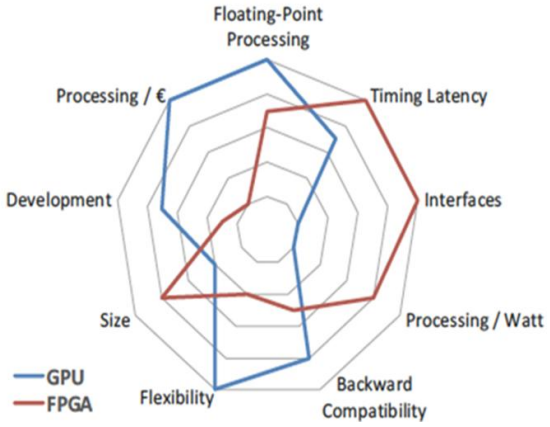


Figure 1.GPU and FPGA qualitative comparison [31]

For large models (e.g. with billions of parameters, like GPT-3), FPGAs are inferior to GPU due to resource limitations. GPUs are great at large-scale matrix operations typical for CNN training, but there is higher latency compared to FPGAs and less flexibility in optimizing for specific tasks. Power consumption becomes an issue in large data centers. GPUs require powerful cooling systems, which increases the overall cost of operation. High power consumption always affects the cost of operation.
FPGAs provide high-energy efficiency when implementing optimized models, especially for binary neural networks. This makes them ideal for IoT and mobile devices. Table 4 provides additional processor characteristics for comparison FPGA and GPU.

Table 4. Characteristic comparison of GPU and FPGA

Characteristics	**FPGA**	**GPU**
Applications	Cameras, LIDAR, autonomous vehicles, industrial equipment's, Audio sensors, Deep neural networks, Natural Language Processing etc.	Video editing, Video encoding, 3D graphic rendering, Machine learning, EDA, Image Processing and computer vision etc.
Performance	High performance	High performance
Processing speed	Fast real time processing	Extremely fast processing
Interface	No need interfacing for software	Interface for software
Programming model	Flexible	Restricted
Programming method	Vivado, Quartus	Cuda/C
Optimal processing patter	Pipeline, feedback loops	SIMD array
Cybersecurity	Good	no
Performance benefit	Throughput, latency	Throughput

An important conclusion can be drawn from Table 4 - FPGAs do not require a software interface, which simplifies their use in hardware-oriented tasks. GPUs depend on a software interface (e.g., CUDA/C), which increases flexibility but adds complexity to integration..
The enormous computing power required for AI applications poses a major challenge. Traditional devices such as processors (CPUs) and graphics processing units (GPUs), while offering significant performance, are often unable to cope with these tasks effectively. This is where field-programmable gate arrays (FPGAs) come into play. Comparative analysis has shown their transformative role in accelerating AI processes, with FPGAs providing a flexible, energy-efficient, and high-performance solution for such tasks. Modern application methods for field-programmable gate arrays (FPGAs) in the field of AI demonstrate their outstanding performance, low power consumption, and adaptability. FPGAs are widely used to accelerate neural network computations, process data in real time, and create specialized architectures optimized for AI tasks.

6 Discussion and future directions

The issue of incomplete recognition of the usefulness of FPGA arose in the process of incomplete attention to this equipment. In terms of lifetime, FPGA is much older than GPU. Even by such a common important characteristic of equipment as service life, FPGA suppliers have 15 years, while GPUs are replaced after 3-5 years due to performance requirements or obsolescence.
Most criticism was addressed to the complexity of programmability of FPGA hardware, since AI algorithms on GPUs are easier to program and the GPU has a fixed architecture. But today, vendors are improving the adaptability of FPGA for programming

infrastructure using high-level synthesis languages C / C ++ instead of HDL languages such as Verilog or VHDL for this there is support for the Open CL standard in FPGA. Here, programming means specifying neural network schemes through the use of millions of platform gates as neural network cores. If the GPU processes computation algorithms at high speed, then the FPGA is “flashed” with algorithm paths.
The GPU functionality does not work separately by itself, but is connected to the CPU coprocessor, so its communication with other parts of the system is quite limited. In addition, it needs to connect to the system RAM, the display frame buffer, and the camera. For any practical use, the GPU needs help from the CPU. At this point, the FPGA has everything on its platform in a set, which work in real time. Therefore, the speed of access to memory and the processor is much lower!
FPGA platform vendors are always working on improving them. For example, the Napatech N3070X AI/SmartNIC FPGA platform is designed for the complex requirements of AI infrastructure and is fundamental to the long-term viability of clouds and data centers. AI infrastructure creates its own requirements for data center servers and processors to support machine learning, large language models, and generative AI. Once big data is processed by AI tools, it must be transferred to its intended purpose, stored, and protected. At this point, GPUs have already finished their work and pass the baton to the networks. Cloud technologies and data centers are not quite ready for speeds. Only unique combinations of 200 and 400 Gigabit Ethernet characteristics will be able to support the transfer of huge volumes of data between clusters and AI nodes [intel]. At the same time, it is necessary to ensure minimal latency for parallel high-performance computing of AI machine learning data in real time and ensure highly deterministic delivery with zero packet loss. Here, ISPs have faced the problem of vendor lock-in based on proprietary implementations of InfiniBand and RDMA over Converged Ethernet (RoCE). Intel's Napatech solutions provide standard Ethernet infrastructure using Ultra Ethernet.
The second disadvantage of FPGAs is the high cost of the platforms. However, mid-range FPGAs (e.g. Lattice Avant) are becoming more affordable. They are used in niche applications (automotive, telecommunications, IoT), but are less common in mainstream AI. For example, FPGAs are often used for ASIC prototyping, which reduces development costs.
One interesting point is the use of FPGAs for edge computing - this is the inference of AI models on edge devices where compactness and energy efficiency are important (smart cameras or medical devices). Tools such as Xilinx Vitis AI or Intel OpenVINO simplify the deployment of AI on FPGAs, making them more accessible to developers [32].
The future of FPGAs looks promising as they continue to play a key role in various high-tech areas due to their flexibility, performance, and parallel computing capabilities. The outlook is as follows:
FPGAs are expected to grow in energy-efficient data centers, especially for big data and cloud computing.
They are increasingly used for edge computing for energy-efficient, real-time AI processing. They can perform specialized tasks such as image processing or sensor data analysis directly on the device. In the future, FPGAs are expected to grow in smart cities, medical devices, and industrial automation
FPGAs are ideal for custom models with reduced accuracy or face recognition, license plate recognition in power-constrained cameras.
With the development of 6G and new communication standards, FPGAs will be adapted to support more complex protocols and high-speed signal processing, due to their ability to process large amounts of data with low latency. They are used in base stations, network equipment, and for implementing communication protocols.
FPGAs in support of quantum computing It is possible to process data from quantum computers in niche areas such as cryptography, real-time signal processing, and scientific research. Today, FPGAs are increasingly integrated with CPUs and GPUs in heterogeneous systems. For example, Intel and AMD release chips with integrated FPGAs, which simplifies their use in mass products [31]. Hybrid solutions with specialized performance are being developed today. Swiss scientists from the Institute of Embedded Systems at the Zurich University of Applied Sciences have developed a method for direct data transfer between a GPU and an FPGA for real-time visualization [33]. Simplification of development tools and integration with other architectures will make FPGAs more accessible, which will expand their use in new industries. In the coming years, we can expect their popularity to continue to grow, especially in areas where low latency, energy efficiency, and adaptability are important.

7 Conclusion

All the above devices are hardware and software systems that provide high-speed computing of the AI infrastructure. The hardware architecture, design technologies, and area of use are all different. Each architecture demonstrates different characteristics that achieve certain loads for their best performance. Each brand has its own programming and optimization technologies. It is impossible to identify one clear technology leader from a comparative analysis of these characteristics. If one piece of equipment is better in one area, it has shortcomings in another. If the GPU is the most powerful in performance, then you have to spend effort and resources to overcome the "overheating".
Comparative analysis showed that FPGA for AI has important advantages in such characteristics as performance and parallelism. FPGA is much more energy efficient, can support AI infrastructures in networks, flexible configurability and security. There are also facts of simplification of the FPGA programming toolkit for AI, allowing integration with popular frameworks. The HLS (High-Level Synthesis) toolkit will allow programming FPGA using languages such as C++ or Python. In conclusion, it could been say with confidence that the hypothesis of incomplete use of FPGA has no basis based on the analysis of the use of FPGA platforms in solving AI problems. It is necessary to continue training students in this area with the complication of laboratory work at the level of industrial projects.

References

[1] Tim Dettmers, Deep Learning with GPUs: A Full Hardware Guide, 2023, https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/
[2] Norman P. Jouppi, Cliff Young, Nishant Patil, In-Datacenter Performance Analysis of a Tensor Processing Unit, Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA), 2017.
[3] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel Emer, Efficient Processing of Deep Neural Networks, 2020, Morgan & Claypool
[4] Derek Jones, Jonathan, E. Allen, Yue Yang, William Drew Bennett, William F. Drew Bennett, Maya Gokhale, Niema Moshiri, Tajana S. Rosing, Accelerators for Classical Molecular Dynamics Simulations of Biomolecules, *Journal of Chemical Theory and Computation*, 18(70), 2022
[5] The Growing Role of FPGAs for Accelerating AI Workloads, TechTarget, 2025, https://www.techtarget.com/searchenterpriseai/tip/The-growing-role-of-FPGAs-for-accelerating-AI-workloads
[6] Machine Learning Algorithms for FPGA Implementation in Biomedical Engineering, Science Direct, 2024, www.sciencedirect.com, https://www.sciencedirect.com/science/article/pii/S2405844024026835
[7] Ahmad Mouri, Zadeh Khaki, Ahyoung Choi, Optimizing Deep Learning Acceleration on FPGA for Real-Time and Resource-Efficient Image Classification, Applied Sciences, 15(1), 2025,:422; https://doi.org/10.3390/app15010422
[8] Andrew Boutros, Aman Arora, Vaughn Betz, Field-Programmable Gate Array Architecture for Deep Learning: Survey & Future Directions, *Journal: arXiv*, 2024, https://arxiv.org/abs/2404.10076
[9] Charu C. Aggarwal, Neural Networks and Deep Learning: A Textbook, 2018, Springer
[10] Pedro Antunes, Artur Podobas, FPGA-Based Neural Network Accelerators for Space Applications: A Survey, License: arXiv :2504. 16173v2 [cs.AR], 2025,
[11] Vasileios Kitsakis, Konstantina Kanta, Ioannis Stratakos, Giannis Giannoulis, Dimitrios Apostolopoulos, George Lentaris, Hercules Avramopoulos, Dimitrios Soudris, Dionysios I. Reisis, Design of a Real-Time DSP Engine on RF-SoC FPGA for 5G Networks, Optical Network Design and Modeling, : 540-551, 2020, DOI:10.1007/978-3-030-38085-4_46
[12] Bassam J. Mohd, Khalil M. Ahmad Yousef, Anas Al Majali and Thaier Hayajneh, Quantization-Based Optimization Algorithm for Hardware Implementation of Convolution Neural Networks, *Electronics* 13(9),:1727, https://doi.org/10.3390/electronics13091727, 2024.
[13] Oleksandr Hryshchuk, Sergiy Zagorodnyuk, Managing energy consumption in FPGA-based edge computing systems with soft-core CPU, *Journal of Edge Computing*, 4, (1), :57–72, 2025,
[14] Mangali Sravanthi, Sravan Kumar Gunturi, Mangali Chinna Chinnaiah, Siew-Kei Lam, Divya Vani, Mudasar Basha, Narambhatla Janardhan, Dodde Hari Krishna and Sanjay Dubey, Adaptive FPGA-Based Accelerators for Human–Robot Interaction in Indoor Environments, *Sensors* 24(21): 6986; 2024, https://doi.org/10.3390/s24216986
[15] Muhammad Sufyan, Verilog vs VHDL: A Comprehensive Comparison, 2024, https://www.wevolver.com/article/verilog-vs-vhdl-a-comprehensive-comparison
[16] https://www.intel.com/content/www/us/en/products/details/fpga/development-tools/quartus-prime.html
[17] https://www.xilinx.com/support/download/index.html/content/xilinx/en/downloadNav/vivado-design-tools/archive.html
[18] Francisco de Assis Tavares Ferreira da Silva, Magno Prudêncio de Almeida Filho, Antonio Macilio Pereira de Lucena, Alexandre Guirland Nowosad, Pattern recognition on FPGA for aerospace applications, *Research Society and Development* 10(12), 2021, DOI:10.33448/rsd-v10i12.19181
[19] Ryan Kastner, Janarbek Matai, Stephen Neuendorffer, Parallel Programming for FPGAs, arXiv, 2018, https://arxiv.org/abs/1805.03648
[20] Yuhao Wu, Review on FPGA-Based Accelerators in Deep learning, Proceedings 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2023, DOI: 10.1109/ITNEC56291.2023.10082175
[21] https://fidus.com/
[22] Ryan Allen, Jacob Young, Gary King, Charles James, Scalability of FPGA-Based High-Performance Computing Systems, Stanford University, 2025
[23] https://fidus.com/blog/the-role-of-fpgas-in-ai-acceleration/#introduction-to-ai-and-fpga-technolog
[24] https://theneildave.medium.com/gpu-vs-fpga-vs-asic-vs-tpu-vs-npu-benefits-use-cases-and-manufacturers-4419115069df
[25] https://www.jakelectronics.com/blog/fpga-vs-gpu-comparison-for-highperformance-computing-and-ai
[26] Soufien Gdaim, Mtibaa Abdellatif, Accelerating convolutional neural networks on FPGA platforms: a high-performance design methodology using OpenCL, *Journal of Real-Time Image Processing*, 22, 2025
[27] Hasan Irmak and Nikolaos Alachiotis, An Energy-Efficient FPGA-based Convolutional Neural Network Implementation, 29th Signal Processing and Communications Applications Conference (SIU), 2021, DOI:10.1109/SIU53274.2021.9477823
[28] Hasan Irmak; Daniel Ziener; Nikolaos Alachiotis, Increasing Flexibility of FPGA-based CNN. Accelerators with Dynamic Partial Reconfiguration, 31st International Conference on Field-Programmable Logic and Applications (FPL), 2021, DOI: 10.1109/FPL53798.2021.00061
[29] Lester Kalms, Pedram Amini Rad, Muhammad Ali, Arsany Iskander & Diana Göhringer, A Parametrizable High-Level Synthesis Library for Accelerating Neural Networks on FPGAs, *Signal Processing Systems*, 93,:513–529, 2021
[30] Meriam Dhouibi, Ahmed Karim Ben Salem, Slim Ben Saoud, Accelerating Deep Neural Networks implementation: A survey, *IET Computers & Digital Techniques*, 2021, doi.org/10.1049/cdt2.12016
[31] https://www.tadviser.ru/
[32] https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/overview.htm
[33] https://www.imveurope.com/white-paper/direct-communication-between-fpgas-and-gpus]

Chapter 7

Milk Prediction Using Conformation Traits as Privileged Information

Contribution #27
Cite as:
Awa Samaké, Abdoulaye Baniré Diallo. Milk Prediction Using Conformation Traits as Privileged Information. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-27,
  title ={Milk Prediction Using Conformation Traits as
     Privileged Information},
  author={Awa Samaké and  Abdoulaye Baniré Diallo},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Milk Prediction Using Conformation Traits as Privileged Information

Awa Samaké [1,*], Abdoulaye Baniré Diallo [1]

[1] Université du Québec à Montréal (UQAM), Montreal, Canada
[*] Corresponding author
{samake.awa}@courrier.uqam.ca

Abstract

Predicting milk production is essential in agriculture, driving advancements in herd management and optimization of animal performance. Traditional approaches include statistical models (autoregressive moving average ARMA, seasonal autoregressive integrated moving average SARIMA) and deep learning (DL) methods such as recurrent networks and multilayer perceptrons. However, these models often address regression or forecasting separately, limiting their adaptability. DL models lack flexibility due to fixed input sizes, while statistical models face scalability issues as population size increases. To overcome these limitations, we propose LSTMDropout, a novel model that leverages training-exclusive features through the learning using privileged information paradigm. Combining long-short-term memory (LSTM) networks with heteroscedastic dropout improves variance estimation and uncertainty modeling, enhancing prediction robustness. Furthermore, we investigate the role of conformation traits as privileged information in predicting single- and multi-output milk yield. In the benchmark dataset, the LSTMDropout model surpasses ARMA, SARIMA, and N-BEATS in regression (errors RMSE: 10.55 and MAE: 8.42) and outperforms the temporal fusion transformer in forecasting, achieving lower errors in both single-output (10.55, 8.42) and multi-output tasks (5.49, 4.48). These results highlight the effectiveness of incorporating privileged information with heteroscedastic dropout, offering accuracy, scalability, and computational efficiency for the dairy industry.
Our code is available at: https://github.com/AwaSamake/LSTMDropout

Keywords: Dairy farming; conformation traits; time series; privileged information; deep learning; heteroscedastic dropout.

1 Introduction

In the domain of dairy farming, prediction of milk production represents a crucial aspect of dairy cattle breeding, facilitating advancements in animal monitoring, herd management, and decision-making for farmers [1, 2]. The dairy industry represents a major part of Canada's agricultural sector, ranking second in terms of the economic value of its agricultural income.[1] Several studies have shown that dairy production is influenced by multiple factors, including health, environment, management, and genetics [3, 4, 5, 6, 7, 8]. In order to improve the profitability of dairy production and the living conditions of the animals, models have been developed that help predict the income (value) or yields of dairy farms [1, 9, 10, 11, 12], determine

[1] https://agriculture.canada.ca/en/sector/animal-industry/canadian-dairy-information-centre/dairy-industry

the impact of the necessary nutritional and energy consumption [6, 7, 13, 14], improve disease detection [15, 16], optimize the breeding decisions [17, 2], and genetic predictions [18, 4, 19, 8].The models employed in these studies consist of linear regression, traditional statistical techniques (ARMA, ARIMA, VARMA), machine learning models (decision tree, random forest), and deep learning models (ANN, multilayer perceptron, LSTMs). Linear regression-based methods assume a linear relationship between features, which represents a simplistic assumption. Moreover, the complexity of statistical methods increases as the population size grows. Deep learning models, on the other hand, require identical input sizes for both training and testing. In our case, conformation traits are only available at the beginning of the first lactation.

Our study aims to develop a model able to predict future lactation curves using historical lactation data and phenotypic information, such as estimated breeding values (EBVs) and conformation traits (or body measurements) while minimizing computational expenses. More specifically, we predict two sets of target variables depending on the type of task: either a single-output or a multi-output formulation. In the single output case, the target variable corresponds to the daily milk value (**DMV**). In contrast, for the multi-output setting, the target variables include the daily yields of milk (**DMY**), fat (**DFY**) and protein (**DPY**). We also address two problems: regression [20] and forecasting [21].

To achieve this, we propose a new architecture using conformation traits available only during training to counter future model overfitting. We employ the learning paradigm using *privileged information* (**LuPI**) [22], which leverages information available exclusively during training. The application of this paradigm to temporal data forecasting remains relatively underexplored. With this paradigm, high-performance models can be trained by exploiting data available only at the beginning of the first lactation. The incorporation of this kind of information exclusively during the training phase not only enhances model performance but also contributes to reducing the computational resource requirements in terms of parameters. We extend this framework by integrating heteroscedastic dropout [23] and by using privileged information to estimate variance within Long Short-Term Models (LSTMs) neural networks. Our results demonstrate that the proposed LSTMDropout model significantly outperforms traditional approaches in multi-output forecasting scenarios, compared to single-output models. These findings highlight the potential for enhanced predictive accuracy in dairy farming through the incorporation of privileged information and heteroscedastic dropout in predictive modeling.

The structure of the paper is as follows. Section 2 provides a review of existing research on regression and forecasting problems, and introduces the concept of privileged information. In Section 3, we outline our proposed methodology. The details of the dataset, experimental setup, and the performance evaluation of our model are provided in Sections 4 and 5. Finally, Section 6 conclude the paper by summarizing our findings and discussing their implications.

2 Related works

In this section, we review existing studies and categorize them into two main approaches: regression and forecasting methods. Additionally, we explore previous research that applies the LuPI paradigm.

2.1 Regression methods

Heinrichs *et al.* (1992) [24] utilized regression analysis to derive optimal prediction equations by incorporating various body measurements, including heart girth, body length, hip width, and wither height. This approach established a framework for predictive modeling in dairy cattle research. Furthermore, Dallago *et al.* (2019) [25] identified key Dairy Herd Improvement (DHI)

metrics that significantly impact the first test day milk yield of first lactation dairy cows. They developed and compared three predictive models: multivariate linear regression, random forest, and artificial neural networks. Salamone *et al.* (2022) [10] developed a random forest methodology to predict the first test day milk yield after calving using information from the previous lactation.

2.2 Forecasting methods

Recent research has emphasized the integration of advanced technologies and data analytics to improve both accuracy and efficiency in predicting milk yields. Traditional lactation models typically forecast milk production based on an initial fixed quantity of milk at the start of the lactation period. Liseune *et al.* (2021) [9] introduced a deep learning framework designed to encode the input of the model, predict latent representations of milk yield sequences, and generate the corresponding lactation curves for dairy cows. This approach leverages historical milk yield data from previous lactation cycles to enhance predictive performance. The authors of [1] have proposed a methodology for the design of extensive predictive models that reflect a wider range of factors, centered around the LSTM neural network: univariate and multivariate models. Braginets *et al.* (2022) [11] developed a neural network model to predict milk production and fat content by incorporating ancestral lineage data, including information from mothers, grandmothers, and great-grandmothers. Furthermore, Li *et al.* (2023) [12] applied convolutional neural network-based models to forecast raw milk prices based on consumer behavior. Franceschini *et al.* (2022) [16] used Fourier transform mid-infrared (FT-MIR) spectroscopy to explore whether unsupervised learning methods could offer new insights into dairy cow health [2].

2.3 Learning using Privileged Information LuPI

The current paradigm in machine learning follows a straightforward approach: given a set of learning examples, the goal is to identify a function from a predefined set that best approximates the unknown decision rule. In human learning, the teacher plays a crucial role by providing additional insights such as explanations, comments, comparisons, and metaphors, collectively referred to as *privileged information* (**PI**). A central aspect of this paradigm is that privileged information is only available during the training phase when the teacher interacts with the learner, and not during the testing phase when the learner operates independently [26]. It is important to recognize that PI is relevant to nearly all learning problems and can significantly accelerate the learning process. Vapnik *et al.* introduced the *LuPI* paradigm [26, 27], initially applied to the support vector machine SVM+ method. Recent studies [28, 29, 30] have demonstrated the utility of LuPI for enhancing early predictions using time-series data. This approach utilizes observed data between the time of prediction and the future outcome as privileged information. Hayashi *et al.* (2019) [30] applied this method to both synthetic datasets and real-world atmospheric datasets with eight variables. Karlsson *et al.* (2021) [28] formalized the framework for time series under the term *learning using privileged time-series* (LuPTS), proving that learning efficiency is improved when time series are drawn from a *non-stationary Gaussian-linear dynamical system*. For time series prediction, distillation has emerged as a standard technique for leveraging PI [30, 28, 29]. Lambert *et al.* (2018) [23] proposed a novel LuPI algorithm tailored for convolutional neural networks (CNNs) and recurrent neural networks (RNNs). They employed pre-trained models (such as VGG networks) with dropout layers, utilizing the *information bottleneck re-parametrization trick*. To manage dropout variance, they suggested computing it as a function of PI, leading to what is termed heteroscedastic dropout. This approach uses PI to control uncertainty in the model output, significantly enhancing sample efficiency and accuracy, even with limited training data. Their method was tested on CNNs and RNNs for *multi-modal machine translation* and

image classification.

3 Methodology

3.1 Problem settings

The problem of predicting lifetime milk production, i.e. DMY, DFY, DPY and DMV, in dairy farming involves longitudinal data collection in lactation cycles over time. In this prediction task, we are dealing with multivariate and multiple time series.

Multivariate aspect: This refers to the inclusion of multiple variables beyond the primary output of interest. These variables may have significant effects on the forecasting problem.

Multiple individuals: The dataset includes temporal data for millions of individuals, each represented by a time series. Our dataset contains records collected over multiple lactation cycles for millions of animals. During each lactation cycle, data collection occurs ten (10) times a year.

Our objective is *to predict future lactation performance using historical data from previous lactation while incorporating PI*, such as EBVs and conformation traits (other phenotypic data). In contrast to [28, 29, 30], we define our PI in terms of features in our study.

3.2 LSTMDropout

In the context of learning hidden representations, the bottleneck problem is a challenge particularly when stacking multiple layers. The information bottleneck (IB) issue arises as intermediate layers struggle to retain relevant information while discarding noise. The most commonly adopted solution in the literature is the introduction of attention mechanisms which have proven effective across various domains. When learning time series with missing intermediate information during inference, discarding the missing field is the default option. In this work, we propose an alternative approach by leveraging LuPI with heteroscedastic dropout, as introduced by Lambert *et al.* (2018) [23]. This method combines the IB framework with privileged information (PI) to improve predictive accuracy, replacing attention mechanisms with a customized dropout strategy. Compared to attention-based models such as the Temporal Fusion Transformer (TFT) [31], our method avoids the computational overhead associated with attention mechanisms, making it suitable for large-scale datasets such as those used in dairy farming.

Our architecture, illustrated in Figure 1, builds on the LuPI paradigm [22, 23] and consists of three main steps:

Handling dimensionality differences The first step addresses the dimensional mismatch between the baseline data $X_{base} \in \mathcal{R}^{n_{samples}, n_{seqlength}, n_{features_{base}}}$ and privileged data $X_{priv} \in \mathcal{R}^{n_{samples}, n_{seqlength}, n_{features_{priv}}}$. Separate LSTM blocks are employed to learn representations from these two inputs. Once transformed, these representations are passed through a linear projection layer, a normalization layer, and a ReLU activation layer, ensuring that both inputs are aligned for subsequent processing. Unlike Lambert *et al.* (2018) [23], where X_{base} and X_{priv} are derived from the same input (respectively a full image and cropped regions), our approach handles heterogeneous inputs, time series data (X_{base}) and phenotypic traits (conformation traits & EBVs, X_{priv}). This adaptation adds complexity but ensures that each type of data is processed optimally.

Estimating variance using heteroscedastic dropout This second step involves estimating the variance of the heteroscedastic dropout, proposed by Lambert *et al.* (2018) [23], from the two previously learned representations. Note that in the reference paper [23], they extract the baseline

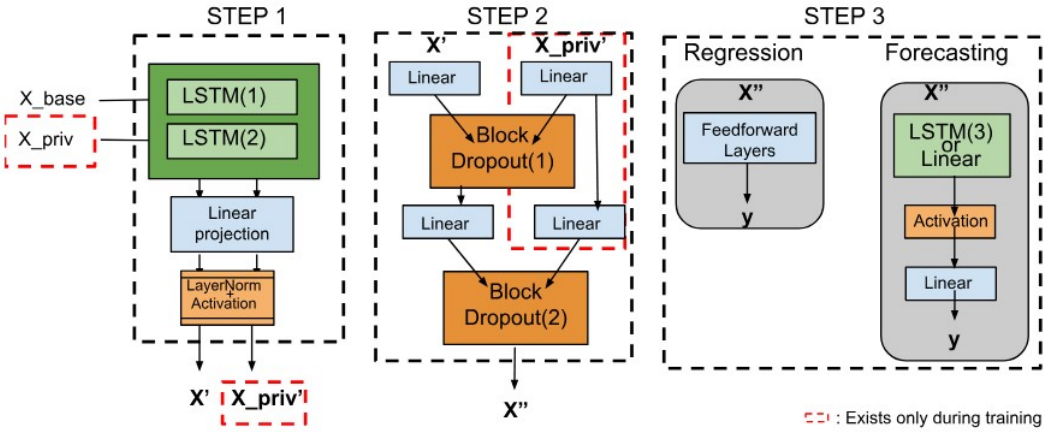


Figure 1: General architecture of our model. Baseline data $X_{base} \in \mathcal{R}^{n_{samples},n_{seqlength},n_{features_{base}}}$; privileged data $X_{priv} \in \mathcal{R}^{n_{samples},n_{seqlength},n_{features_{priv}}}$ and outputs for a regression task $y \in \mathcal{R}^{n_{samples},n_{outputs}}$; a forecasting task $y \in \mathcal{R}^{n_{samples},n_{seqlength}=10,n_{outputs}}$ where $n_{samples}$ is the number of individuals, $n_{seqlength}$ the sequence length, $n_{features_{base}}$ and $n_{features_{priv}}$ the numbers of baseline and privileged features, respectively, and $n_{outputs} \in \{1, 3\}$ the number of output.

and privileged information from the same input, so they do not have to deal with the dimensionality problem encountered in the first step. To calculate the variance of the heteroscedastic dropout during training, we apply two distinct linear layers to the two representations learned in the first step. These learned representations will be the inputs for the third step. Note that during testing, we will only have access to the X_{base} representations as inputs in the second step. We will duplicate these representations before applying the linear layers and the block dropout.

Final step The third and final step is common to both training and testing. However, it varies depending on the nature of the task:

- *Regression.* We use feed-forward layers and output a single value.
- *Forecasting.* First we apply either a linear layer (feed-forward simple) or an LSTM layer (feed-forward RNN). At the end, to obtain the final desired output dimension we apply another linear layer.

In the following sections, we present the experimental settings and evaluate the performance of our LSTMDropout model using real-world milk production data.

4 Experimental settings

4.1 Data and preprocessing

Our experiments utilize dairy production datasets and phenotypic dataset (e.g. conformation traits/body measurements) provided by Lactanet [1]. These datasets cover production records for more than a million Holstein cows. Our focus is specifically on cows that have completed at least three lactation cycles, resulting in a dataset of 11,000 individuals with 100,000 rows and 58 features. Each selected individual has at least three lactation cycles, with one cycle corresponding to 10 data collections per year (10 months per year). The features of the dataset are categorized into several main components.

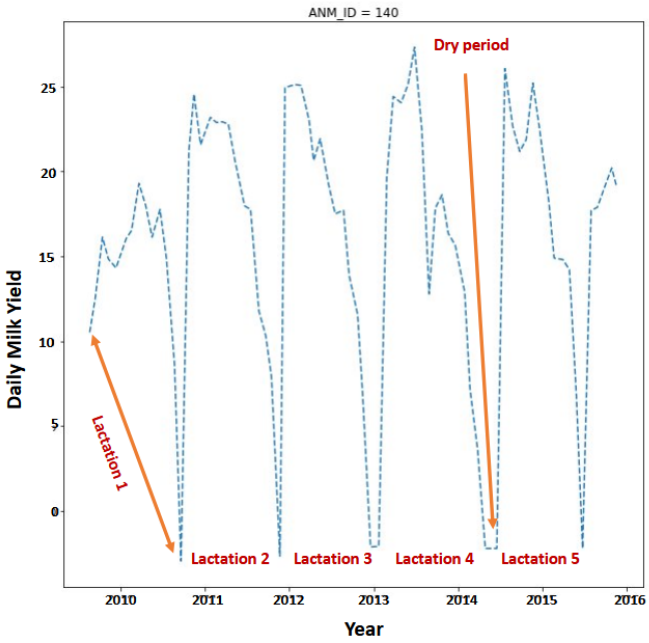


Figure 2: Different lactation cycles of cow 140.

- **Animal**: Contains information about individual animals including birthdate, entry and exit dates from the herd, breed, country, as well as details about their parents such as breed and country. Number of individuals = $2,055,052$ & Rows = $2,245,954$ & Features = 22.
- **Production**: Provides details on lactation cycles, including the quantity of milk, protein and fat on test day, lactose, number of lactation cycles per cow, test dates, daily milk values, daily feed costs and health characteristics such as somatic cell count, β-hydroxybutyrate (BhB). Number of individuals = $1,471,312$ & Rows = $35,976,274$ & Features = 24. Figure 2 shows the different lactation cycles of a cow. As can be seen, a cow can reach a very high peak during lactation and then see her production levels gradually decline at 305 days in milk (i.e. 10 months). We can also see that the curve differs from one lactation to the next for the same cow.
- **Phenotypes**: Includes information on phenotypic markers, taken before the first lactation cycle, such as conformation traits (or body measurements) (e.g. mammary measures, chest width, body depth, loin strength, rump angle) and estimated breeding values (EBVs). Number of individuals = $1,035,532$ & Rows = $1,035,532$ (1 per individual) & Features = 42. In Table 1, we can find more details on our phenotypic features.

The dataset *Production* has been structured as a time series collected between 2006 and 2017, enabling a comprehensive analysis in various dimensions of dairy production and health management. After selecting cows that had completed their first three lactations from the *Production* dataset, we merged the datasets based on animal IDs. The final dataset was split into 60% training, 20% test and 20% validation sets.

Based on their relevance as established by expert knowledge in the dairy production domain, we then deleted the features with more than 50% missing values. Other missing values were filled

Table 1: Repartition of Conformation Traits

Names	Conformation Traits
Estimated Breeding Values EBVs	EBV Milk; EBV Protein; EBV Fat; EBV Protein Percent; EBV Fat Percent
Body Capacity	Stature; Height at Front End; Body Condition BCS; MR; Chest Width; Body Depth; Loin Strength
Rump	Rump Angle; Pin Width; Angularity
Feet & Legs	Hock Lesion HL; Foot Angle; Heel Depth; Bone Quality; Rear Leg Side View; Rear Leg Rear View; Thurl Place Area over the Pelvis
Udder (Mammary)	Udder Depth; Udder Texture; Teat Length; Fore Attach; Fore Teat Place; Rear Teat Place; Median Susp; Rear Attach Height; Rear Attach Width
Dairy Strength	Lactase Persistence LP; Minimum Support Price MSP; MT; Daugther fertility; Calving Ability; Daily Cow Average; Multi-Drug Resistant
Others	Inbreeding; Rvalue; Somatic Cell Score (SCS)

in for numerical features by global mean (like daily milk value) or local mean (per individual's records, like daily milk yield); and for categorical features by the mode. Finally, we proceeded to standardize the data.

Note that *Production* & *Animal* datasets is X_{base}, and *phenotype* one (EBVs and conformation traits) is X_{priv}.

4.2 Baselines models

We benchmark our method against several baseline models, categorized into single-output and multi-output problems:

Single-output problems: ARMA [32]: An autoregressive moving average model that handles stationary time series data. ARMA combines autoregressive (AR) and moving average (MA) components.**SARIMA**: Seasonal ARIMA, which extends ARIMA, autoregressive integrated moving average model, to account for seasonal patterns in time series data. **NBEATS** [33]: An univariate and single-output model designed for time series forecasting. **MuMu+Attention**: An extension of the MuMu model [1] incorporating an attention mechanism [34].

Multi-Output problems: Information Dropout [35]: Originally a regularization method involving multiplicative noise injection in neural network activations. Adapted for forecasting tasks, we modify its architecture to integrate the concept of privileged information $\mathbf{X}_{priv}$ during training, distinguishing it as **Information Dropout (with privileged information)**. This model is referred to as **Information Dropout new** in Table 3. **Temporal Fusion Transformer** (TFT) [31] – an attention-based architecture with interpretable insights into temporal dynamics.

4.3 Evaluation metrics and hyperparameter tuning

Evaluation metrics In general, the evaluation of regression or forecasting models is based on standard metrics such as root mean squared error (RMSE) and mean absolute error (MAE). *Root Mean Squared Error (RMSE)*: Measures the average magnitude of the error. *Mean Absolute Error (MAE)*: Provides the average absolute difference between predicted and actual values. MAE and RMSE are complementary metrics. MAE is more robust to outliers, while RMSE is more sensitive

Table 2: Hyperparameters's tuning.

Hyperparameters	Values
Number of epochs	{ 25; 50; 100; 250 }
Batch size	{ 10; 30; 50; 100; 250 }
Learning rate	{ 10^{-2}; 10^{-3}; 10^{-4}; 10^{-5} }
Activation	{ReLU, Tanh, SoftSign }
Optimizer	Adam
Early stopping	patience ∈ { 5; 8; 10 }
	criterion = loss
Scheduler	patience ∈ { 5; 8; 10 }
	decrease factor
Hidden size	{64; 128; 512; 1024 }
Context size	{16; 32; 64; 128; 512 }
FC size	{64; 128; 512; 1024 }
Number of layers	{ 5; 8; 10 }
Dropout rate	{ 0.15; 0.2; 0.3 }

to outliers or large deviations. While MAE takes into account the mean of the errors, RMSE takes into account their variance. Moreover, RMSE is more relevant in situations where larger errors have a greater impact and require minimization, as in our case (regression and forecasting tasks).

Hyperparameter tuning To optimize the hyperparameters (Table 2), we employ a grid search algorithm. The learning rate is set to 10^{-3}, with a scheduler that allows up to five (5) epochs of non-decreasing validation error before applying a decay factor of 5×10^{-3}. To mitigate overfitting, we implement *early stopping* with a patience of 10 based on a loss criterion. Additionally, we utilize the ReLU activation function and the Adam optimizer.

5 Results

To evaluate the performance of our *LSTMDropout* and *information dropout with privileged information*, we evaluated it across two distinct types of problems: regression and forecasting. Each type further comprises specific tasks categorized into single-output and multi-output scenarios.

- Targets **y** of **single-output** task: *daily milk value* (**DMV**). In dairy farming, the DMV represents the average dollar value of the milk components.
- Targets **y** of **multi-output** task: daily yields of *milk* (**DMY**), *fat* (**DFY**), and *protein* (**DPY**).

■ ***Regression task*** In Table 3, we present our results compared to NBEATS, ARMA (2,2), SARIMA (1,0,1,11), MuMu + Attention, Information Dropout [35], for *single output* regression tasks. It is important to note that in this case, the baseline MuMu+Attention model has access to EBVs and conformation traits data during both training and testing. To determine the parameters of the ARMA and SARIMA models, we used the ACF (autocorrelation function) and

Table 3: Comparison of our results. Regression versus forecasting tasks. **M** = multivariate models. **U** = univariate models. **RNN** = feed-forward RNN. **FF** = feed-forward simple.

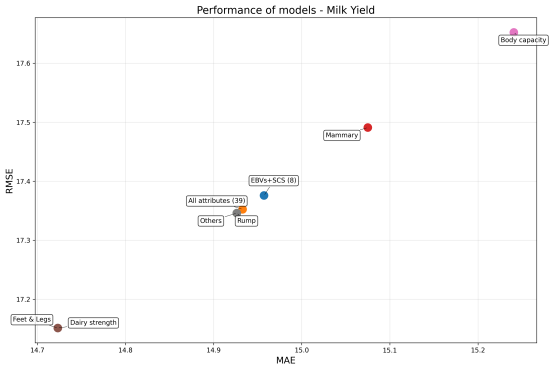
Task	Models	Single		Multi-output	
		RMSE	MAE	RMSE	MAE
Regression	LSTMDropout [M]	**10.5516**	**8.4237**	8.6232	6.9874
	Information Dropout [M]	10.8508	8.6177	-	-
	Information Dropout new [M]	10.8483	8.6234	-	-
	ARMA(2,2) [U]	21.0400	18.0087	-	-
	SARIMA(1,0,1,10) [M]	21.0392	18.0120	-	-
	NBEATS [U]	21.7107	18.7554	-	-
Forecasting	LSTMDropout [RNN]	**10.5517**	**8.4237**	**5.4873**	**4.4752**
	LSTMDropout (without PI) [RNN]	11.2615	9.6288	11.5358	10.4416
	LSTMDropout [FF]	10.6870	8.5263	**5.4832**	**4.4626**
	LSTMDropout (without PI) [FF]	11.2615	9.6288	11.6316	10.4691
	Information Dropout [RNN]	10.8508	8.6177	7.3014	6.4017
	Information Dropout [FF]	10.9178	8.6238	5.8689	4.8780
	Information Dropout new [RNN]	10.8483	8.6234	7.3918	6.3502
	Information Dropout new [FF]	10.9012	8.6179	**5.7492**	**4.7723**
	Temporal Fusion Transformer TFT	14.9444	10.3264	12.4981	11.1652

PACF (partial autocorrelation function). Additionally, we examined the stationarity of our time series, confirming that they are stationary. We observe that our LSTMDropout model has the smallest RMSE and MAE errors. This shows that privileged information, in particular EBVs and conformation traits (X_{priv}), has a positive effect. As a result, their incorporation during training improves the performance of prediction models (see Table 3). In other words, EBVs and conformation traits (X_{priv}) influence the daily value of milk. This is also confirmed by the work [1, 36]. Integrating EBVs and conformation traits only during training improves both the prediction performance of our regression model and the reduction of learning resources. On the other hand, the model *information dropout with privileged information* shows performance similar to the original model in [35]. This suggests that integrating privileged information with this model does not add significant value here.

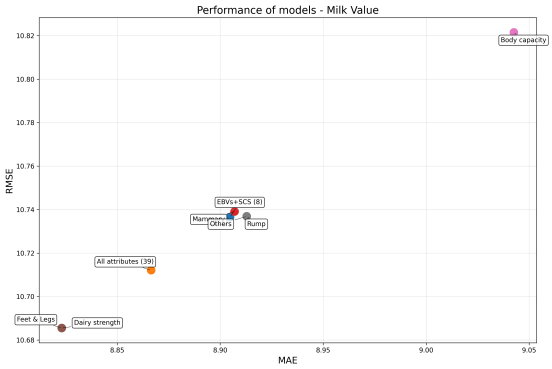
Let us now analyze the LSTMDropout model for single- and multi-output forecasting tasks. We will compare the performance of our two feedforwards, simple and RNN.

■ ***Forecasting task*** For multi-output forecasting (see Table 3), the results show that our LSTMDropout model performs better than baselines regardless of the type of decoder chosen, RNN feedforward or simple feedforward. We also observe that our model outperforms for multi-output problems compared with single-output problems. This can be explained by the important role played by certain production traits, such as DMY, DMF, DMP, and somatic cell count, on DMV and in assessing a cow's production performance. The comparaison between LSTMDropout and LSTMDropout (without PI) shows that the use of PI has real impact on the performance. Table 3 also demonstrates that our LSTMDropout model with a RNN feedforward outperforms the

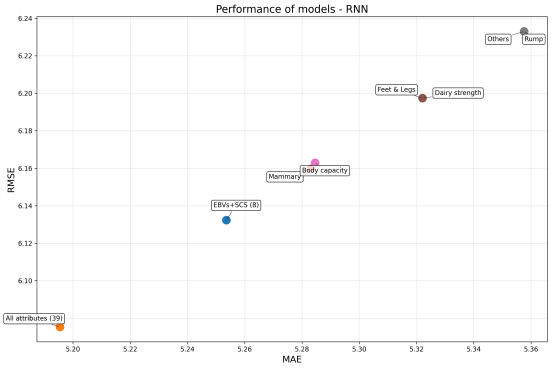
one with a simple feedforward architecture for single-output forecasting. The model *information dropout with privileged information*, when utilizing a simple feed-forward architecture, outperforms the approach in [35] for multi-output forecasting.



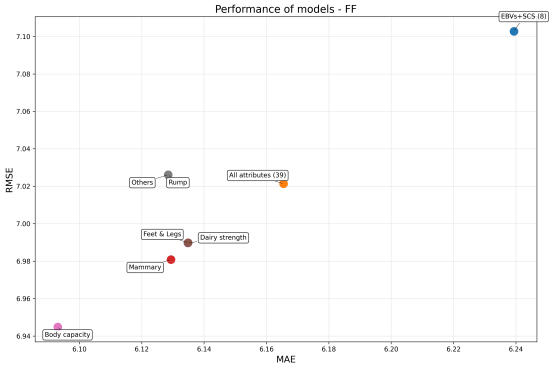
(a) Single-output forecasting - LSTMDropout: target = daily milk yield (DMY).



(b) Single-output forecasting - LSTMDropout: target = daily milk value (DMV).



(c) Multi-output forecasting - LSTMDropout [RNN]: targets = daily yields of milk, fat & protein.



(d) Multi-output forecasting - LSTMDropout [FF]: targets = daily yields of milk, fat & protein.

Figure 3: Ablation study evaluating the impact of different partitions of conformation traits, as PI, on forecasting performance: **(a)** & **(b)** Single-output forecasting targeting daily milk yield (DMY) and daily milk value (DMV). **(c)** & **(d)** Multi-output forecasting using feedforward RNN and feed-forward simple (FF) architectures, with targets including daily yields of milk (DMY), fat (DFY), and protein (DPY).

■ ***Ablation studies*** Given the demonstrated efficiency of our LSTMDropout model in forecasting tasks, we focus our ablation study exclusively on this architecture. This analysis aims to evaluate the impact of input features on the performance of LSTMDropout. In Table 1, we can find more details on our features.

Figure 3 highlights that the features *dairy strength* and *feet & legs* significantly enhance the performance in single-output forecasting tasks. Furthermore, the study reveals that selecting *body capacity* and *all attributes (39)* as privileged information yields improved performance in multi-output forecasting tasks, particularly with simple and RNN feedforward architectures,

respectively. These findings underscore the importance of feature selection and the role of privileged information in optimizing model performance.

6 Conclusion

Integrating privileged information such as estimated breeding values (EBVs) and phenotypic data (conformation traits) significantly improves the performance of forecasting and regression models in dairy farming applications. Estimating the variance using heteroscedastic dropout from privileged information represents a promising approach to improve prediction accuracy in time series analysis. This innovative methodology not only improves predictive models, but also underscores the potential to advance dairy farming practices through data-driven insights. Future research will focus on conducting a rigorous theoretical analysis of the proposed algorithms. Additionally, we plan to generalize the LSTMDropout model for broader applications across various types of time series data. Our upcoming efforts will include benchmarking the model against existing methods that leverage privileged information for regression & forecasting tasks.

References

[1] Charlotte GonÇalves Frasco., Maxime Radmacher., René Lacroix., Roger Cue., Petko Valtchev., Claude Robert., Mounir Boukadoum., Marc-André Sirard., and Abdoulaye Banire Diallo. Towards an effective decision-making system based on cow profitability using deep learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 949–958. INSTICC, SciTePress, 2020.

[2] Andrea Vergani, Alessandro Bagnato, and Marco Masseroli. Predicting bovine daily milk yield by leveraging genomic breeding values. *Computers and Electronics in Agriculture*, 219:108777, 04 2024.

[3] A. Sawa, M. Bogucki, S. Krężel-Czopek, and W. Neja. Relationship between conformation traits and lifetime production efficiency of cows. *International Scholarly Research Notices*, 2013(1):124690, 2013.

[4] H. N. Kadarmideen. Genetic correlations among body condition score, somatic cell score, milk production, fertility and conformation traits in dairy cows. *Animal Science*, 79(2):191–201, 2004.

[5] Fan Zhang, John Upton, Laurence Shalloo, Philip Shine, and Michael D. Murphy. Effect of introducing weather parameters on the accuracy of milk production forecast models. *Information Processing in Agriculture*, 7(1):120–138, 2020.

[6] Marija Špehar, Jelena Ramljak, and Ante Kasap. Estimation of genetic parameters and the effect of inbreeding on dairy traits in istrian sheep. *Italian Journal of Animal Science*, 21(1):331–342, 2022.

[7] JB Cole, BO Makanjuola, CM Rochus, N van Staaveren, and C. Baes. The effects of breeding and selection on lactation in dairy cattle. *Anim Front*, 2023.

[8] Zsolt J. Korösi, László Bognár, Szabolcs Bene, and Ferenc Szabó. The role of the conformation of holstein cows in the sustainability of milk production. *Chemical Engineering Transactions*, 114:745–750, 12 2024.

[9] Arno Liseune, Matthieu Salamone, Dirk Van den Poel, Bonifacius van Ranst, and Miel Hostens. Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning. *Computers and Electronics in Agriculture*, 180:105904, 2021.

[10] M. Salamone, I. Adriaens, A. Vervaet, G. Opsomer, H. Atashi, V. Fievez, B. Aernouts, and M. Hostens. Prediction of first test day milk yield using historical records in dairy cows. *animal*, 16(11):100658, 2022.

[11] Svetlana A. Braginets, Olga V. Galanina, and Vadim S. Grachev. *Neural Networks and Artificial Intelligence in Stockbreeding and Forecasting Dairy Cattle Productivity*, pages 199–207. Springer Nature Singapore, Singapore, 2022.

[12] Zongyu Li, Anmin Zuo, and Cuixia Li. Predicting raw milk price based on depth time series features for consumer behavior analysis. *Sustainability*, 15(8):6647, 2023.

[13] Y. M. Lee, C. G. Dang, M. Z. Alam, Y. S. Kim, K. H. Cho, K. D. Park, and J. J Kim. The effectiveness of genomic selection for milk production traits of holstein dairy cattle. *Asian-Australasian journal of animal sciences*, 2020.

[14] C.S. Parra, J.M. Bragatto, F.A. Piran Filho, S.M.S. Silva, B.F. Tuzzi, C.C. Jobim, and J.L.P. Daniel. Effect of sealing strategy on the feeding value of corn silage for growing dairy heifers. *Journal of Dairy Science*, 104(6):6792–6802, 2021.

[15] M.T.M. King, S.J. LeBlanc, E.A. Pajor, T.C. Wright, and T.J. DeVries. Behavior and productivity of cows milked in automated systems before diagnosis of health disorders in early lactation. *Journal of Dairy Science*, 101(5):4343–4356, 2018.

[16] S. Franceschini, C. Grelet, J. Leblois, N. Gengler, and H. Soyeurt. Can unsupervised learning methods applied to milk recording big data provide new insights into dairy cow health? *Journal of Dairy Science*, 105(8):6760–6772, 2022.

[17] F Zhang, KA Weigel, and VE Cabrera. Predicting daily milk yield for primiparous cows using data of within-herd relatives to capture genotype-by-environment interactions. *Journal of Dairy Science*, 105(8):6739–6748, 2022.

[18] T.H. Short and T.J. Lawlor. Genetic parameters of conformation traits, milk yield, and herd life in holsteins. *Journal of Dairy Science*, 75(7):1987–1998, 1992.

[19] M. Khansefid, M. Haile-Mariam, and J.E. Pryce. Including milk production, conformation, and functional traits in multivariate models for genetic evaluation of lameness. *Journal of Dairy Science*, 104(10):10905–10920, 2021.

[20] K.A. Venkatesh, K. Mohanasundaram, and V. Pothyachi. 8 - regression tasks for machine learning. In Tilottama Goswami and G.R. Sinha, editors, *Statistical Modeling in Machine Learning*, pages 133–157. Academic Press, 2023.

[21] Casper Solheim Bojer. Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities. *International Journal of Forecasting*, 38(4):1555–1561, 2022. Special Issue: M5 competition.

[22] Vladimir Vapnik, Akshay Vashist, and Natalya Pavlovitch. Learning using hidden information (learning with teacher). In *2009 International Joint Conference on Neural Networks*, pages 3188–3195. IEEE, 2009.

[23] John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8886–8895, Los Alamitos, CA, USA, 06 2018. IEEE Computer Society.

[24] AJ Heinrichs, GW Rogers, and JB Cooper. Predicting body weight and wither height in holstein heifers using body measurements. *Journal of dairy science*, 75(12):3576–3581, 1992.

[25] Gabriel Machado Dallago, Darcilene Maria de Figueiredo, Paulo César de Resende Andrade, Roseli Aparecida dos Santos, René Lacroix, Débora E. Santschi, and Daniel M. Lefebvre. Predicting first test day milk yield of dairy heifers. *Computers and Electronics in Agriculture*, 166:105032, 2019.

[26] Vladimir Vapnik. *Estimation of dependences based on empirical data: Springer series in statistics*. Springer-Verlag, Berlin, Heidelberg, 1982.

[27] Vladimir Vapnik and Rauf Izmailov. Learning with intelligent teacher: Similarity control and knowledge transfer. In Alexander Gammerman, Vladimir Vovk, and Harris Papadopoulos, editors, *Statistical Learning and Data Sciences*, pages 3–32, Cham, 2015. Springer International Publishing.

[28] Rickard K. A. Karlsson, Martin Willbo, Zeshan Hussain, Rahul G. Krishnan, David Sontag, and Fredrik D. Johansson. Using time-series privileged information for provably efficient learning of prediction models. *CoRR*, abs/2110.14993, 2021.

[29] Bastian Jung and Fredrik Daniel Johansson. Efficient learning of nonlinear prediction models with time-series privileged information. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[30] Shogo Hayashi, Akira Tanimoto, and Hisashi Kashima. Long-term prediction of small time-series data using generalized distillation. *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019.

[31] Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.

[32] J. Contreras, R. Espinola, F.J. Nogales, and A.J. Conejo. Arima models to predict next-day electricity prices. *IEEE Transactions on Power Systems*, 18(3):1014–1020, 2003.

[33] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

[34] Vahid Naghashi, Gabriel Dallago, Abdoulaye Banire Diallo, and Mounir Boukadoum. Univariate and multivariate time-series methods to forecast dairy income. In *2nd AAAI Workshop on AI for Agriculture and Food Systems*, 2023.

[35] Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12), 2018.

[36] Gordon Atkins. How is functional conformation affecting your herd's longevity and profitability? *AgProud*, 2018.

Chapter 8

Deep Metric Learning for Face Recognition with Focus on Twins: Triplet Network Approach

Contribution #28
Cite as:
Kapysh Nurseiit, Akbayan Bekarystankyzy. Deep Metric Learning for Face Recognition with Focus on Twins: Triplet Network Approach. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-28,
  title ={Deep Metric Learning for Face Recognition
     with Focus on Twins: Triplet Network Approach},
  author={Kapysh Nurseiit and  Akbayan Bekarystankyzy},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Deep Metric Learning for Face Recognition with Focus on Twins: Triplet Network Approach

Kapysh Nurseiit[1], Akbayan Bekarystankyzy[1, 2]
[1]Satbayev University, Institute of Automation and Information Technologies, Almaty, Kazakhstan
[2]Narxoz University, School Of Digital Technologies, Almaty, Kazakhstan
[1]Kapysh.N@stud.satbayev.university, [2]Akbayan.b@gmail.com

***Abstract*—Facial recognition systems still have difficulty recognizing very similar people, such as identical twins. In this study, we investigated the effectiveness of deep metric learning, specifically Triplet Networks trained on FaceNet embeddings, to solve the twins problem. We use the CelebA dataset, supplemented with manually collected images of twins, for experiments and present a comprehensive evaluation through ROC-AUC, FAR/FRR and t-SNE visualization. Our method achieves an AUC value of 0.89 and demonstrates strong clustering by identity, setting a foundation for further evaluation on dedicated twin datasets.**

***Index Terms*—face recognition, triplet network, twins, FaceNet, biometric verification, deep metric learning**

I. Introduction

Facial recognition systems are becoming increasingly common in modern biometric applications such as access control, identity verification, and video surveillance. Despite their success, these systems often encounter difficulties when faced with unusual situations such as identical twins whose biometric characteristics are remarkably similar. This issue is crucial because false acceptances in high-security environments can compromise system integrity, especially when standard convolutional neural network (CNN) classifiers are used. [1]

To improve the discriminative power of face embeddings, recent studies have proposed deep metric learning methods, particularly using Triplet Networks.[2] According to Schroff et al. [1], Triplet Loss can effectively structure the embedding space by reducing intra-class distance and increasing inter-class separation. Koch et al.[3] demonstrated that Siamese and Triplet architectures are suitable for one-shot learning and face verification tasks. However, few works have examined how these networks perform in distinguishing between identical twins.

Public datasets such as CelebA offer a large number of face images but lack labeled twin identities [4]. To explore this limitation, we augment the CelebA dataset with a custom collection of identical twin face images. Our research addresses the gap in facial recognition by investigating how well a Triplet-based deep metric learning model distinguishes between twin and non-twin identities. The objectives of this study are:

- To assess the performance of Triplet Networks in distinguishing between twins.
- To evaluate the model's discriminative power through metrics such as ROC-AUC, FAR/FRR, and Equal Error Rate (EER).
- To visualize the embedding space using t-SNE.
- To test real face pairs manually using the trained model.

We use the FaceNet architecture to extract embeddings and train a small fully connected Triplet Network using these features.

II. State of the Art

To gather information on relevant research on facial recognition and twin identification, we conducted a literature review using Google Scholar, IEEE Xplore, and arXiv. We searched for terms such as "triplet facial recognition," "Siamese network verification," and "identical twin facial recognition." The selected articles were used to define our architecture, implementation strategy, and evaluation.

The foundation for our network design is the Triplet Network introduced by Hoffer and Ailon [2], which learns to minimize distances between similar samples while maximizing those from different classes. This was extended in FaceNet by Schroff et al. [1], which used triplet loss for facial embedding learning, achieving state-of-the-art verification performance. We adopted this embedding structure in our pipeline.

Koch et al. [3] proposed Siamese networks for one-shot learning, relevant for situations with limited labeled data such as twin datasets. In addition, ArcFace introduced by Deng et al. [5] improved inter-class separability using angular margin loss, inspiring the design of more discriminative loss functions.

Wang et al. introduced CosFace[6], a margin-based loss function that focuses on improving inter-class separability in deep facial recognition models, complementing previous metric learning such as Triplet Loss and ArcFace.

To improve the reliability of the dataset, Cao et al. created VGGFace2[7], which contains a large number of images of faces in a wide range of poses and ages, which served as the main resource for learning facial representation.

In addition, Xu et al. proposed further improvements to face recognition models by introducing X2-Softmax, a margin adaptive loss function that dynamically adjusts angular margins based on inter-class angles, thereby enhancing the discriminative power of deep face recognition systems in complex scenarios [8].

For data, we relied on the CelebA dataset compiled by Liu et al. [9], which is widely used in face analysis benchmarks. However, it lacks labeled twin identities. To overcome this, we manually collected images of twins from publicly available

sources, including the pinterest, freepik, pixabay,NBC News Twin Study and etc.[10][11][12][13].

Zhang et al. [14] introduced MTCNN, which we used for facial detection and alignment before embedding extraction. The application of Triplet Loss in re-identification was supported by Hermans et al. [15], reinforcing its viability for biometric learning.

We also reviewed DeepFace by Taigman and co-authors [16], which emphasized the effectiveness of end-to-end face recognition training, and the IARPA Janus benchmark test [17], which solved the problem of unrestricted recognition by showing problems related to differences in posture and lighting.

To ensure broader contextual relevance, we included recent surveys such as Fuad et al. [18], which summarized the evolution and advancements of deep learning techniques in facial recognition.

Our use of FaceNet embeddings, triplet loss, and twin-based evaluation is grounded in these previous works and combines best practices from the literature on deep metric learning and face verification.

III. Methodology

This section presents the materials and methods used in our experimental setup. The methodology consists of four main stages: data preparation, face alignment, embedding extraction, and Triplet Network training.

A. 3.1 Dataset and Preprocessing

We used the CelebA dataset[4], which contains more than 200,000 face images labeled by identity, but we take only 170,000.

To better evaluate genetically similar twins, we manually collected images from open sources and turned to the Twins Days [19] dataset, which is a specialized large-scale twin dataset designed specifically for biometric assessment.

The Twins Days dataset, although not directly used here, defines a standard for future evaluations of Twin verification systems.

To introduce more challenging cases, we manually collected and pre-aligned 100 additional images of identical twins publicly available sources such as Pinterest, Freepik, Pixabay, and NBC's Twin Study[10][11][12][13] . Each twin pair was cropped and saved in a structured folder format.

All images were preprocessed using the Multi-task Cascaded Convolutional Neural Network (MTCNN) for face detection and alignment, then resized to 160×160 pixels[14]. Embeddings were extracted using a pre-trained FaceNet model and stored along with image paths.

Initially, we trained the model using only the CelebA dataset without twin-specific images. The model achieved an AUC of 0.9212, which indicates strong general recognition ability. However, further inspection revealed that this version performed poorly on near-duplicate identities and failed to distinguish twins, with FAR remaining low only at the cost of extremely high FRR (e.g., FAR = 0.0119, FRR = 0.4350 at threshold 8.0). These limitations motivated the inclusion of twin-specific samples to enhance discrimination.

After embedding extraction, the dataset was split into training and validation subsets based on person identity, ensuring no overlap. We used an 80/20 split across unique person IDs. The final dataset consisted of approximately 18,400 triplets.

B. 3.2 Embedding Extraction

We used the `InceptionResnetV1` model from the `facenet-pytorch` library, pretrained on the VGGFace2 dataset, to extract 512-dimensional feature vectors (embeddings) for each face. Each embedding was then passed through a fully connected Triplet Network for training. Before training, all embeddings were normalized to unit length using L2 normalization.

C. 3.3 Triplet Network and Training

The Triplet Network used a fully connected architecture with the following structure:

- FC(512 → 256) + ReLU + BatchNorm
- FC(256 → 128) + ReLU + BatchNorm
- FC(128 → 64) + L2 normalization

The network was trained using the Triplet Margin Loss function:

$$L = \sum_{i=1}^{m} \max\Big(\Big\| f\left(A^{(i)}\right) - f\left(P^{(i)}\right) \Big\|_2^2 - \Big\| f\left(A^{(i)}\right) - f\left(N^{(i)}\right) \Big\|_2^2 + \alpha,\ 0 \Big)$$

[1]

where:

- $f(A)$: embedding of the anchor image
- $f(P)$: embedding of the positive image (same identity)
- $f(N)$: embedding of the negative image (different identity)
- $\alpha = 0.5$: margin hyperparameter

The model was trained for 20 epochs using the Adam optimizer with a learning rate of 1×10^{-3}. Both training and validation losses were monitored each epoch to prevent overfitting.

D. 3.4 Evaluation Procedure

After training, the model was evaluated using multiple metrics:

- Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC)
- Equal Error Rate (EER), where False Accept Rate (FAR) = False Reject Rate (FRR)
- FAR/FRR at various thresholds (5–10)
- t-SNE visualization of embedding space

A balanced test set of pairs of the same identity and different identity was used for the evaluation, including CelebA identities and twins.[4][11][10][12] The best threshold was selected by minimizing the EER in the validation set.

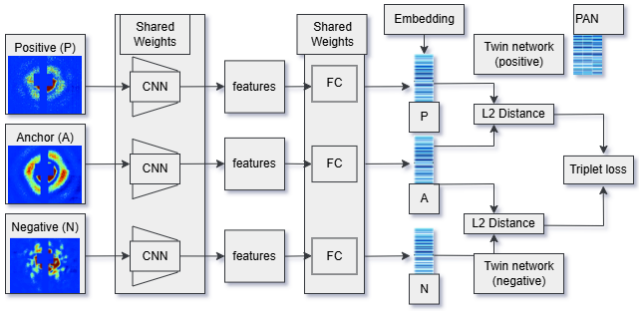


Fig. 1. Triplet Network Architecture: The model processes anchor (A), positive (P), and negative (N) samples using shared CNN and FC layers, compares embeddings using L2 distance, and applies the triplet loss.

IV. Results

This section presents the experimental results obtained using the trained Triplet Network model. The evaluation was carried out on a validation set consisting of pairs of same and different identity, including twins.

A. 4.1 Distance and Discrimination

The model was evaluated on a validation set composed of equal numbers of same-identity and different-identity face pairs, including twins. The Euclidean distances between embeddings were computed. A histogram of these distances is presented in Figure 2.

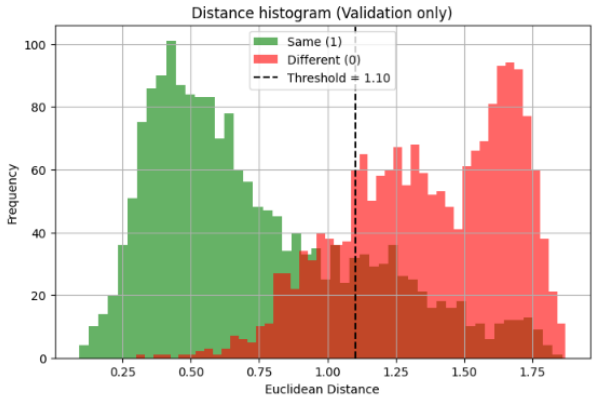


Fig. 2. Distribution of Euclidean distances between validation pairs. Green: same-identity, Red: different-identity.

B. 4.2 ROC Curve and Threshold Selection

To evaluate the discriminative power of the model, we calculated the ROC curve (Figure 3). The Area Under the Curve (AUC) reached 0.8847, which shows strong discriminative capability.

To determine whether two face embeddings belong to the same person, we use the Euclidean distance between them. A threshold must be selected to make this decision: if the distance is less than the threshold, the pair is classified as 'same identity'; otherwise, it is classified as 'different identities'.

We performed an empirical analysis of distances on the validation set (both twin and non-twin pairs) and plotted the distributions. Based on the Equal Error Rate (EER) point 0.1952 where the False Acceptance Rate (FAR) 0.1974 equals the False Rejection Rate (FRR) 0.1931 we selected a threshold of **1.09**. This threshold corresponds to the crossing point in the ROC curve and histogram where both error rates are balanced.

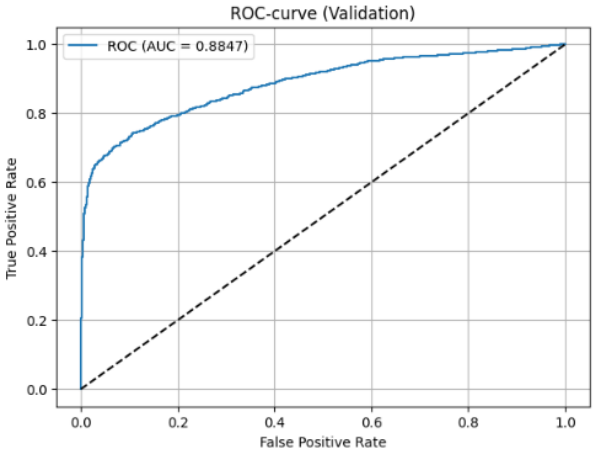


Fig. 3. ROC curve on validation set with AUC = 0.8847.

C. 4.3 Confusion Matrix on Validation Set

Furthermore, the confusion matrix (Figure 4) and the validation metrics were computed using this threshold. While 1.09 does not guarantee the highest possible precision or recall individually, it provides the best balance between the two. In real-world deployments, this threshold could be tuned further depending on whether *security* (favoring low FAR, i.e., high precision) or *convenience* (favoring low FRR, i.e., high recall) is prioritized.

In our study, we aimed for **balanced performance**, as the goal was to show the system's general discriminative ability, especially under difficult conditions such as distinguishing identical twins.

- **True Positives (TP)**: 1486 (same identity, correctly predicted)
- **True Negatives (TN)**: 1461 (different identity, correctly predicted)
- **False Positives (FP)**: 358 (different identity, predicted as same)
- **False Negatives (FN)**: 383 (same identity, predicted as different)

D. 4.4 Clustering and Visualization

To further evaluate the embedding space, we applied t-SNE dimensionality reduction on randomly sampled face embeddings from the validation set. As seen in Figure 5, the network learned to cluster same-identity faces (including twins) closely, with visible separation between identities.

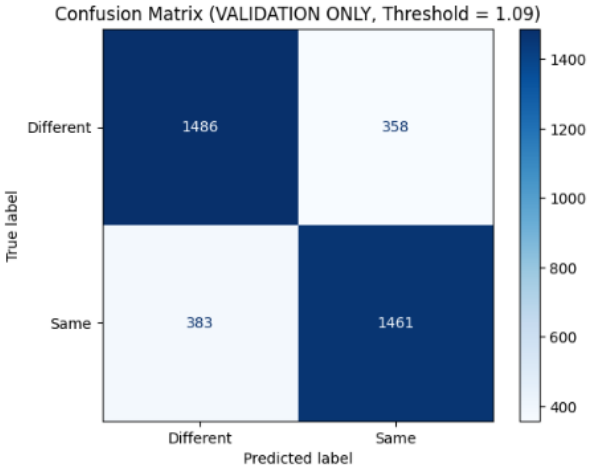


Fig. 4. Confusion matrix on validation set with threshold = 1.09.

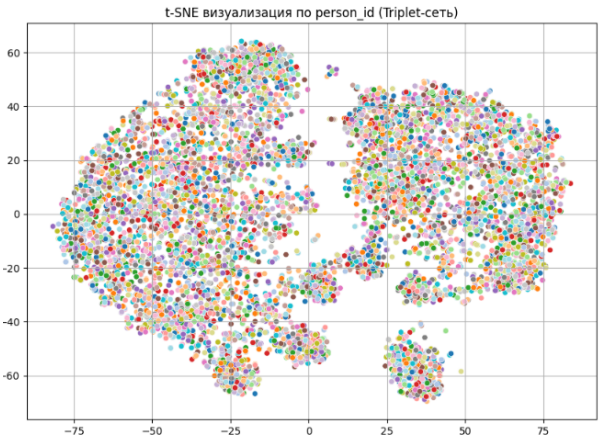


Fig. 5. t-SNE visualization of the learned embedding space.

E. 4.5 Manual Face Pair Testing

We also manually tested the model using four real cropped images:

- Two photos of identical twins (expected similar).
- Two photos of unrelated individuals (expected different).

The model correctly recognized the twins with distance = 0.0388 (Similar) and misclassified the different-identity pair with distance = 0.1152 as similar, indicating a need for tighter threshold tuning or hard negative mining in training.

F. 4.6 Summary of Validation Metrics

Table I summarizes the main evaluation metrics obtained from the validation set. These metrics provide a comprehensive overview of the model's verification performance.

The area under the curve (AUC) of 0.8847 demonstrates good overall recognition ability. The Equal Error Rate (EER) of 0.1952 indicates a balanced point where the model's False Accept Rate (FAR) and False Reject Rate (FRR) are nearly equal. A low FAR of 0.1974 and FRR of 0.1931 confirm the system's ability to differentiate between same and different identities with high reliability. The threshold value of 1.11 was selected based on minimizing the EER during validation.

TABLE I
SUMMARY OF EVALUATION METRICS (VALIDATION)

Metric	Value
AUC	0.8847
EER	0.1952
FAR	0.1974
FRR	0.1931
Threshold	1.11

V. DISCUSSION

The results of our experiments confirm that the proposed Triplet Network architecture, when used in combination with pretrained FaceNet embeddings, demonstrates strong capabilities in distinguishing between highly similar faces — including identical twins. With an achieved AUC of 0.8847 and Equal Error Rate (EER) of 0.1952, our model performs competitively with prior state-of-the-art methods, despite being trained on a comparatively smaller and more challenging dataset.

Unlike traditional face verification benchmarks, our work specifically incorporated manually collected images of identical twins, introducing hard-positive samples during training. This exposure allowed the network to learn finer distinctions in facial structure and texture, contributing to the increased generalization capability observed in validation performance.

To test the applicability of the model in real conditions, we further tested the model using manually selected face images of twins and unrelated individuals. The model accurately identified the same-twin pairs as similar and different identities as dissimilar based on the learned threshold (approximately 1.09), which allows it to be used in practical verification tasks.

In addition, the analysis of the confusion matrix on the validation set confirmed a well-balanced classification results. True positive and true negative rates remained high, while false predictions were limited and symmetrically distributed — suggesting no systematic bias in the decision boundary. The t-SNE visualization further confirmed the effectiveness of the learned embedding space, with identity clusters clearly formed, including well-separated pairs of twins.

To revisit our original research objectives:

1) The model demonstrated the ability to separate visually similar individuals, including identical twins, in the learned embedding space.
2) Quantitative evaluation yielded high verification accuracy with AUC = 0.8897 and EER = 0.1952.
3) Clustering analysis through t-SNE reinforced the model's ability to form discriminative representations for each identity.
4) Manual validation with cropped real-world images showed the practical viability of our approach for pairwise face comparison.

These outcomes highlight the promise of lightweight deep metric learning systems to address one of the most challenging

aspects of biometric authentication: recognizing genetically identical individuals. While real-world deployment is yet to be realized, the current study lays a solid foundation for future implementation within biometric systems — such as identity verification in secure government services, airport border control, or advanced mobile authentication systems.

VI. Conclusion and Future Work

A. Conclusion

This research explored the problem of distinguishing between highly similar individuals, such as identical twins, using a combination of FaceNet embeddings and a lightweight Triplet Network. By augmenting the CelebA dataset with manually collected images of twins, the model was exposed to hard-positive samples that enhanced its ability to learn discriminative representations.

The proposed system achieved strong results, with an AUC of 0.8897 and an EER of 0.1952 on the validation set. t-SNE confirmed the formation of compact clusters for individuals, including close but separable representations for twin pairs. Manual testing with real image pairs further demonstrated the practical effectiveness of the model.

These results show that even simple and computationally efficient models, carefully designed and trained using appropriate data, can contribute meaningfully to solving complex biometric problems. The work also highlights the importance of including domain-specific edge cases in training to ensure generalization beyond common benchmarks.

B. Future Work

Future research will focus on several directions:

- **Twin-specific dataset expansion:** Collecting a larger set of labeled twin images, including from datasets such as Twins Days or real-world passport imagery.
- **Hard negative mining:** Integrating dynamic mining strategies during training to improve robustness against near-duplicate identities.
- **Hybrid loss functions:** Combining Triplet Loss with classification-based objectives (e.g., ArcFace or CosFace) to enforce both intra-class compactness and inter-class separability.
- **Deployment and testing:** Investigating how the model performs in real-world authentication systems, such as mobile biometrics, secure facility access, or forensic identification.
- **Exploring lightweight architectures:** Testing whether more compact or transformer-based models can retain performance while reducing inference latency in deployment environments.
- **Benchmarking on Twins-specific datasets:** In future work, we plan to benchmark our model against more specialized datasets such as Twins Days [19], which focus on recognizing identical twins and thus would validate the practical robustness of our approach.
- **Integration into Biometric Authentication Systems:** Based on the principles discussed by Jain et al. [20], future implementations will aim to integrate lightweight twin-aware face verification into mobile devices, airport border control systems, and high-security identity management platforms.

This study provides a foundation for further research and deployment of twin-aware biometric recognition systems — addressing one of the most challenging cases in facial verification today.

Acknowledgments

The author would like to thank the academic advisors at Narhoz University, Satbayev University. This work was conducted as part of a graduate thesis project.

References

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[2] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition (SIMBAD)*. Springer, 2015, pp. 84–92.

[3] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015.

[4] Z. Liu, P. Luo, X. Wang, and X. Tang, "Celeba dataset," http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, accessed: 2025-04-19.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[6] H. Wang, Y. Wang, Z. Zhou, and X. Tang, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74, 2018.

[8] J. Xu, X. Liu, X. Zhang, Y.-W. Si, X. Li, Z. Shi, K. Wang, and X. Gong, "X2-softmax: Margin adaptive loss function for face recognition," *arXiv preprint arXiv:2312.05281*, 2023.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738.

[10] "Identical portraits of twins by photographer martin schoeller," https://www.pinterest.com/pin/426997608393027302/, accessed: 2025-04-19.

[11] "Freepik image search – identical twins face," https://www.freepik.com/search?ai=excluded&format=search&last_filter=page&last_value=65&page=65&

query=identical+twins+face&type=photo, accessed: 2025-04-19.
[12] "Pixabay image search – identical twins face," https://pixabay.com/images/search/identical%20twins%20face/, accessed: 2025-04-19.
[13] NBC News, "Twin study: Effects of smoking, sun and stress on the skin of twins," https://www.nbcnews.com/slideshow/health/effects-of-smoking-sun-and-stress-on-the-skin-of-twins-33418710, 2011, accessed: 2025-04-19.
[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[15] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, arXiv preprint.
[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.
[17] B. F. Klare, B. Klein, E. Taborsky, A. Jain, and et al., "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1931–1939.
[18] M. T. H. Fuad, A. A. Fime, D. Sikder, M. A. R. Iftee, J. Rabbi, M. S. Al-Rakhami, A. Gumaei, O. Sen, M. Fuad, and M. N. Islam, "Recent advances in deep learning techniques for face recognition," *arXiv preprint arXiv:2103.10492*, 2021.
[19] "Twins days festival official website," https://twinsdays.org/, accessed: 2025-04-26.
[20] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

Chapter 9

Integrating Machine Learning for Predictive Maintenance in Urban Bus Fleets using Oil Analysis

Contribution #29
Cite as:
Margarida Duarte, Mateus Mendes, Torres Farinha, Jorge Almeida. Integrating Machine Learning for Predictive Maintenance in Urban Bus Fleets using Oil Analysis. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-29,
  title ={Integrating Machine Learning for Predictive
     Maintenance in Urban Bus Fleets using Oil
     Analysis},
  author={Margarida Duarte and  Mateus Mendes and
     Torres Farinha and  Jorge Almeida},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Integrating Machine Learning for Predictive Maintenance in Urban Bus Fleets using Oil Analysis

Margarida Duarte[1,*], Jorge Almeida[1,2], Mateus Mendes[1,2], Torres Farinha[1,2]

[1]Instituto Superior de Engenharia de Coimbra, Instituto Politécnico de Coimbra, Rua Pedro Nunes, 3030-199 Coimbra, Portugal
[2]RCM2+ Research Centre in Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Higher Institute of Engineering of Coimbra, Rua Pedro Nunes-Quinta da Nora, 3030-199 Coimbra, Portugal
{a21280328*, jalmeida, mmendes, tfarinha}@isec.pt

Abstract

The growing requirements for high level of service with low operational costs makes it necessary for all public transport operators to introduce more complex and effective maintenance strategies. Historical times-based and reactive maintenances are found to be inadequate to mitigate the occurrence of unplanned failures and frequent service interruptions. Within this category, the paradigm of Condition-Based Maintenance (CBM) appears as a proactive strategy, associated to a Data-Driven Maintenance (DDM), to support the maintenance actions to be made - in real-time or retrospectively - using information provided by the assets themselves, representing their real lifetime operating activities. On the available diagnostic techniques, engine lubricating oil analysis is a powerful tool for monitoring engine health, with the capability of early detection of degradation processes including additive consumption, metals wear, and contamination.

This study proposes an oil analysis-based approach combined with machine learning techniques for predictive maintenance in urban bus fleets. The methodology integrates Principal Component Analysis (PCA) for dimensionality reduction and K-Means clustering for the identification of oil degradation profiles. Four distinct clusters were identified, each representing a different degradation mechanism: chemical oxidation, mechanical wear with fuel dilution, preserved oil condition, and severe multifactorial degradation. Subsequently, a Single-Layer Perceptron was employed to interpret the most influential factors in the formation of each cluster, enabling a deeper understanding of the degradation patterns observed.

By combining structured oil monitoring with intelligent data analysis, the proposed framework supports smarter, adaptive maintenance planning and contributes to the digital transformation of fleet management systems.

Keywords: Machine Learning; Condition-Based Maintenance; Oil Analysis; Principal Component Analysis; K-Means Clustering.

1 Introduction

Aiming the increasing requirements for vehicle availability, operational performance and cost control, public transport systems are facing great challenges. Many service interruptions and operational costs can be attributed to inflexible maintenance strategies such as Corrective Maintenance (CM). In this context, the use of Condition-Based Maintenance (CBM) has been presented as a proactive maintenance policy that uses historical or real-time data and diagnostic tools to perform interventions based on the actual condition of assets, increasing the reliability and reducing the downtime of vehicles.

This kind of CBM is considerably important in the domain of public transportation, where different

condition monitoring services - for example, lubricating oil analysis - can be made to work together for a more complete understanding of the vehicles' health. It enables early fault detection and more efficient decision making on oil change and component inspection. Normative frameworks as the NP EN 13306 [1] standard defines fundamental concepts and terminology for maintenance management, which are a fundamental issue to support implementation of organized and efficient processes in maintenance of public fleets.

Meanwhile, recent developments in data science and machine learning have paved the way to improve and optimize predictive maintenance approaches. The machine learning models can detect complex, non-linear patterns in high-dimensional datasets generated from oil samples. By combining information on oil condition with various learning ML methods, maintenance programs can be adjusted to actual wear states.

This study presents a new oil condition monitoring framework that integrates oil analysis with Machine Learning (ML) based simulation models to characterize and predict the lubrication degradation profiles in the urban bus fleet. This paper seeks to explore the applicability of supervised learning techniques toward the classification of states of degradation from which further intelligent and adaptive maintenance decisions may be made. Eventually, it is about bringing in digitalization on a higher level in fleet management and open the door for more sophisticated systems such as self-learning maintenance solutions.

To guide this investigation, the following research questions are addressed:

- Is it possible to classify lubrication degradation profiles via ML methods on the basis of data of urban bus fleets?
- How to augment the oil condition data interpretation in a maintenance environment by utilizing PCA, clustering and supervised learning?
- To what extent can this method be used for intelligent and adaptive maintenance strategy generation?

This paper is organized as follows: Section 2 reviews the related work, section 3 outlines the methodology employed, i.e. data collection, pre-processing and implementation of PCA, K-Means clustering and the Single-Layer Perceptron. Section 4 describes and discusses the analysis results and section 5 finishes with the main conclusions and a few tips for future work.

2 State of the Art

Oil analysis has historically provided a cornerstone foundation for monitoring the health of the engine and predicting failures in, at least, demanding public transport fleet operation. Its importance is highlighted by the numerous evidence showing that start-and-stop driving patterns speed up the lubricant aging, and consequently, more intelligent and efficient maintenance policies are needed.

Conventional methods for monitoring oils pay attention to physicochemical & elemental parameters such as viscosity, TBN (Total Base Number), TAN (Total Acid Number), Oxidation, Nitration, wear metals and other contamination. Gołębiowski *et al.* [2] and Lenza et al. [3], used techniques such as Fourier Transform Infrared spectroscopy (FTIR) and multivariate statistical analysis to establish the relationship between oil condition, as well as engine usage and operational environment. They concluded that the effect of the urban traffic operations on the oil deterioration is much greater and that the fixed mileage intervals for oil change are not quite effective.

For instance, Raposo *et al.* [4] utilized statistical and exponential smoothing for the prediction of the degradation of components like iron (Fe) wear and the increase of the oil change service led to cost savings. Similarly, Gołębiowski *et al*. [5] had shown that the trends in TBN and viscosity were more reliable indicators of oil life than distance alone. Schutz [6] and Rodrigues [7] emphasized that monitoring the levels of additives, soot, silicates is critical for maintenance prevention of mechanical failure and reliability.

Recently, new analytical data-based methodologies have been developed to extend the capabilities of conventional methods. To seek the characteristics of oil analysis data and make way for predictive maintenance, Artificial Intelligence (AI) such as Machine Learning (ML) and Principal Component Analysis

(PCA) have been employed. Raposo *et al.* [8] proposed to develop prediction models using time series of oil pollutants (e.g., soot) and provided better precision of degradation forecasts. Other works, such as [9] and Camba, *et al.* [10], Artificial Neural Networks (ANN) and PCA, which are then combined with correlation and change point analysis in order to optimize oil change intervals, as well as early warning signs for system degradation. Moreover, Sousa [11] built models to predict the oil degradation based on the emissions.

In line with these works, Syed [11] developed an approach for predictive maintenance specially designed for school bus diesel powered fleets. His research utilized real-time telematics and machine learning algorithms to track emissions, idle time, and component wear, which provided the capacity to intervene preemptively and thereby decrease total maintenance cost. Pilot deployments in Los Angeles showed the proposed system could operate successfully with notable improvements in operation and the environment. Through detailed high-resolution GPS data, engine logs and environmental telemetry, the approach allows self-diagnostic and scalable analyses on other categories of vehicles (e.g. emergency or public transport fleet).

Additionally, Tormos *et al.* [12] proposed an explainable Artificial Intelligence (AI) (XAI) methodology for urban buses that identifies the underlying features responsible for diesel particulate filter (DPF) regeneration. Based on the OBDII data and combination of machine learning classifiers (XGBoost, LightGBM, Random Forest and SVM), they obtained a precise classification of DPF statuses and used SHAP values and partial dependence plots to interpret the influence mechanism of the operating conditions. The inclusion of XAI brought transparency and practice advice to maintenance personnel in order to understand and optimize the regeneration cycles, thus minimizing the risk of early oil aging and improving system reliability.

There is a clear trend toward hybrid methodologies that integrate conventional oil analysis techniques with machine learning and statistical modeling approaches. These integrated models provide deeper insights into oil degradation processes and assist in making more accurate and data-driven maintenance decisions. However, the combined application of these techniques in the context of urban bus fleet systems remains insufficiently explored in the current literature gap that this paper seeks to address. Furthermore, most existing research tends to focus on fuel consumption prediction using artificial intelligence [13][14], while aspects such as oil degradation and predictive maintenance receive considerably less attention.

3 Methodology

The study relies on a database obtained after condition monitoring of lubricating oils from vehicles of urban and interurban bus transport in Slovakia. These bus models include Autosan, Iveco, Solaris and SOR running on FPT Industrial, Cummins and Iveco engines. A total of 165 used oil samples, originating from different brands (*Lukoil, Urania* and *Orlen*) as well as different viscosities (15W40, 10W40 and 5W30) were initially gathered, like showed in Table 1.

Table 1: Distribution of Lubricating Oil Samples

Oil Type	Number of samples
Lukoil Advantg Ultra 15W40	31
Lukoil Advantg XLA 10W40	58
Urania Ecosynth 10W40	15
Lukoil Advantg Prof 5W30	9
Urania FE LS 5W30	31
Orlen Oil Platinum Extreme 10W40	21
Total	165

Each sample is accompanied by detailed data such as oil type, vehicle model, production year, mileage

and includes a wide range of physicochemical parameters. These parameters measure the health of the oil and evaluate its performance: Viscosity (at 40°C and 100°C), Viscosity index (VI), Total Base Number (TBN), Total Acid Number (TAN), Oxidation, Nitration, Sulfation, Several contamination chemicals (soot, diesel fuel, water, ethylene glycol), wear metals (such as iron, copper, lead) and additive elements (for example calcium, zinc, molybdenum) are also monitored, which gives a more complete picture of oil degradation and engine wear. The distribution of the oil samples is shown in Table 2.

To avoid heterogeneity and formulation variability, the present work is restricted to 58 samples of *Lukoil Advantg XLA* 10W40, a semi-synthetic oil widely used in high performance diesel engines. The lubricant has a moderate viscosity index (133.5) with high oxidation and nitration (20.7 and 21.3 A/cm) and a high molybdenum content (388.3 ppm), which implies superior wear protection properties.

Table 2 - Summary of Oil Sample Data (Lukoil 10W40)

Sample data		
Parameter	**Value**	
Vehicle Model	Solaris Urbino, SOR C, Sor CN 10,5	
Engine	FPT Industrial SPA, Cummins ISB6, FPT Industrial Tector 7, Iveco	
Year of production	$\mu = 2019$	
Overall Mileage [km]	296037.38	
Mileage on oil [km]	315542.28	
Oil properties		
	μ	**σ**
Viscosity 40ºC	113.9	19.31
Viscosity 100ºC	13.7	1.15
TBN	4.4	1.50
TAN	2.4	0.68
Oxidation	20.7	8.71
Nitration	21.3	7.66
Sulfation	15.6	4.37
Phos. Antiwear	-13.2	2.03
Pollutions		
	μ	**σ**
Diesel	0.0	0.24
Soot	0.5	0.28
Water	0.1	0.03
Ethyl. Glycol	0.4	0.12
Contaminants		
	μ	**σ**
Coolant leaks	11.7	5.45

Abrasive metals		
	μ	**σ**
Content in Mn	0.9	0.85
Content in Cr	1.4	0.48
Content in Cu	10.8	32.66
Content in Fe	69.8	25.87
Content in Ni	0.6	0.22
Content in Pb	0.8	0.82
Additives		
	μ	**σ**
Ca	1151.1	196.22
Mo	388.3	19.62
P	923.6	71.81
S	2671.9	243.74
Zn	1239.9	90.21

PCA was used to project the data into a reduced dimension subspace that maximally preserved the variance. The PCA technique allows to simplify interpretation of complex multivariate input data by transforming correlated input variables into a collection of uncorrelated linear combinations (called principal components). The first several components usually account for most of the variance so, it can be visualized and clustered more effectively.

After dimensionality reduction, the dataset was segmented through K-Means clustering, an unsupervised learning method that separates data into clusters, minimizing the variance inside clusters. The number of ideal clusters was chosen based on silhouette scores as well as visual evaluation of separation of clusters in PCA space.

A Single-Layer Perceptron (SLP) model was in addition implemented to support the unsupervised technique by classifying oil sample degradation clusters. Perceptron is a kind of supervised neural network which uses iterative optimization to learn a set of weight coefficients and assigns class labels to data. Despite more advanced deep learning models, SLP has been selected due to its simplicity and interpretability, and especially because interpretability plays an important role in operational decision-making.

These methodologies were selected for their ability to provide both a proactive assessment of potential failure risks and a data-driven classification of oil degradation profiles, enabling a comprehensive evaluation of vehicle reliability and maintenance needs.

4 Results

To uncover hidden degradation patterns, PCA was first applied to the dataset. The first five components captured 71.15% of total variance, enabling meaningful dimensionality reduction while retaining the key data structure. The explained variance of each principal component is detailed in Table 3.

Table 3 - Distribution of Explained Variance Across Principal Components

PCA Component	Explained Variance (%)
PC1	23.95%
PC2	19.62%
PC3	12.03%
PC4	8.08%
PC5	7.46%
Total	71.15%

The loadings analysis, shown in Figure 1, revealed that variables related to additives (Zn, P, S), wear metals (Cu, Fe, Mn), and chemical degradation (TAN, nitration) were the most influential in the data's variance structure.

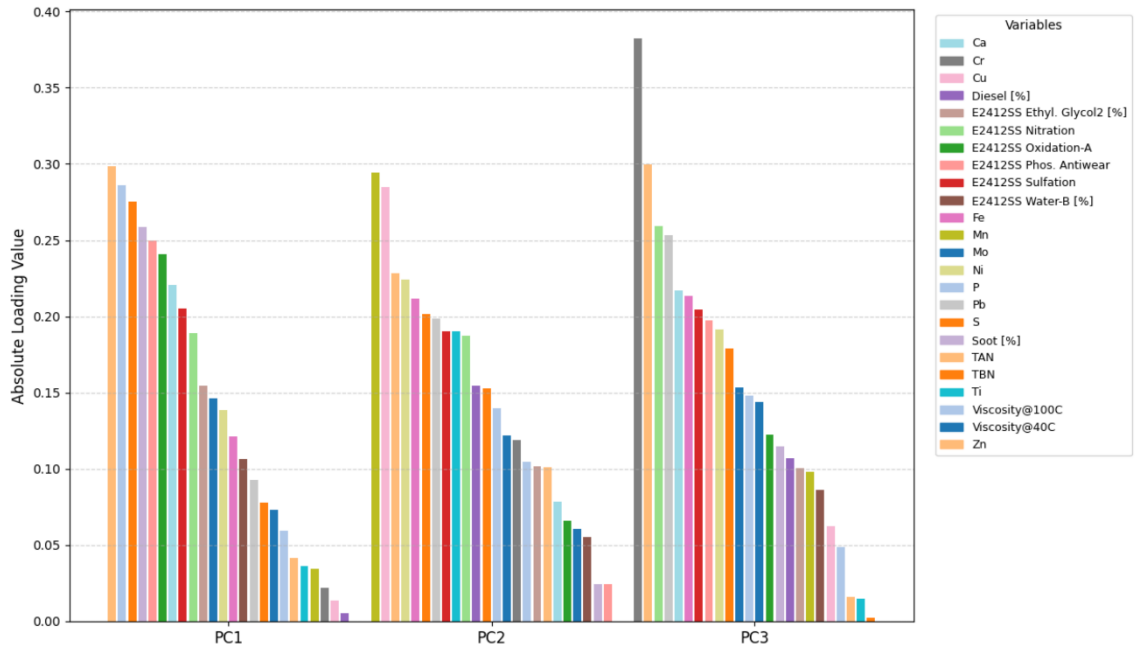


Figure 1 - Loading values of variables per Principal Component

Principal component 1 (PC1) have zinc (Zn) and phosphorus (P) with the highest contributions of 0.298 and 0.286, respectively, followed by sulfur (S) with 0.274, and soot with 0.261. These parameters are mainly related to the additives content and to the combustion product. Elements with the highest loading factors to the second principal component (PC2) included Mn with 0.296, Cu with 0.287, TAN (total acid number) with 0.224 and Ni with 0.222, indicating metallic wear and chemical degradation wear indicators. The third principal component (PC3) was dominated by Cr (0.384), TAN (0.297), Nitration (0.256), Pb (0.255), and Ca (0.255) showing a mixed influence of wear, contamination, and additives depletion.

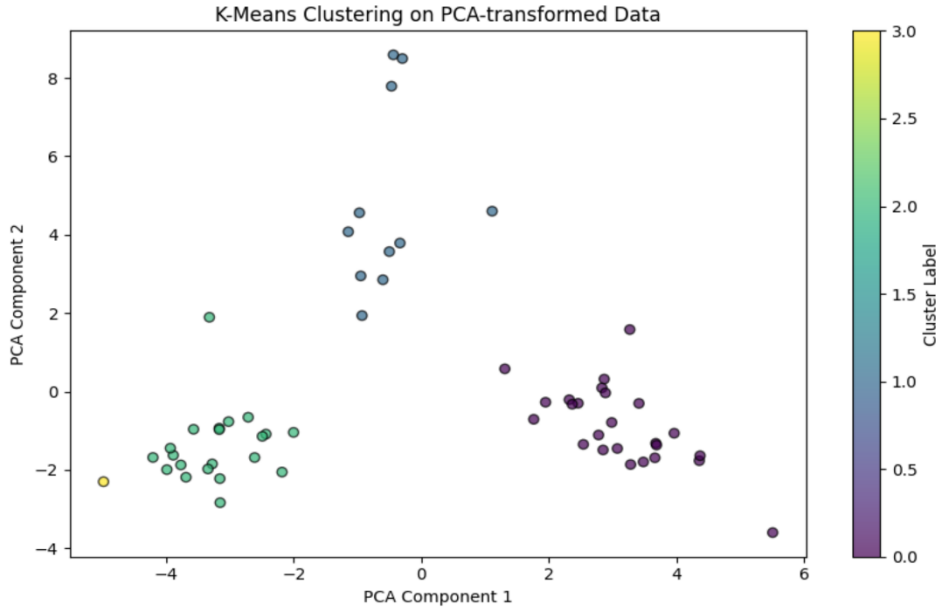


Figure 2 - Clustering Results (with K-Means) in PCA Space

Subsequently, K-Means clustering was applied to the reduced data. Figure 2 displays the results of K-Means clustering, resulting in four distinct clusters, with a silhouette score like 0.4932, which means a moderate level of separation between groups. Each dot represents an oil sample, colored by its assigned cluster. The clear separation between groups indicates that K-Means effectively identified distinct degradation patterns in the PCA-reduced dataset.

Table 4 - Cluster profiles identified by K-means clustering

Cluster	Number of records	Mileage on oil (mean)	Overall mileage (mean)	Year of production (bus)			Engine	Operating conditions
				mean	Oldest bus	Newest bus		
0	25	32222.08	270413.60	2019	2018	2020	Cummins ISB6	Urban
1	11	29909.09	65138.55	2022	2022	2023	FPT Industrial Tector 7	Interurban
2	21	31661.90	425443.05	2017	2016	2022	FPT Industrial S.P.A, FPT Industrial Tector 7	Interurban
3	1	30000.00	759000.00	2009	2009	2009	Iveco	Interurban

Table 4 shows the main properties for each cluster identified using K-Means clustering, including number of samples, average oil mileage, total vehicle mileage, year of production, and type of engine. Although no single variable fully explains the segmentation, some patterns emerge that suggest underlying influencing factors. Cluster 1 includes newer vehicles (2022–2023) with low mileage and FPT Tector 7 engines. Cluster 0 consists of moderately old buses (2018–2020) which are operated with Cummins ISB6 engines. Cluster 2 has

a more heterogeneous composition, including a larger variety of FPT powered buses from the wider production years (2016–2022) and with higher mileage. On the other extreme, Cluster 3 is an absolute outlier, which stands for one aged IVECO vehicle with the highest accumulated mileage.

To complement the unsupervised segmentation of oil samples, a supervised classification model based on a Single-Layer Perceptron (SLP) was implemented to enhance the interpretability of cluster formation. The SLP, trained to classify data points into the defined clusters, achieved 100% accuracy, confirming its suitability for this task. The connection weight method was then applied to interpret the model, enabling the identification of the most influential input variables for each cluster. This approach provided valuable insights into the key factors distinguishing the groups of oil samples and supported a more transparent understanding of engine condition patterns, essential for predictive maintenance strategies in urban bus fleets.

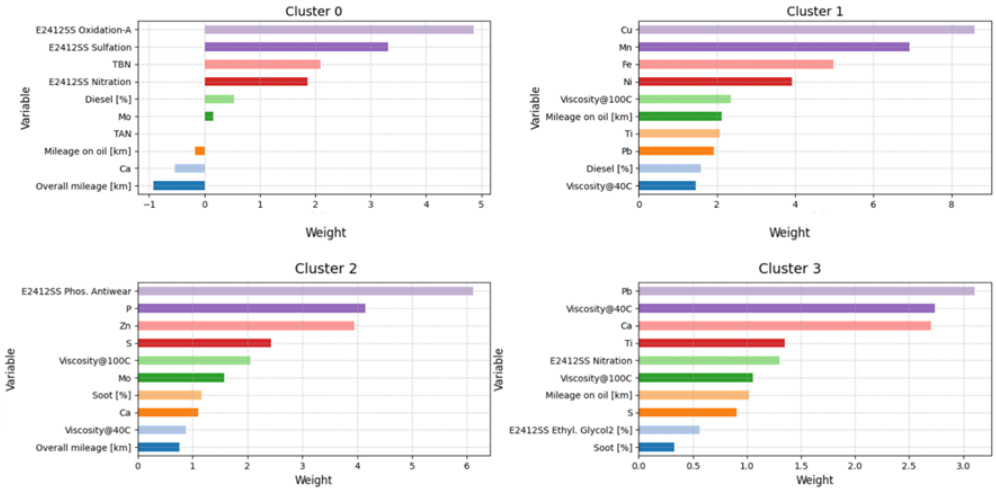


Figure 3 - Importance of variables per cluster based on Perceptron Input Weights

The perceptron weight analysis, shown in Figure 3, highlights the most influential variables for each cluster. These key factors help characterize the distinct oil degradation profiles previously identified, enhancing the interpretation of the clusters and supporting more targeted maintenance decisions.

The analysis confirmed that each cluster has a distinct signature defined by key variables. Cluster 0 is dominated by oxidation-related indicators, suggesting chemical degradation even in low-mileage vehicles. Cluster 1 is characterized by high levels of wear metals and fuel dilution, pointing to mechanical wear and internal faults. Cluster 2 reflects oils in good condition, marked by preserved additives and low contamination, typical of early-life or well-maintained engines. Cluster 3 shows severe, multifactorial degradation, associated with coolant contamination, and soot accumulation. Thus, we can consider cluster 3 as an outlier.

5 Discussion

The results herein prove how powerful is the combination of machine learning and oil condition monitoring for predictive maintenance of urban bus fleets. Because the above, PCA, K-Means, and SLP could simplify the data and help to obtain clean oil degradation patterns based on the various kinds of operating conditions.

The clustering in four classes led to the identification of four clearly distinguished oil degradation profiles: chemical one due to the oxidation (cluster 0), heavy mechanical wear sustained degradation with fuel dilution (cluster 1), in good mechanical conditions oil with active additives still active (cluster 2), strongly degraded oil characterized by severe multifactorial damages due to potential cooling system failure and contamination from

combustion by-products (cluster 3).

The supervised model used (Perceptron) supported the conclusions of the unsupervised analyses by providing a clear interpretation of the most relevant variables in defining each cluster and its diagnostic potential regarding lubrication state. At the same time, it also empowers the case for CBM, having the ability to be used to make better interventions in line with the engine's actual condition rather than just clocked mileage.

This paper helps to close the gap using artificial intelligence to oil degradation with an implementable, interpretable, and applicable methodology that can be integrated into maintenance management systems.

Furthermore, the findings point towards the fact that substantial oil degradation can be experienced in vehicles with relatively low mileage users, again reinforcing the importance of the ongoing monitoring the oil condition as an important tool to assist in maintaining the health of the operational fleet.

Finally, as the dataset considered was for only one oil formulation (*Lukoil Advantg XLA* 10W40), the approach taken can be applied to other formulas, providing the parameters considered are the same.

6 Conclusion

This study demonstrated the effectiveness of combining oil analysis with machine learning techniques for predictive maintenance in urban bus fleets. By applying PCA, K-Means, and a Single-Layer Perceptron, it was possible to identify distinct oil degradation profiles and key influencing variables, enabling more accurate diagnostics and maintenance planning.

The results confirmed that oil degradation can occur even in low-mileage vehicles, highlighting the limitations of time or distance-based maintenance strategies. The proposed interpretable and data-driven approach supports the implementation of Condition-Based Maintenance (CBM) and offers a scalable solution for integration into digital fleet management systems.

In future studies this approach could be applied to other types of oils as well as to different kind of engines considering operational parameters in real time like telemetry and driving style in order to improve the prediction models.

References

[1] NP EN 13306, “Norma Portuguesa - NP EN 13306: 2021 - Manutenção Terminologia de manutenção,” *Inst. Port. da Qual.*, pp. 1–44, 2021.

[2] W. Gołębiowski, A. Wolak, and B. Šarkan, “Engine Oil Degradation in the Real-World Bus Fleet Test Based on Two Consecutive Operational Intervals,” *Lubricants*, vol. 12, no. 3, 2024, doi: 10.3390/lubricants12030101.

[3] T. L. L. Lenza, A. Ruggiero, A. Senatore, and P. Siano, “Some results on the used lubricants analysis of urban bus diesel engines,” *Adv. Transp.*, vol. 16, pp. 541–550, 2004.

[4] H. Raposo, J. T. Farinha, I. Fonseca, and L. A. Ferreira, “Condition monitoring with prediction based on diesel engine oil analysis: A case study for urban buses,” *Actuators*, vol. 8, no. 1, 2019, doi: 10.3390/act8010014.

[5] W. Gołębiowski, G. Zając, M. Sejkorová, and A. Wolak, “Assessment of oil change intervals in urban buses based on the selected physicochemical properties of used engine oils,” *Combust. Engines*, vol. 196, no. 1, pp. 15–23, 2024, doi: 10.19206/CE-169807.

[6] DIONLENO BORGES SCHUTZ, “UTILIZAÇÃO DA ANÁLISE DE ÓLEO LUBRIFICANTE COMO FERRAMENTA DA ENGENHARIA DE MANUTENÇÃO,” Universidade Federal do Rio Grande do Sul, 2022. [Online]. Available: http://www.ifpri.org/themes/gssp/gssp.htm%0Ahttp://files/171/Cardon - 2008 - Coaching d’équipe.pdf%0Ahttp://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203%0Ahttp://mpoc.org.my/malaysian-palm-oil-industry/%0Ahttps://doi.org/10.1080/23322039.2017

[7] J. C. Antunes Rodrigues, “Análise Multivariada de Óleos Lubrificantes,” Instituto Superior de Engenharia de Coimbra, 2019.

[8] H. Raposo and D. Galar, “Prediction condition based on oil analysis- a case study,” 2019, doi: https://doi.org/10.1016/j.triboint.2019.01.041.

[9] J. Rodrigues, I. Costa, J. T. Farinha, M. Mendes, and L. Margalho, “Predicting motor oil condition using artificial neural networks and principal component analysis | Prognozowanie stanu oleju silnikowego za pomocą sztucznych sieci neuronowych i analizy składowych głównych,” *Eksploat. i Niezawodn.*, vol. 22, no. 3, pp. 440–448, 2020.

[10] R. Ramirez Camba, C. Garcia Garcia, M. Garcia Tobar, and J. Fajardo Merchan, “An Integrated Methodological Approach for Interpreting Used Oil Analysis in Diesel Engines,” *Lubricants*, vol. 13, no. 4, 2025, doi: 10.3390/lubricants13040169.

[11] S. Syed, “Enhancing School Bus Engine Performance : Predictive Maintenance and Analytics for Sustainable Fleet Operations,” vol. 44, no. 3, pp. 17765–17775, 2024.

[12] B. Tormos, B. Pla, R. Sánchez-Márquez, and J. L. Carballo, “Explainable AI Using On-Board Diagnostics Data for Urban Buses Maintenance Management: A Study Case,” *Inf.*, vol. 16, no. 2, 2025, doi: 10.3390/info16020074.

[13] M. B. Data, “Predicting Gasoline Vehicle Fuel Consumption in Energy and,” 2022.

[14] A. Lorenc, “Predicting Fuel Consumption by Artificial Neural Network (ANN) Based on the Regular City Bus Lines,” *Sustain.*, vol. 17, no. 4, 2025, doi: 10.3390/su17041678.

Chapter 10

Analysis of the state of a plastic injection machine

Contribution #30
Cite as:
Joao Costa, Rui Silva, Goncalo Martins, Jorge Barreiros, Mateus Mendes. Analysis of the state of a plastic injection machine. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-30,
  title ={Analysis of the state of a plastic injection
     machine},
  author={João Costa and  Rui Silva and  Gonçalo
     Martins and  Jorge Barreiros and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Analysis of the state of a plastic injection machine

João Costa[1], Rui Silva[2], Gonçalo Martins[2], Jorge Barreiros[1,3], Mateus Mendes[1,3,4,*]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[2] Sinmetro LDA, Rua dos Costas, Lote 19, Loja 74, R/C 2415-567 Leiria
[3] RCM^{2+} Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[4] Institute of Systems and Robotics,
Department of Electrical and Computer Engineering, University of Coimbra,
3030-290 Coimbra, Portugal
[*] Corresponding author
mmendes@isec.pt jmsousa@isec.pt rui.silva@sinmetro.pt gmartins@sinmetro.pt

Abstract

Predictive maintenance is essential for minimizing unplanned downtime and optimizing industrial processes. In the case of plastic injection molding machines, failures that lead to downtime cause costly delays. As industrialization advances, proactive equipment management enhances cost efficiency, reliability, and operational continuity.

This study aims to detect failures before they occur using sensors, data science techniques, and statistical analysis. A case study was conducted, including machine characterization and multi-source data collection. Clustering methods identified operational patterns and anomalies, classifying machine behavior into distinct states.

By applying unsupervised learning, we developed an early anomaly detection methodology, enabling timely interventions to prevent failures. Results show that clustering techniques effectively differentiate normal from abnormal conditions, offering a robust data-driven predictive maintenance approach. This method improves failure prediction accuracy and optimizes maintenance strategies, enhancing equipment reliability and efficiency.

These findings advance the state of the art by providing a scalable, proactive monitoring solution that supports high operational standards in modern manufacturing. Implementing this methodology helps to reduce unplanned downtime, increase machine reliability, and improve production efficiency.

Keywords: Predictive maintenance; Plastic Injection Machine; Anomaly detection

1 Introduction

Unplanned downtime in industrial machinery poses a significant challenge to manufacturing efficiency. Unexpected machine failures can lead to severe financial losses, production delays, and increased maintenance costs. In plastic injection molding, where precision and continuity are crucial, predicting and preventing such failures is essential to maintain productivity and reduce operational risks. Failures in these machines can stem from various sources, including mechanical wear, electrical faults and human errors, making early detection a complex but necessary task.

Plastic Injection Molding (PIM) machines are heavy industrial equipment that require specialized maintenance interventions. Ideally the operate during many hours, days or even weeks, in order to maximize production. Nonetheless, they can suffer numerous problems that require qualified maintenance interventions. A number of possible problems and solutions are discussed below. The problems include fixed plate deformation, obstructions in the injection system, pressure and temperature variations, mold cooling failures, and other common challenges in this type of equipment.

Recent advancements in data science and machine learning have enabled the development of predictive maintenance strategies aimed at reducing unplanned stoppages. Clustering techniques and anomaly detection have shown promising results in identifying patterns associated with machine failures, allowing for early interventions. However, despite the growing body of research, many industrial applications still rely on reactive maintenance, leading to inefficiencies and high costs. This study aims to bridge this gap by applying data-driven clustering methods to detect failures in a plastic injection machine before they occur.

The main objectives of this research are: (1) to analyze machine behavior through clustering techniques, (2) to classify operational states and detect early warning signs of failure, and (3) to evaluate the effectiveness of these methods in a real-world industrial setting. The study follows a structured methodology, starting with data collection and machine characterization, followed by clustering analysis and validation of failure detection models.

The remainder of this paper is organized as follows: Section 3 presents a review of the state of the art. Section 3.2 describes what is predictive maintenance and its importance in industrial settings. Section 3.3 describes some data mining techniques used in predictive maintenance models. Section 3.4 present a study on a machine learning model designed to predict failures in plastic injection machines, as well as techniques for enhancing the dataset. Section 4 details the methodology used in the study. Section 4.1 describes the original dataset and the dataset utilized in the study, along with the statistics and illustrative graphics. Section 4.2 presents the techniques and technologies employed to enhance the dataset. Section 4.3 provides an overview of Principal Component Analysis (PCA) and its application in the study. Section 4.4 explains the functioning of DBSCAN and its application in the study. Section 5 discusses the results, and Section 7 presents conclusions and future research directions.

2 Plastic Injection Machines

Figure 2 shows a diagram of a PIM and the main constituting parts. A more detailed description of the most important parts is given in Table 2.

PIM are complex devices which require careful optimization and maintenance to operate smoothly and safely. According to industry operators who perform daily maintenance on injection molding machines, a number of issues are common [2].

One of the most common problems in PIM machines is the obstruction of the injection devices due to plastic residue accumulation or contaminants. This issue may arise from improper cleaning procedures, low-quality materials, or incorrect processing conditions. Additionally, incorrect material selection and improper processing parameters can lead to material degradation, overheating, and nozzle blockage. Regular maintenance and cleaning of injection devices are essential to prevent obstructions. Operators should use appropriate cleaning agents to remove accumulated residues. Using compatible materials and optimizing processing parameters also significantly reduces obstruction risks and enhances machine efficiency.

The Mold Cooling System is another sensitive part. Inefficient mold cooling can result from poor cooling channel design and inadequate heat dissipation capacity, leading to uneven tem-

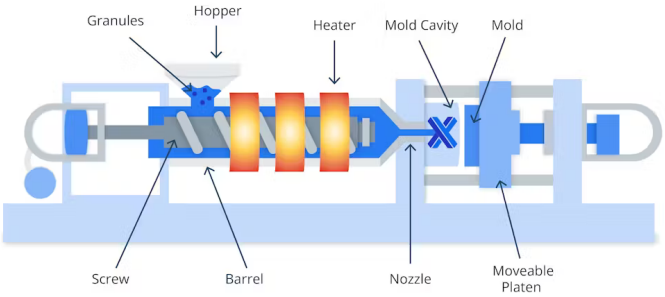


Figure 1: Diagram of a Plastic Injection Machine [1]

Table 1: Name, location, function and appearance of the most important parts of a PIM

Name	**Location**	**Function**	**Appearance**
Barrel	In the middle of the machine, surrounds the screw	Captures and mixes the plastic, maintaining uniformity and pressure during injection	Long and cylindrical; pointed at one end; covered by heating bands
Gates	Between channels and mold cavity	Control the flow of plastic going into the cavity	Small openings
Heaters	Around the barrel	Heats the barrel to melt the plastic	Metal bands around the barrel
Hopper	At the top of the machine	The place where plastic is introduced in its solid form. Sometimes contains a dryer to remove moisture	Funnel (Conical Shape)
Mold	Connected to the Fixed and Movable Platens	Gives the final shape to the molten plastic, forming the desired part	Metal block, typically two parts, containing a cavity, cooling channels, and vents
Mold Cavity	Inside the mold	Creates the final shape of the mold and contains cooling cavities	A space in the mold that forms the desired shape
Movable Platen	Connected to one half of the mold	Presses one half of the mold against the other during part manufacturing and releases it once the part is finished and cooled	Flat, rectangular, and metallic
Nozzle	At the end of the barrel, near the mold	Directs the plastic into the mold cavity and prevents it from cooling before entering the mold	Tapered outlet
Pellets	Inside the barrel and nozzle	Plastic material inserted into the machine for molding. Common plastics include ABS, PP, or Nylon, sometimes with additives	Small plastic granules
Reciprocating Screw	Inside the barrel	Mixes and compresses the plastic coming from the hopper	A metal spiral
Runners	Inside the mold	Directs the molten plastic, maintains a uniform flow, and reduces plastic waste	Long and narrow channels
Sprue	Central channel that connects the nozzle to the entry point of the molten plastic	Directs the molten plastic from the nozzle to the runners	Cone-shaped channel

perature distribution, extended cycle times, higher defect risks, and reduced product quality. Optimizing the layout and configuration of cooling channels ensures uniform temperature distribution throughout the mold. Proper positioning, consistent channel diameters, and appropriate spacing improve cooling efficiency, leading to better productivity and higher-quality injected parts.

Pressure and Temperature Variations in Injection can also be a source of problems. Pressure variations in the injection process are influenced by changes in material viscosity and polymer flow behavior. Temperature fluctuations and humidity content can alter material flow properties, causing injection pressure instability. Controlling material properties through proper storage, handling, and humidity monitoring is critical. Injection parameters must be adjusted based on material type, product shape, and mold design. Proper pressure calibration minimizes variations and ensures consistent production quality.

Part Adhesion and Removal Issues are another typical problem. Part adhesion to the mold can result from inadequate use of mold release agents. Insufficient extraction force or poorly designed ejector pins may lead to deformation or incomplete removal of molded parts. Applying mold release agents correctly and optimizing surface finish reduce friction and facilitate part removal. Adjusting extraction force and modifying ejector pin design ensures efficient part ejection, minimizing defects and material waste.

The Hydraulic System can also cause frequent failures. Hydraulic failures, including oil leaks due to worn seals, damaged hoses, or loose connections, can significantly impact machine performance. Insufficient hydraulic pressure caused by pump failures, valve blockages, or fluid contamination also disrupts operation. Regular hydraulic system inspections prevent oil leaks and ensure all connections are secure. Monitoring pressure gauges and performing preventive maintenance on pumps, valves, and filters, help maintain optimal hydraulic performance and avoid machine failures.

The Electrical and Control System must also be monitored to prevent potential failures Electrical issues, such as power fluctuations or wiring failures, can cause unexpected machine shutdowns and production delays. Malfunctions in the control system, including software or hardware issues, can impact machine performance and process regulation. Installing surge protectors, voltage regulators, and uninterruptible power supplies (UPS) stabilizes power supply and prevents sudden failures. Routine inspections of electrical wiring, terminals, and connectors, help prevent malfunctions. Diagnosing and resolving control system issues in advance ensures stable operation and high-quality production.

By addressing these issues through proper maintenance and optimization strategies, the efficiency, durability, and quality of plastic injection molding machines can be significantly improved.

3 Literature Review

A comprehensive literature review was conducted, searching scientific databases such as Scopus, IEEE Xplore, and ScienceDirect. Keywords used included “plastic injection molding failures,” “fixed plate optimization in PIM,” “mold cooling efficiency,” and “hydraulic system failures in injection molding machines.” The papers to review were selected based on their relevance in improving machine performance, efficiency, and durability. Before presenting these findings, a description of the machine will be provided, along with a figure for better comprehension.

3.1 Plastic Injection Optimization

Shu Sun [3] studied the optimization of the fixed plate structure in a plastic injection molding machine, aiming to reduce deformation and minimize the fatigue of tie bars. The study focuses

on reducing the deformation of the fixed plate caused by compressive forces applied by the moving plate and the plastic injection process. Over prolonged operations, this deformation leads to increased stress on the tie bars, causing fatigue and premature wear. The solution proposed employs a self-organization method that redistributes material efficiently, reinforcing areas under high stress while minimizing material waste. This method adjusts the Young's modulus of structural elements according to the applied forces, improving resistance without significantly increasing weight or cost. After 52 optimization iterations, the study demonstrated a significant reduction in fixed plate deformation, increasing the lifespan of tie bars. The results suggest that this method can be applied to other structural components in industrial machines to enhance their efficiency and durability.

3.2 Predictive Maintenance in Plastic Injection Machine

Pierleoni *et al.* [4] discuss the evolution of maintenance strategies in industrial settings, emphasizing the importance of Predictive Maintenance (PdM) within Industry 4.0. Effective maintenance management is critical to avoid unexpected failures, which can result in high costs and negatively impact product quality and system reliability. Despite its importance, many companies have yet to adopt advanced strategies to optimize their maintenance budgets. The study focuses on the application of PdM in four electric plastic injection molding machines equipped for Industry 4.0. Analyzing failure patterns and monitoring process variables, a methodology was developed to predict component wear and prevent complete breakdowns. Data were collected from machine sensors, measuring variables such as injection pressure, plasticization volume, cycle time, temperature, and motor force. Since most companies follow preventive maintenance strategies that avoid obvious failures, obtaining real defect data was challenging. Instead, failures were inferred based on qualitative analysis of historical data and expert interviews. The collected data were labeled into two categories: 'optimal' operation and 'functional limit,' allowing the development of predictive models for twelve different machine-product combinations.

Seven machine learning classification models were tested: Fine Tree, Medium Tree, Coarse Tree, Fine K-Nearest Neighbor, Medium K-Nearest Neighbor, Coarse K-Nearest Neighbor, and Logistic Regression. A 10-fold cross-validation method was used to prevent overfitting. Models were evaluated using accuracy, sensitivity, and specificity. At least one model for each of the twelve scenarios achieved over 90% performance across most metrics. During testing, Fine Tree models with 100 split points were applied to new machine data. While accuracy and sensitivity remained similar to cross-validation results, specificity declined, likely due to overfitting. A key limitation of the study is the scarcity of real failure data, which affects the accuracy of adverse condition classifications.

3.3 Data Mining for Fault Diagnosis

Kozjek *et al.* [5] explore the application of data mining techniques for fault diagnosis in PIM processes. The study examines how algorithms such as J48, Random Forest, and k-Nearest Neighbors can identify operational failure patterns, aiming to improve production efficiency and maintenance planning. This study demonstrates how Data Mining (DM) can uncover defective operational patterns and improve quality and productivity in PIM. DM provides an alternative to traditional methods like statistical process control and experimental design by discovering significant patterns and relationships in industrial data.The dataset consists of six months of operational records from five European PIM machines, including Process parameter logs input and output values per cycle, Alarm logs, and Tool change records. Approximately 2.2 million cycles were recorded, with over ten different tools used per machine. Alarms were classified into

First-degree alarms (immediate machine stoppage) or Second-degree alarms (non-critical issues). The final dataset contained 62 numerical attributes and a binary classification label. To ensure an unbiased classification baseline of 50%, the dataset included equal instances of normal and faulty cycles. A Python-based system was developed for data processing, including reading, encoding, filtering, transformation, and querying. / Classification models were implemented using the Weka tool, with 10-fold cross-validation. The results indicate that the J48, Random Forest, JRip, Naïve Bayes, and k-NN algorithms effectively identify patterns related to defective operating conditions. All tested algorithms outperformed the standard accuracy benchmark of 50.0%. J48, Random Forest, JRip, and Naïve Bayes exhibited higher classification accuracy than k-NN, as they leverage target attribute information during model induction. Random Forest offers the advantage of easy parameter tuning and relatively high predictive performance. However, its interpretability is limited, whereas J48 and JRip provide interpretable rule-based models. Key parameters affecting defective operating conditions in UMSs for a selected tool and machine include temperature, opening time, and cycle time.

3.4 Fault Estimation in PIM

Aslantaş *et al.* [6] propose a machine learning model for identifying and predicting failures in plastic injection molding (PIM) machines. They research on sensor data analysis to anticipate failures before they occur, enabling more efficient maintenance planning and preventing unexpected production downtimes. To achieve this goal, classification algorithms such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were applied to data collected from three machines in a home appliance manufacturing plant. Additionally, the SMOTE technique was employed to balance data distribution and enhance model accuracy. This study introduces a machine learning model designed to predict failure types in PIM machines based on sensor data. The model is constructed using classification algorithms to analyze sensor readings and forecast machine failures before they occur. Failure identification in PIM machines follows multiple stages, from raw data collection to classification. The estimation of Remaining Useful Life (RUL) plays a crucial role in identifying machines with a higher likelihood of failure. RUL, defined as the time interval between the present moment and the point of failure or maintenance requirement, is computed using historical maintenance and failure data. Several factors, such as clamping force, cycle time, and oil temperature, influence this interval. These data are obtained through sensors, process parameter logs, alarms, and maintenance records. Two classification algorithms were employed to develop predictive maintenance models: Random Forest and XGBoost. Raw sensor data alone do not sufficiently describe failure types, making feature extraction a crucial step. Extracted features include minimum, maximum, mean, skewness, kurtosis, and entropy. In this study, raw data were aggregated into 60-minute intervals, and features were automatically extracted for each interval. Data were collected from three plastic injection molding machines (HEP3204, HEP3207, and HEP3213) in a Turkish home appliance factory. The dataset sizes range from 205,000 to 913,000 records, covering the period from 2018 to 2021. A separate predictive model was constructed for each machine using variables such as clamping force, cycle time, and oil temperature. Experimental results indicate that RF and XGBoost, with and without SMOTE, demonstrated strong performances. XGBoost with SMOTE achieved the best performance, with an average accuracy of 98%. The highest F-score was also obtained using XGBoost with SMOTE (0.98), compared to RF (0.88) and other variations. This study estimated the types of failures in plastic injection molding machines, addressing the issue as a multi-class classification problem for predictive maintenance. Handling missing values and extracting relevant features were crucial steps in the process. The XGBoost and Random Forest (RF) algorithms were evaluated, with XGBoost demonstrating superior performance, especially

Table 2: Description of variables with their respective statistical values.

Variable	Number of values	Mean	Max	Min	Std. Dev.
Specific_injection_pressure_peak_value	170645	1710.18	2071.20	0.0	363.77
Switchover_volume_actual_value	170642	159.31	502.76	0.0	29.71
Specific_pressure_at_switch_over	170644	1631.39	2000.60	0.0	338.75
Specific_holding_pressure_peak_value	170642	1629.60	1997.90	0.0	337.83
Material_cushion_smallest_value	170642	35.22	495.08	0.0	23.65
Material_cushion_after_holding_pressure	170643	55.38	697.64	0.0	25.63
Material_cushion_end_holding_pressure	170642	36.23	538.11	0.0	28.06
Shot_volume	170644	991.61	1096.76	0.0	161.45
Injection_time	170644	3.72	20.50	0.0	0.42
Speed_peak_value	170642	0.43	0.65	0.0	0.06
Specific_back_pressure_peak_value	170644	78.77	301.30	0.0	10.78
Plasticizing_volume	170642	1049.19	1132.39	0.0	161.97
Plasticizing_time	170644	14.83	330.21	0.0	2.76
Clamping_force_peak_value	170644	5077.88	5255.10	0.0	465.12
Mold_opening_stroke_peak_value	170642	644.68	652.20	0.0	57.50
Cycle_time	170644	59.11	2335.19	0.0	25.31
Cooling_time	170642	24.48	333.04	0.0	3.57
Cycle_time_holding_pressure	170641	14.41	20.00	0.0	2.13
Barrel_Temperature_Zone_Actual_Temperatures_INJ1_Z01	170641	200.39	280.20	22.90	27.33

when combined with the SMOTE technique. The achieved accuracy was 98% with XGBoost + SMOTE.

4 Data and Methodology

In this study, an unsupervised learning approach was employed to analyze the dataset, using Principal Component Analysis (PCA) for dimensionality reduction and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for clustering. Unsupervised learning is a type of machine learning that analyzes data without human supervision. Unlike supervised learning, it works with unlabeled data, allowing models to discover patterns and insights independently. This method is particularly useful for organizing large volumes of information into clusters and identifying previously unknown patterns. The methodology followed a structured workflow involving data preprocessing, feature extraction, and clustering to uncover meaningful patterns in the data.

4.1 Dataset

The dataset consisted of sensor values recorded by the machine over time. Due to its high-dimensional nature, a preprocessing step was necessary to reduce the dataset size while preserving crucial information to improve clustering performance.

The original dataset comprised 48 million records across 129 variables and 7 machines. This was subsequently reduced to 19 critical variables, containing 3,242,214 records from a single machine, Machine 76.

This dataset encompasses records from April 2024 to January 2025, covering a sampling period of 274 days. The records were sampled at a frequency of milliseconds.

It is possible to observe in Table 2, as well as in Figures 2 and 3, the statistics of the variables used for clustering and the plots of the records for two of the variables utilized.

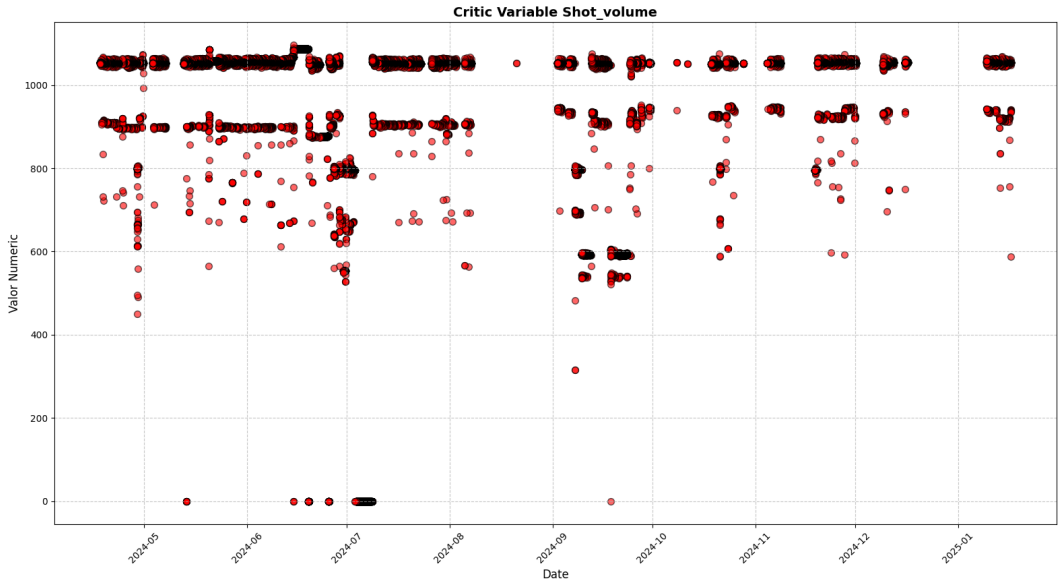


Figure 2: Critical variable analysis of shot volume for Machine 76.

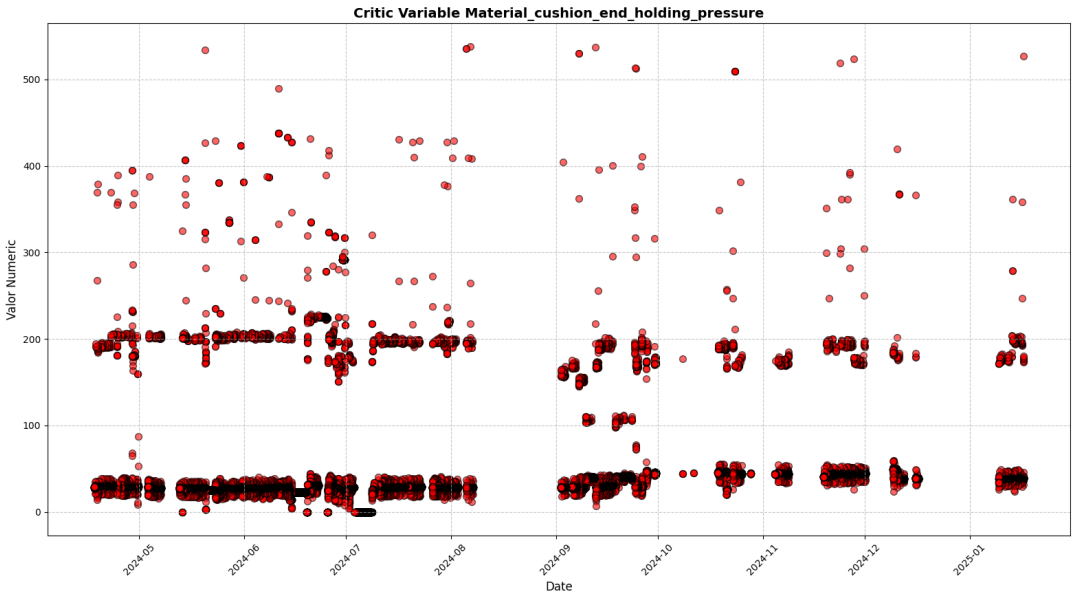


Figure 3: Critical variable analysis of material cushion end holding pressure for Machine 76.

4.2 Data Preprocessing

The dataset had areas of missing records, in all the critical variables used. As can be seen in figure 2, in some areas it is possible to see the lack of records. They values were replaced using the forward-fill (ffill) method in python pandas library. This approach was chosen because sensor values often remain constant over short periods, making the previous value a reasonable replacement.

Since PCA is sensitive to scale, all numerical features were standardized using z-score normalization, ensuring a mean of zero and a standard deviation of one. This step prevented features with large magnitudes from dominating the analysis.

4.3 Dimensionality Reduction Using PCA

Principal Component Analysis (PCA) is a statistical technique that transforms high-dimensional data into a lower-dimensional space while retaining most of its variance. This process is particularly useful in clustering, as it removes noise and redundant information, leading to improved performance. PCA identifies directions, known as principal components, along which data exhibits the most variance and projects the data onto these components.

4.4 Clustering with DBSCAN

DBSCAN is a density-based clustering algorithm that groups data points based on their density, making it well-suited for datasets with irregular cluster shapes. Unlike K-Means, DBSCAN does not require specifying the number of clusters beforehand. Instead, it identifies clusters as dense regions separated by areas of lower density.

DBSCAN assigns each data point to a cluster or labels it as noise if it does not meet the density criteria — *i.e.*, the number of neighbouring points in a given radius is less than a predefined minimum. A core point is defined as one having at least min_samples neighbors within the radius Points that did not have a minimum of neighbours within the radius are classified as noise.

5 Results

5.1 Difference Between "Mode" and "State" of the Machine

The machine has six configuration modes. However, these modes do not necessarily correspond to the possible states of the machine. This implies that the number of machine states may be either greater or smaller than the number of configuration modes. In other words, the machine may exhibit state variations that are not directly reflected in the defined modes, or conversely, different modes may result in the same machine state.

Furthermore, it is important to highlight that altering a configuration mode does not necessarily imply a change in the machine state. In certain situations, the machine may remain in the same state despite the modification of the mode, depending on operational variables or specific system conditions.

This distinction between modes and states is crucial for the correct interpretation of the results, particularly in analyzing the machine's behavior over time.

Table 3: Description of clusters, their respective colors, and meanings. ID is the label of the cluster. Number -1 refers to noise points, which are isolated and do not form any cluster. Sil is the silhouette score.

ID	Color	Description	Sil.
-1	Light Blue	*Outliers*, possible noise points or very severe faults.	79%
0	Dark Blue	Reduced or anomalous operating condition.	79%
1	Light Green	Normal operation with a 4-cavity mold.	79%
2	Dark Green	Normal operation with a mold different from the light green one.	79%
3	Pink	Machine operating with fewer than 4 cavities.	79%
4	Red	Machine stopped.	79%
5	Yellow	Anomalous operation, capturing anomalies from the pink *cluster* (3 cavities).	79%
6	Orange	Severe anomaly, but with low occurrence, possibly irrelevant.	79%

5.2 DBSCAN results

After optimization, DBSCAN identified for the machine a total of seven different states. The results show seven clusters plus noise. Table 3 shows a description of each cluster. Points assigned label -1 are not included in any cluster, so they are considered noise.

Figure 4 shows a visualization of the results of the clustering algorithm. The figure illustrates how the algorithm identifies different clusters over the time, for variable Shot Volume. Figure 5 shows the same for variable Injection Pressure. Those clusters were formed using four PCA components, which explain up to 99.5 % of the variance in the dataset.

As the figures and the table show, DBSCAN was able to successfully separate the working states of the machine. Points assigned labels -1 (noise), or 6, may require urgent attention because they refer to possible anomalous states. Cluster 5 refers to a state which also requires attention, because the machine is not operating in good condition.

Since the machine operates with a four-cavity mold, a specific cluster immediately forms, represented in light green in Figure [4]. If one of the cavities exhibits a defect or malfunction, the machine operates with three cavities. This operating condition forms a separate cluster. Another example is the replacement of the mold with one that has only a single cavity, represented in the graph by the dark green color.

The clusters with the highest potential to represent an anomalous machine state were identified as dark blue and yellow clusters. As shown in Figure [5], this cluster corresponds to values that fall above and below the ranges considered normal, thus indicating a potential irregularity in the machine's operation. Additionally, the light blue points, considered outliers, may be classified as anomalous points.

5.3 Analysis of the clusters

Through detailed analysis of the clusters, it was possible to identify that one of them is directly related to the machine's idle state. This cluster, represented in red, corresponds to the moment when the machine is in standby mode. In this state, the machine is not performing any active operations, awaiting its next task or activation. Distinguishing this cluster is essential for understanding machine inactivity periods and effectively monitoring its operational cycles.

A more detailed description of the clusters is given below. Table 4 also show some statistical information of each cluster.

- ***Points labeled* -1: *Outliers* / Possible Severe Faults**. Those points are extremely dispersed. Considered *outliers*, they potentially represent noise in the data or severe machine

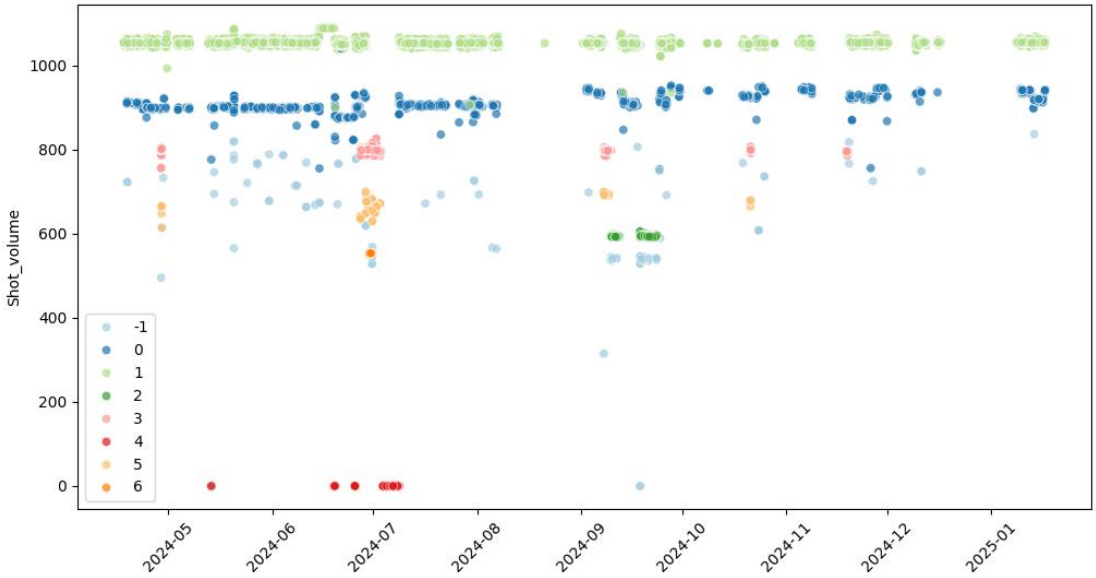


Figure 4: Illustration of DBSCAN clustering results for variable Shot Volume. There are 7 clusters with labels 0-6. Noise points are assigned label -1.

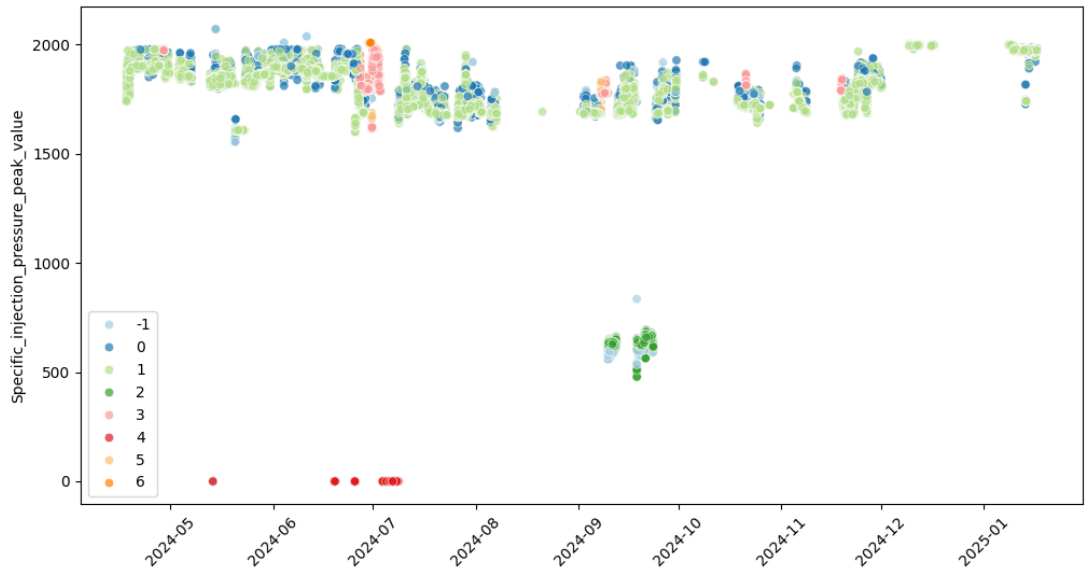


Figure 5: Illustration of DBSCAN clustering results for variable Specific Injection Pressure Peak Value. There are 7 clusters with labels 0-6. Noise points are assigned label -1.

Table 4: Statistical data of the principal components of each cluster and noise. ID is point label assigned by DBSCAN

ID	PCA1				PCA2				Description
	min	max	mean	std	min	max	mean	std	
-1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Noise. Possible Severe Faults
0	-105.31	506.25	217.51	135.43	-300.29	238.49	-43.62	77.99	Reduced or Anomalous Operation
1	-37.88	540.91	229.67	137.53	-312.02	173.47	-48.17	89.29	Normal Operation with 4-Cavity Mold
2	-1880.73	-1614.96	-1688.68	24.43	688.59	860.92	735.20	16.29	Normal Operation with 1-Cavity Mold
3	-167.88	370.22	134.89	97.70	-154.71	241.31	-20.66	52.02	Operation with Less than 4 Cavities
4	-5411.40	-5409.07	-5410.74	0.72	-2640.40	-2632.49	-2635.88	2.67	Machine Stopped
5	-210.95	332.50	107.44	131.40	-149.69	250.29	-20.42	76.00	Reduced or Anomalous Operation
6	428.60	435.49	429.91	1.03	-218.35	-203.47	-213.67	3.25	Possible Anomaly

failures.

- ***Cluster* 0: Reduced or Anomalous Operation**. There is lower variability compared to *outliers*, with PCA1 values from -105.30 to 506.25. Indicates reduced or anomalous machine operation, possibly associated with an alert condition or low operational efficiency.
- ***Cluster* 1: Normal Operation with 4-Cavity Mold**. Well-concentrated values, with a PCA1 mean of 229.67 and controlled dispersion. Represents normal machine operation with its most common 4-cavity mold, indicating smooth process running.
- **Cluster 2: Normal Operation with 1-Cavity Mold**, PCA1 values concentrated in a negative region (-1880.73 to -1614.95) with low standard deviation. Indicates normal machine operation with a different mold compared to Cluster 1. Useful for classifying different operational modes of the machine.
- ***Cluster* 3: Operation with less than 4 Cavities**. Moderate variability, PCA1 values from -167.87 to 370.21, greater dispersion in PCA3 (-462.86 to -127.14). Represents an operational state where the machine runs with fewer than 4 cavities, expected in certain production cycles.
- ***Cluster* 4: Machine Stopped**. Extremely concentrated values in PCA1 (-5411.39 to -5409.07) and PCA2 (-2640.39 to -2632.49), with negligible standard deviation. Indicates that the machine is stopped, with no significant operational variations.
- ***Cluster* 5: Reduced or Anomalous Operation**. Intermediate values, PCA1 ranging from -210.95 to 332.50, with controlled dispersion. Suggests anomalous machine operation, representing low production or failure state associated with specific conditions.
- ***Cluster* 6: Possible Anomaly**. Values concentrated in PCA1 (428.60 to 435.49), slightly negative variation in PCA2 (-218.35 to -203.47). Low presence in data, considered irrelevant for overall analysis. May represent a rare anomalous situation.

6 Discussion

The results obtained from the DBSCAN clustering model provide valuable insights into the operational states of the plastic injection molding machine. A comparison with the state-of-the-art methodologies, particularly the optimization of the fixed plate in plastic injection machines proposed by Shu Sun [3], highlights both similarities and advancements in our approach. While Sun's study focused on structural optimization to reduce deformation and enhance the durability of

machine components, our research explores machine states and anomalies through advanced clustering techniques, providing a complementary perspective on machine efficiency and predictive maintenance.

One of the primary contributions of this study is the identification and differentiation of machine states beyond the predefined configuration modes. Unlike conventional analyses that rely on fixed operational modes, our approach demonstrates that the machine exhibits nuanced states that are not strictly tied to these predefined modes. This distinction is crucial in improving machine diagnostics and optimizing operational parameters. By leveraging Principal Component Analysis to reduce data dimensionality and integrating the Silhouette Score metric to assess clustering quality, our methodology ensures robust segmentation of machine states, contributing to enhanced monitoring and predictive maintenance strategies.

A key insight from our findings is the ability to detect and classify anomalous states effectively. The identification of outliers (label -1) as potential severe malfunctions aligns with industry concerns regarding common failures in plastic injection molding machines [2]. These include issues such as injection device obstruction, mold cooling inefficiencies, and variations in pressure and temperature. By clustering machine states based on real-time operational data, our method offers a data-driven approach to early fault detection, thereby complementing conventional preventive maintenance practices.

Moreover, the segmentation of machine states into seven distinct clusters provides a more granular understanding of operational behaviors. For instance, the differentiation between normal operation with a four-cavity mold (light green cluster) and operation with fewer cavities (pink and yellow clusters) enables more precise monitoring of mold performance and defect rates. Similarly, the identification of idle states (red cluster) allows for a more accurate assessment of machine utilization and efficiency.

Our study also introduces an optimized parameter selection process for DBSCAN, where the variance adjustments in PCA significantly improved clustering results. The achieved Silhouette Score of approximately 80% indicates a high-quality clustering performance, validating our model's effectiveness in distinguishing meaningful machine states. Compared to existing methods that rely on static thresholds or rule-based classifications, our data-driven approach offers greater adaptability to varying operational conditions.

In conclusion, this research advances the understanding of plastic injection molding machine behavior by introducing a clustering-based approach to state identification and anomaly detection. While structural optimizations, such as those proposed by Sun [3], enhance machine longevity, our approach contributes to real-time monitoring and predictive maintenance. Future work may involve integrating additional sensor data and refining clustering techniques to further enhance predictive capabilities and operational efficiency.

7 Conclusion

This study addressed the challenge of accurately identifying and monitoring the operational states of plastic injection molding machines. Unsupervised learning DBSCAN clustering algorithm was applied in order to determine machine states automatically.

The results demonstrated that the machine operates in more states than predefined configuration modes suggest. Identifying seven distinct clusters, including normal operation, idle periods, and potential anomalies, our approach provides a more detailed and data-driven method for assessing machine behavior. The ability to detect outliers and abnormal states further enhances machine diagnostics and predictive maintenance strategies.

This research contributes to the state of the art by moving beyond structural optimizations

and incorporating machine learning techniques to analyze machine performance dynamically.

Despite its advantages, this study has certain limitations. The clustering model's effectiveness depends on the quality and quantity of available data. Additionally, external factors such as material variations and environmental conditions may influence clustering outcomes, necessitating further refinements.

Future work should explore integrating additional sensor data, refining clustering algorithms, and incorporating real-time adaptive models to enhance predictive capabilities. Expanding this approach to other industrial machinery could further validate its applicability and effectiveness in improving manufacturing efficiency.

References

[1] Xometry. Basics of plastic injection molding, 2025. The process, the machine's components, and compatible materials. Available online at https://www.xometry.com/resources/injection-molding/basics-of-plastic-injection-molding/ Last accessed on Feb. 13, 2025.

[2] TopStar Machine. Common operating problems and solutions for several plastic injection machine, 2025. Plastic injection machine play a vital role in modern injection molding manufacturing processes. Available at https://www.topstarmachine.com/common-operating-problems-and-solutions-for-several-plastic-injection-machine/. Last accessed on: Feb. 13, 2025.

[3] Shu Huang Sun. Optimum topology design for the stationary platen of a plastic injection machine. *Computers in Industry*, 55(2):147–158, 2004.

[4] Paola Pierleoni, Lorenzo Palma, Alberto Belli, and Luisiana Sabbatini. Using plastic injection moulding machine process parameters for predictive maintenance purposes. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 115–120, 2020.

[5] Dominik Kozjek, Rok Vrabič, David Kralj, Peter Butala, and Nada Lavrač. Data mining for fault diagnostics: A case for plastic injection molding. *Procedia CIRP*, 81:809–814, 2019. 52nd CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019.

[6] Gökhan Aslantaş, Tarik Alaygut, Mustafa Rumelli, Mehmet Özsaraç, et al. Estimating types of faults on plastic injection molding machines from sensor data for predictive maintenance. *Artificial Intelligence Theory and Applications*, 3(1):1–11, 2023.

Chapter 11

Development of a Machine Learning model to classify different postures for the sitting position

Contribution #31
Cite as:
Maria Coutinho, Beatriz Costa, Fernanda Coutinho, Mateus Mendes. Development of a Machine Learning model to classify different postures for the sitting position. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-31,
  title ={Development of a Machine Learning model to
     classify different postures for the sitting
     position},
  author={Maria Coutinho and  Beatriz Costa and
     Fernanda Coutinho and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Development of a Machine Learning model to classify different postures for the sitting position

Maria Coutinho[1,*], Beatriz Costa[2], Fernanda Coutinho[1,3], Mateus Mendes[1,3,4]
[1] Polytechnic Univ. of Coimbra, Coimbra Institute of Engineering
Coimbra, Portugal
[2] Sensing Future Technologies, Coimbra, Portugal
[3] RCM^{2+} Research Centre, Coimbra, Portugal
[4] Institute of Systems and Robotics, Univ. of Coimbra, Coimbra, Portugal
[*] Corresponding author a2022142012@isec.pt
beatrizcosta@sensingfuture.pt, {fermaco, mmendes}@isec.pt

Abstract

This paper presents the development of a Machine Learning model aimed at classifying sitting postures, with the broader goal of promoting ergonomic practices and preventing musculoskeletal disorders in the workplace. The motivation behind this study lies in the growing concern over health issues related to prolonged incorrect postures during desk work. Data were collected using pressure sensors embedded in a prototype chair for testing purposes, capturing ten distinct sitting postures, each sustained for 30 seconds. The data analysis was conducted on the Google Colab platform with the PyCaret library, enabling efficient testing and comparison of five supervised classification algorithms: Random Forest (RF), eXtreme Gradient Boosting (XGB), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and k-Nearest Neighbors (KNN). The best performance was achieved with the XGB model, particularly when trained with filtered pressure images and Body Mass Index (BMI) as an additional variable, reaching an F1-Score of 0.9750. In parallel, correlation analysis was conducted between variables such as average pressure, maximum pressure, area of contact, and percentage of pressure, providing insights into variable redundancy and complementarity across postures. The findings highlight the feasibility of integrating intelligent posture-monitoring systems into everyday office environments, contributing to the advancement of workplace ergonomics and proactive health monitoring.

Keywords: Machine Learning, Pressure Maps, Ergonomics, Posture Classification, Pressure Sensors, Occupational Health

1 Introduction

Adopting inappropriate postures for long periods of sitting is a significant factor in the development of musculoskeletal disorders (MSDs), that may compromise the quality of life and performance in the workplace. These disorders represent a growing concern in various areas, especially those involving prolonged sedentary activities, such as offices and operation centers [1].

In recent years, advances in technologies based on pressure sensors, combined with machine learning algorithms, have enabled new approaches to postural monitoring [2]. These solutions offer the potential to identify patterns of postural behavior in real time, enabling preventive and corrective interventions.

In this context, this work proposes the development of a computational model for the automatic classification of different sitting postures, based on data obtained by pressure sensors. These sensors are integrated in a prototype instrumented chair, developed by an industrial partner.

The main objective is to analyze the suitability of using different machine learning algorithms to identify user posture in the prototype system. Therefore, aspects such as identification accuracy and suitability for real-time operation will be considered.

By fulfilling this objective, the prototype may evolve into a system that can contribute to the prevention of musculoskeletal injuries and diseases, optimization of ergonomic practices and, consequently, improvement of physical well-being in the workplace [3].

The methodology adopted in this study involves collecting postural data using pressure sensors installed in a chair, followed by the application of machine learning techniques to create and evaluate classification models.

This article is organized as follows: Section 2 presents a review of the state of the art on postural monitoring and machine learning applications in this context. Section 3 describes the methodology used, including data collection, model development, and evaluation procedures, providing a comprehensive overview of the experimental workflow. Section 4 presents the results obtained. Section 5 discusses these results in detail. Finally, Section 6 outlines the conclusions and suggestions for future work.

2 State of the art

The increasing prevalence of sedentary lifestyles, particularly in occupational settings, has heightened concerns regarding the health implications of prolonged sitting. Poor sitting posture is a significant contributor to musculoskeletal disorders (MSDs), which can lead to chronic pain and decreased quality of life. To address these issues, researchers have explored various technologies for monitoring and improving sitting posture.

2.1 Sensor-Based Posture Monitoring

Among the technologies investigated, pressure sensor-based systems have gained prominence due to their non-intrusive nature and ability to capture detailed pressure distribution data. These systems typically involve embedding pressure sensors within seating surfaces to monitor the user's posture in real-time. For instance, the Sitting Posture Recognition System (SPRS) utilizes a matrix of pressure sensors to classify sitting postures with an accuracy of up to 99.1%, demonstrating the potential of such systems in real-world applications [4] [5].

Additionally, smart sensing chairs equipped with Force Sensing Resistors (FSRs) have been developed to detect and classify various sitting postures. These chairs leverage machine learning algorithms, such as Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs), to analyze pressure data and provide feedback to users. However, challenges remain in achieving high classification accuracy, particularly due to limitations in training data diversity [6].

2.2 Machine Learning Approaches

Machine learning techniques have been extensively applied to enhance the accuracy and efficiency of posture classification systems. A comprehensive literature review highlights the utilization of various algorithms, including Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests, in classifying sitting postures based on pressure sensor data

[4]. These algorithms have shown promising results, with some systems achieving classification accuracies exceeding 90% [7].

Furthermore, advancements in deep learning have introduced more sophisticated models capable of handling complex and high-dimensional data. For example, the integration of Spiking Neural Networks (SNNs) with pressure sensor data has demonstrated improved performance in recognizing subtle posture variations [5]. Such approaches offer the potential for more adaptive and personalized posture monitoring solutions [6].

2.3 Challenges in Postural Identification

Despite the progress in sensor technologies and machine learning algorithms, several challenges persist in the development of effective posture monitoring systems [6]. Key issues include sensor placement optimization, data variability among users, efficient real-time operation, and the need for large, diverse datasets to train robust models. Moreover, ensuring user comfort and system unobtrusiveness remains critical for widespread adoption [8].

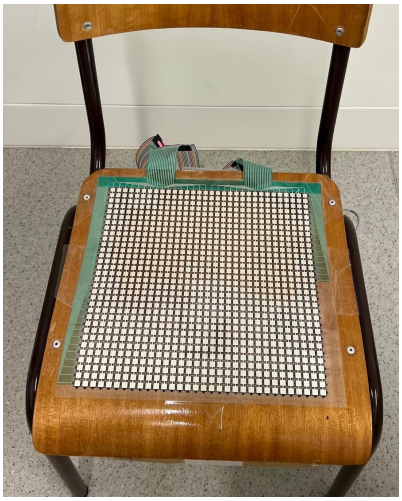
3 Methodology

This section outlines the experimental procedures, data acquisition strategy, and modeling approach used in this study. A systematic framework was adopted to ensure consistent data collection and to evaluate the performance of various machine learning algorithms in classifying seated postures based on pressure sensor input.

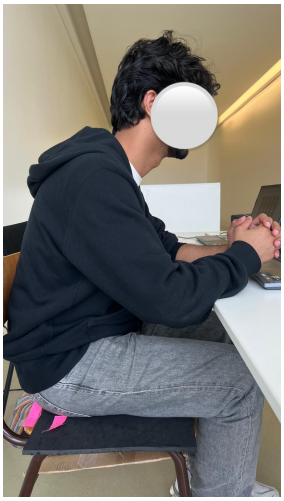
3.1 Data Collection

Following approval from the Ethics Committee of the Polytechnic Institute of Coimbra, a total of 40 volunteers participated in the study (Figure 1b), helping collect data using a prototype instrumented chair equipped with pressure sensors strategically distributed across the seating surface (Figure 1a). These sensors were connected to a pressure mapping system, integrated with an arduino and supported by a customizable software, which enabled both real-time visualization and continuous recording of the data.

Each volunteer was instructed to simulate 10 distinct sitting postures. Each posture was held for a duration of 30 seconds. The postures selected focus on pressure-based ergonomic assessment for seated configurations (Figure 2).



(a) Instrumented chair setup.



(b) Example of participant during data collection.

Figure 1: Experimental setup: (a) pressure-sensor chair; (b) data collection with participant.



Figure 2: Set of 10 simulated sitting postures adopted by participants [9].

3.2 Model Development and Evaluation

The classification pipeline was developed using the PyCaret machine learning framework within a Google Colab environment, offering a cloud-based infrastructure that ensured both computa-

tional efficiency and experimental reproducibility. This framework streamlined the processes of data preprocessing, feature engineering, model training, and performance evaluation, allowing for rapid prototyping and consistent comparisons across multiple algorithms. The raw data collected from the pressure sensors consisted of time-series matrices representing force distribution (in kg/cm^2) across a 32×32 grid, along with additional variables such as average pressure, maximum pressure, center of pressure (CoP), percentage of total pressure per quadrant, and contact area.

As part of the preprocessing stage, a filtering procedure was applied to reduce high-frequency noise caused by small involuntary body movements and sensor fluctuations. This filter consisted of a median filter applied frame by frame across the pressure maps, replacing each cell's value by the median of its 3×3 neighborhood. This operation preserved edges while smoothing out isolated spikes, resulting in more stable input patterns for the classification algorithms.

A total of five supervised classification algorithms were evaluated in this study: Random Forest (RF), eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and k-Nearest Neighbors (KNN). These algorithms were selected due to their demonstrated effectiveness in prior works involving sensor-based classification tasks.

Each model was trained under multiple configurations, using both raw and preprocessed pressure data. Additionally, the inclusion of various features tested. These are the Body Mass Index (BMI), Center of Pressure (CoP), and Percentage of Pressure distributed across the sensing matrix.

To ensure robust generalization and minimize the risk of overfitting, all models were evaluated using stratified 10-fold cross-validation. This approach allowed each posture class to be proportionally represented in both training and validation folds, enhancing the reliability of the evaluation process.

Performance was assessed using two primary metrics: accuracy and F1-score. Accuracy provides a global view of classification correctness across all posture classes, while F1-score accounts for class imbalance and offers a harmonic mean of precision and recall. This combination was chosen to ensure a balanced evaluation, particularly important in multiclass problems such as seated posture classification, where certain postures may be harder to distinguish due to subtle variations in pressure distribution.

4 Results

able 1 presents the accuracy results of the five evaluated models across eight dataset configurations. These configurations include combinations of raw sensor data, median-filtered data, and engineered features such as Body Mass Index (BMI), Center of Pressure (CoP), and Percentage of Pressure (%Press). This consolidated table enables a direct comparison of model behavior across all experimental conditions.

Among all models, ANN consistently achieved high performance, with its best result (0.9486) observed in the "No Filter + BMI" configuration. This suggests that the model is particularly effective when using raw pressure data augmented with simple physiological features. Even in the absence of any engineered variables, ANN maintained an excellent baseline accuracy of 0.9463 in the "No Filter" setting. While filtering slightly reduced its performance (e.g., 0.9379 in "Filter"), the model remained robust across all conditions, never falling below 0.9355.

Random Forest (RF) also showed strong and stable performance across all configurations. It achieved its highest accuracy of 0.9475 in the "No Filter + BMI" setup, closely followed by 0.9415 in "No Filter + %Press". These results suggest that RF benefits from engineered features and performs well both with and without filtering.

XGBoost (XGB) clearly outperformed all other models in terms of raw accuracy, reaching a peak value of 0.9761 in the "Filter + BMI" configuration. It also performed very well with "No Filter + BMI" (0.9677), demonstrating its ability to capitalize on informative engineered features. However, its performance declined slightly when using CoP or %Press, indicating that BMI is a more discriminative feature in this context.

In contrast, Support Vector Machine (SVM) showed high accuracy with raw unfiltered data (0.9390), but its performance dropped drastically when engineered features were introduced. It fell to 0.2186 with BMI and as low as 0.0454 with %Press. This suggests that SVM is highly sensitive to feature scaling and transformations, and may struggle to generalize in more complex or heterogeneous feature spaces.

K-Nearest Neighbors (KNN) consistently exhibited the lowest classification accuracy among all models. It performed best in the "No Filter + BMI" condition (0.7193), but its accuracy dropped significantly in other setups, particularly when CoP or %Press features were introduced. The poor performance of KNN indicates that it is not well suited for this high-dimensional and pressure-based dataset without substantial optimization.

In summary, ANN and RF proved to be the most balanced and robust models across configurations, while XGB showed superior accuracy when properly tuned with filtered data and BMI. Meanwhile, SVM and KNN revealed substantial limitations, especially in handling engineered features and complex data transformations.

Table 1: Models Accuracy across various dataset configurations

Configurations	**RF**	**XGB**	**SVM**	**ANN**	**KNN**
No Filter	0.9355	0.9152	0.9390	0.9463	0.5247
Filter	0.9343	0.9140	0.8866	0.9379	0.4780
No Filter + BMI	0.9475	0.9677	0.2186	0.9486	0.7193
Filter + BMI	0.9391	0.9761	0.1637	0.9427	0.6727
No Filter + COP	0.9379	0.9164	0.8006	0.9451	0.4387
Filter + COP	0.9331	0.9116	0.8889	0.9367	0.4709
No Filter + %Press	0.9415	0.9164	0.0694	0.9415	0.3358
Filter + %Press	0.9319	0.9068	0.0454	0.9355	0.3382

4.1 Model Performance Based on F1-Score

Table 2 presents the F1-Scores obtained by the evaluated models across eight different dataset configurations, which combine raw sensor data, median filtering, and engineered features including Body Mass Index (BMI), Center of Pressure (CoP), and Percentage of Pressure (%Press). This comprehensive view highlights the models' ability to handle both raw and enriched data sources in posture classification tasks.

Artificial Neural Networks (ANN) emerged as the most consistently high-performing model. Its best F1-score, 0.9449, was achieved in the "No Filter" configuration using only raw data. Even when engineered features or filtering were introduced, ANN maintained excellent performance, never falling below 0.9305. For instance, it achieved 0.9434 with "No Filter + BMI" and 0.9404 with "No Filter + CoP", confirming its robustness and adaptability across different feature representations. The stability and high recall demonstrated by ANN make it well-suited for real-world implementations in dynamic environments.

Random Forest (RF) also demonstrated strong and reliable performance. It reached a peak F1-score of 0.9430 in the "No Filter + BMI" configuration and maintained values above 0.925 in every setting, including more complex feature combinations. Notably, RF scored 0.9361 with

"No Filter + %Press" and 0.9326 with "No Filter + CoP", confirming its resilience to changes in data preprocessing and feature composition. These results reinforce RF's practicality for scenarios demanding efficient, interpretable models with stable behavior across varying input conditions.

XGBoost (XGB) delivered the highest overall F1-score, achieving 0.9750 in the "Filter + BMI" configuration. This underscores its strength in leveraging structured and stable features, especially when median filtering is applied to reduce noise. XGB also performed very well in "No Filter + BMI" (0.9651), suggesting it is particularly effective when provided with anthropometric features like BMI. Although its performance dipped slightly in %Press and CoP-based configurations (ranging from 0.9006 to 0.9096), it consistently outperformed most models in these setups as well. Thus, XGB excels in conditions where preprocessing and feature engineering are feasible.

In contrast, Support Vector Machine (SVM) showed high variance in its performance. While it attained a competitive 0.9343 with raw data, the inclusion of engineered features led to substantial degradation. F1-scores fell dramatically to 0.1580 with "No Filter + BMI", 0.1149 with "Filter + BMI", and even lower with pressure-derived features—dropping to 0.0205 with "Filter + %Press". Although one configuration with filtered CoP yielded a respectable 0.8842, the overall instability suggests SVM's sensitivity to non-linearly distributed or colinear features, limiting its reliability in heterogeneous data environments.

K-Nearest Neighbors (KNN) consistently underperformed compared to the other models. Its highest F1-score, 0.6999, was recorded with "No Filter + BMI", showing some benefit from the addition of anthropometric data. However, all other configurations resulted in scores below 0.5—falling as low as 0.3053 with "Filter + %Press". These outcomes reflect KNN's inherent limitations in managing high-dimensional, noisy data, and its poor scalability in multiclass classification with engineered features.

In summary, ANN and RF are the most robust and versatile models across all tested conditions, offering high performance without over-dependence on specific preprocessing techniques. XGBoost stands out for its top-tier accuracy under optimal configurations but requires careful feature engineering and filtering. Meanwhile, SVM and KNN showed severe limitations, reinforcing the need for tailored model selection depending on feature complexity and deployment constraints.

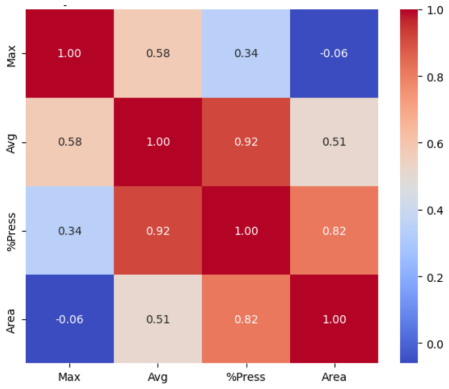
Table 2: Models F1-Score across various dataset configurations

Configurations	**RF**	**XGB**	**SVM**	**ANN**	**KNN**
No Filter	0.9305	0.9085	0.9343	0.9449	0.5003
Filter	0.9282	0.9079	0.8781	0.9328	0.4559
No Filter + BMI	0.9430	0.9651	0.1580	0.9434	0.6999
Filter + BMI	0.9332	0.9750	0.1149	0.9371	0.6551
No Filter + COP	0.9326	0.9089	0.7954	0.9404	0.4207
Filter + COP	0.9276	0.9060	0.8842	0.9314	0.4482
No Filter + %Press	0.9361	0.9088	0.0399	0.9364	0.3071
Filter + %Press	0.9250	0.9006	0.0205	0.9305	0.3053

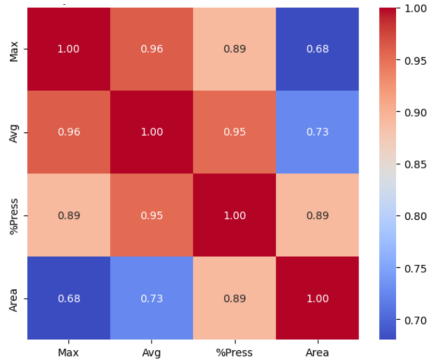
4.2 Correlation Analysis Between Pressure-Derived Features

To better understand the relationships between the pressure-derived varieds used in the classification, a correlation analysis was performed for three postures: P3 (forward-leaning), P7 (back

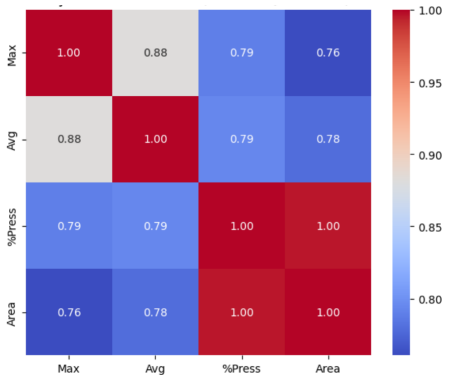
supported, right leg crossed), and P10 (sitting on the edge of the seat). The corresponding correlation matrices are presented in Figures 3a, 3b, and 3c.



(a) Posture P3 – Leaning forward



(b) Posture P7 – Back supported, right leg crossed



(c) Posture P10 – Sitting on the front edge

Figure 3: Correlation matrices between pressure-derived features for selected postures: P3, P7, and P10.

For posture P3 (Figure 3a), the correlation between maximum pressure and contact area was –0.06, which is likely not statistically significant. This suggests that there is no meaningful correlation between these two features, meaning they vary independently across participants. Such lack of association may reflect a more localized pressure distribution—typically in the front of the seat—leading to inter-subject variability that can hinder classification accuracy, especially in models that rely on global spatial patterns.

The features examined in these correlation matrices were:

- Maximum pressure: the highest value recorded across all sensors at a given time;
- Average pressure: the mean pressure value over all active sensors;
- %Pressure: the percentage of total pressure applied within a specific region (quadrant-based);

- Area: the total number of sensors registering pressure above a set threshold, representing the contact surface.

These four features were chosen as representative pressure-derived variables, commonly used in the literature for characterizing postural load distribution. They are computationally efficient, interpretable, and relevant across a wide range of ergonomic monitoring scenarios.

In contrast, posture P7 (Figure 3b) exhibited strong internal correlations among all variables. Maximum, average, and %Pressure values were highly interrelated, reflecting a uniform and stable pressure pattern. This likely results from the natural support of the chair back and leg-crossing, which reduce variability and facilitate accurate classification.

Posture P10 (Figure 3c)—where subjects sat on the front edge of the seat—showed moderate to high correlations, especially between area and %Pressure. Despite the stronger overall consistency, this posture remained sensitive to minor shifts in weight and body placement, which may explain reduced performance in models such as SVM and KNN.

In summary, these feature correlation analyses underscore how internal consistency among pressure-derived metrics can impact model performance. Postures with coherent pressure signatures (like P7) tend to produce higher classification accuracy, while postures with fragmented or unstable signatures (like P3) introduce more classification noise. This highlights the importance of context-aware feature selection in future model iterations.

5 Discussion

From an implementation standpoint, the results reinforce the value of relative pressure metrics, particularly BMI, which consistently improved classification performance while being computationally simple to compute. The inclusion of engineered features proved beneficial for certain models—especially XGBoost and Random Forest—while detrimental for others, such as SVM, which showed notable sensitivity to transformed or composite variables.

Filtering was also assessed as part of the preprocessing strategy. A median filter was applied to the raw sensor matrices to reduce noise caused by involuntary micro-movements or brief fluctuations in contact pressure. While this smoothing technique led to modest gains in classification consistency for some postures—particularly those with dynamic weight shifts like P10—it did not consistently yield statistically significant improvements in overall accuracy across models. For example, ANN showed slightly reduced performance when filtering was applied, suggesting that in some cases, important discriminative details might have been smoothed out.

Furthermore, the feature configurations tested in this study did not include all three engineered features (BMI, Center of Pressure (CoP), and Percentage of Pressure) simultaneously. This decision was made to avoid multicollinearity and retain experimental clarity. Initial correlation analyses showed that BMI and %Press, as well as %Press and Area, exhibited strong interdependence. Including all engineered variables together might have introduced redundant information, compromising both model generalization and interpretability. Additionally, testing combinations in isolation allowed for a more granular understanding of each feature's individual impact on classification performance.

The correlation analysis further revealed that postures with high internal feature coherence, such as P7, were consistently classified with high accuracy. In contrast, postures with more localized or variable pressure signatures, such as P3 and P10, exhibited lower consistency across models. This highlights the need for posture-specific strategies, potentially including tailored models, context-aware features, or adaptive thresholds for filtering and feature selection.

In practical workplace scenarios where continuous posture monitoring is needed, achieving a balance between technical accuracy, computational efficiency, and deployability is crucial. This

study demonstrates that effective seated posture classification is feasible using low-cost sensor arrays and open-source ML frameworks, but successful real-world implementation requires thoughtful system architecture and resource-aware deployment strategies.

In that context, Random Forest emerged as a leading candidate for real-time integration, given its fast inference speed, low computational overhead, and tolerance to noise. Artificial Neural Networks (ANN) also performed robustly, particularly in embedded scenarios where minimal preprocessing is preferred. These models offer promising throughput and maintain both robustness and interpretability.

On the other hand, XGBoost, while delivering the highest accuracy in filtered configurations with engineered features, demands significant computational resources and hyperparameter tuning, making it more suitable for offline batch analysis or hardware-accelerated systems. KNN, despite its conceptual simplicity, was hindered by memory demands and slow prediction times, limiting its scalability. SVM proved highly sensitive to feature composition and scale, which reduced its reliability in dynamic, heterogeneous environments.

Overall, these findings validate the feasibility of implementing posture classification systems that are not only technically robust but also operationally viable, paving the way for real-time ergonomic monitoring with concrete implications for occupational health and safety.

6 Conclusion and Future Work

This study proposed and evaluated a machine learning-based system for the classification of seated postures using pressure sensor data and derived features. A comprehensive set of experiments was conducted using five supervised learning algorithms—Random Forest, XGBoost, Support Vector Machine, Artificial Neural Networks, and k-Nearest Neighbors—applied to multiple feature configurations. The results demonstrated that models such as ANN and RF consistently achieved high classification accuracy and F1-score, particularly when using raw data or percentage-based pressure features.

The correlation analysis further revealed that postural stability and feature interdependence strongly influence classification performance. Postures with high internal feature correlation (e.g., P7) were more easily identified, while those with varied pressure concentration (e.g., P3 and P10) posed greater challenges. These insights reinforce the importance of thoughtful feature engineering and posture-specific analysis in sensor-based ergonomic systems.

From a practical standpoint, the discussion in Section 5 highlights that by selecting the appropriate machine learning algorithm, accurate posture recognition is achievable using relatively simple hardware and open-source tools. Moreover, given the lightweight nature of the trained models — particularly RF and ANN — the system shows strong potential for real-time integration. In future work, these models may be implemented in C# environments, enabling seamless integration with existing ergonomic monitoring systems. Such implementations could support live classification, while also computing performance metrics such as accuracy and F1-score in real time, allowing for adaptive feedback and system self-assessment during data acquisition sessions.

Future research should explore model generalization across diverse populations and extended sitting durations, as well as the inclusion of temporal features and sensor fusion strategies. Additionally, the deployment of this system in workplace environments will allow further validation of its usability, responsiveness, and ergonomic impact under real-world conditions.

References

[1] João Silva and Ana Pereira. Impacto da má postura na saúde ocupacional. *Revista Brasileira de Ergonomia*, 35(2):45–52, 2020.

[2] Luiza Costa and Bruno Almeida. Sensores de pressão aplicados à ergonomia: Uma revisão. *Journal of Smart Systems*, 14(1):77–89, 2021.

[3] Mariana Oliveira and Rafael Santos. Classificação de posturas utilizando aprendizado de máquina. *Computação Aplicada*, 10(4):112–121, 2022.

[4] Ming-Chih Tsai, Edward T.-H. Chu, and Chia-Rong Lee. An automated sitting posture recognition system utilizing pressure sensors. *Sensors*, 23(13):5894, June 2023.

[5] Jianquan Wang, Basim Hafidh, Haiwei Dong, and Abdulmotaleb El Saddik. Sitting posture recognition using a spiking neural network. 2022.

[6] David Faith Odesola, Janusz Kulon, Shiny Verghese, Adam Partlow, and Colin Gibson. Smart sensing chairs for sitting posture detection, classification, and monitoring: A comprehensive review. *Sensors*, 24(9):2940, May 2024.

[7] Muhammad Nadeem, Ersin Elbasi, Aymen I. Zreikat, and Mohammad Sharsheer. Sitting posture recognition systems: Comprehensive literature review and analysis. *Applied Sciences*, 14(18):8557, September 2024.

[8] Luís Fonseca, Fernando Ribeiro, and José Metrôlho. Pressure-based posture classification methods and algorithms: A systematic review. *Computers*, 12(5):104, May 2023.

[9] Ana Silva. Pressure-based ergonomic assessment for sitting postures. 2023.

Chapter 12

Ozone pollution levels in Portugal (2023): geostatistical interpolation

Contribution #32
Cite as:
Luís Margalho, Pascoal Silva. Ozone pollution levels in Portugal (2023): geostatistical interpolation. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-32,
  title ={Ozone pollution levels in Portugal (2023):
     geostatistical interpolation},
  author={Luís Margalho and  Pascoal Silva},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Ozone pollution levels in Portugal (2023): geostatistical interpolation

L.M. Margalho[1,2], P.M. da Silva[1,3,*]

[1] Instituto Politécnico de Coimbra, ISEC, DFM,
Rua Pedro Nunes, 3030-199 Coimbra, Portugal.
[2] RCM2$^+$ - Research Center in Asset Management and System Engineering
[3] CMUC-Centro de Matemática da Universidade de Coimbra.
[*] Corresponding author
{lmelo,pascals}@isec.pt

Abstract

As is well known, the ozone layer is essential to protect life on Earth. The study of the evolution of ozone in a given region can contribute to a better understanding of the evolution of health problems in humans, particularly respiratory problems. Furthermore, it can allow the development of strategies to protect the Ozone Layer. Ozone is measured by air quality monitoring stations, and when critical levels are reached, alerts are issued to the population to ensure their protection and avoid public health problems. This work aims to analyze the measurement data from monitoring stations in relation to ozone (O_3) throughout 2023 in the regions with the largest populations in Portugal, the metropolitan areas of Lisbon and Porto, respectively. Maps of predicted concentrations are created at unsampled locations for different months using geostatistical interpolation.

Keywords: Ozone, Health, Geostatistics, Environmental Monitoring, Data Analysis

1 Introduction

Ozone (O_3) is a highly reactive oxidizing gas composed of three oxygen atoms. It is well known that it plays a crucial role in protecting life on Earth by being able to filter ultraviolet solar radiation. A lot of information about ozone can be found in the literature as for example in [1] [2] [3], namely about its effects on health, biodiversity and its relationship with climate change. This concern should not only be shared by experts and researchers, but also by the community in general. In Portugal, for example, the Portuguese Environment Agency (APA) shares a lot of information on this topic (https://apambiente.pt/ar-e-ruido/ozono-o3).
Ozone is formed by chemical reactions in the atmosphere, from precursors such as hydrocarbons and nitrogen oxides. Primary pollutants and organic compounds from industrial processes and transport contribute significantly to the production of ozone, as does solar radiation. Therefore, higher concentrations of ozone are expected to be reached in areas with more traffic or industrial development. Higher levels of pollution are also expected on hotter days with strong solar radiation, caused by increased ozone concentrations.
High ozone concentrations are highly harmful to human health, so monitoring levels throughout the territory is very important. This near-real-time recording allows a given population to be alerted to the existing danger at a time when critical levels, normally above the reference value

of 120 $\mu g/m^3$, are detected. In Portugal, these levels are monitored by the QualAr information system. Measurements are made by air quality monitoring stations, and when critical levels are reached, alerts are issued to ensure the protection of public health and the environment. Critical levels and alerts are defined by the APA. The long-term target is 120 $\mu g/m^3$, relative to the maximum concentration of the eight-hour average of the day.
Regarding the negative effects of human exposure to high levels of ozone, we can highlight irritation of the eyes, nose, and throat, headaches and chest pains, and breathing difficulties. In fact, one of the main characteristics of this gas is its penetration into the respiratory tract and the negative effects it causes on lung function [4]. It is also harmful to vegetation and biodiversity [5] and causes changes in materials such as rubber and textiles.
The importance of studying the evolution of ozone concentrations over space is related to the fact that high concentrations are not reached and verified in the area of the emission source. High concentrations can easily occur at a long distance from the emission source.
Climate change influences ozone levels primarily through alterations in meteorological factors such as temperature, cloudiness, precipitation, wind speeds, and boundary layer dynamics. These changes affect key processes that determine ozone formation and distribution, including photochemical reaction rates, vertical mixing, horizontal transport, biogenic emissions, and the rates of pollutant removal by wet and dry deposition [1].
In what follows, the structure of the work is described. Section 2 presents the state-of-the-art. The important theoretical part supporting the techniques used is described in Section 3, with reference to autocorrelation, the covariance function, the Variogram, and Kriging. Section 4 presents the results, where an analysis of the data studied is performed, as well as a prediction of the evolution of ozone concentration in the region in focus. The discussion of the results is carried out in Section 5. Finally, some conclusions are presented in Section 6.

2 State of the art

As stated in the Introduction section (Section 1), tropospheric ozone O_3, a secondary air pollutant formed through photochemical reactions involving nitrogen oxides NO_x and volatile organic compounds (VOCs), is known to have harmful effects on human health. Acute and chronic exposure to elevated concentrations of ozone has been associated with respiratory distress, impaired lung function, increased asthma exacerbations, and increased cardiopulmonary morbidity and mortality [6]. Due to the spatial heterogeneity of ozone and the limited distribution of ground-based monitoring stations, there is a growing need for spatial interpolation methods to estimate population-level exposure with greater accuracy [7].
Kriging, a geostatistical technique, provides an effective framework for modeling the spatial distribution of ozone concentrations. Using the spatial autocorrelation structure of the measured data, kriging generates continuous prediction surfaces of ozone levels along with estimation variances, thus facilitating high-resolution exposure assessment [8, 9]. These surfaces are essential for environmental epidemiology, enabling the linkage of ozone concentrations to health outcomes in regions without direct measurements. Furthermore, kriging variance offers a valuable metric for assessing prediction uncertainty and identifying areas requiring additional data collection.
The application of kriging in air pollution studies supports more refined health impact assessments and contributes to evidence-based environmental health policy. This integration is critical to address urban-rural disparities in air quality monitoring and to ensure equitable protection from the health effects of ozone pollution.
In [10], the authors applied kriging to predict daytime ozone levels in a spatial domain, demonstrating its utility in estimating exposure in areas lacking direct monitoring. The study empha-

sized the effectiveness of the method in capturing spatial variability of ozone concentrations.
The authors in [11] explored the feasibility of using kriging to predict monthly mean ozone concentrations, facilitating the estimation of long-term exposure parameters essential for studies of chronic health effects. The statistical analysis indicated that maximum ozone concentrations and regional variations were more prominent in the warmer months, especially in July and August, which showed higher standard deviations, suggesting greater regional variability during these months, and that ozone concentrations tend to be lowest in winter and peak during summer, correlating with seasonal changes in solar radiation and climatic conditions that favor ozone formation.
Also in [12], the kriging methodology was found useful for predicting tropospheric O_3 pollution, with most episodes occurring in urban areas during late afternoon hours. The authors concluded that the use of kriging enhances ozone pollution mapping by providing statistically valid spatial predictions of O_3 concentrations at unmonitored locations based on observed data points, and that the kriging's ability to incorporate spatial correlations and provide uncertainty estimates makes it a powerful tool for accurate, reliable ozone pollution mapping, which is vital for environmental monitoring and public health decision-making.

3 Theoretical framework

Geostatistics is a branch of statistics that deals with spatial or spatiotemporal data, providing tools and techniques for analyzing and interpreting data that are correlated based on the locations where they are collected. Geostatistics is widely used in fields such as mining, agriculture, epidemiology or environmental science, for tasks such as resource estimation, disease mapping or environmental monitoring [13, 14].

3.1 Spatial autocorrelation

Spatial autocorrelation measures the degree to which a spatial phenomenon is correlated to itself in space [15]. It is a measure of how much the values of a variable at different locations are correlated with each other, considering their spatial relationships.
It is often assessed using a metric called the variogram, which quantifies how the similarity between sampled points decreases as the distance between them increases.

3.2 Variogram and Covariance Function

The *variogram* is a key tool in geostatistics. It is a function that describes the degree of spatial variability of a dataset and is defined as:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(s_i) - Z(s_i + h)]^2 \tag{1}$$

where $N(h)$ is the number of data pairs separated by distance h, and $Z(s_i)$ is the variable of interest at location s_i [16].
The *covariance function*, together with the variogram, is a fundamental tool to model dependent data observed over space [17]. It describes how two observations at different locations are related, which is essential for predicting at nonobserved locations using an interpolation method called *kriging*.

3.3 Kriging

Kriging is a geostatistical interpolation method that allows the estimation of unknown values at specific locations based on known values in a spatial domain, taking into account the spatial correlation between those values. It is an optimal estimator that minimizes the prediction error by using weighted averages of the surrounding data points, where the weights are derived from the spatial correlation structure of the data [18].
There are different types of kriging [19], such as ordinary kriging (assumes a constant mean and known variance), simple kriging (assumes a constant unknown mean but known variance), and universal kriging (which allows for a trend in the data together with a stochastic component), each with its own assumptions and suitability for different data types and spatial patterns.

3.3.1 Prediction at unsampled locations

The simple kriging estimate at a location s_0 is expressed as :

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{2}$$

where $\hat{Z}(s_0)$ is the estimated value at location s_0, λ_i are the weights assigned to each known value $Z(s_i)$ at an observed location s_i [20].

3.3.2 Error of predicted values

In kriging, interpolation error refers to the uncertainty in the predicted value at unsampled locations, quantified as a variance (called the kriging variance or estimation variance). After determining the kriging weights λ_i, the kriging variance σ_k^2 at unsampled location s_0 is computed as

$$\sigma_k^2(s_0) = \gamma(0) - \sum_{i=1}^{n} \lambda_i \gamma(s_i, s_0) \tag{3}$$

where $\gamma(0)$ is the variogram at zero distance and $\gamma(s_i, s_0)$ is the semivariance between sample point s_i and the prediction point s_0 [20].

4 Results

4.1 Description of the Study Area

The application of the concepts introduced in Section 3 will now be made using a data set of O_3 concentrations collected during the year of 2023. Figure 1 shows the sampling locations in mainland Portugal. It can be noticed that the distribution of these sampling locations is not uniform over the entire mainland, which, in terms of kriging predictions, would lead to an increase in interpolation errors due to a reduced number of nearby locations. Because of this, the prediction at unsampled locations will be obtained only in regions with higher sampling intensity, represented by the boxes in Figure 1, which are amplified in Figure 2. The regions shown in Figure 2 refer to the North region (left) and the Lisbon region (right) in Portugal. These regions have a higher density of monitoring location because they are more densely populated and have a greater level of industrial development.

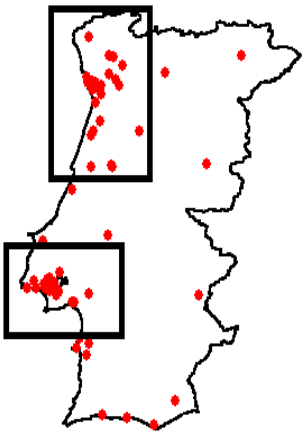


Figure 1: Location of monitoring stations.

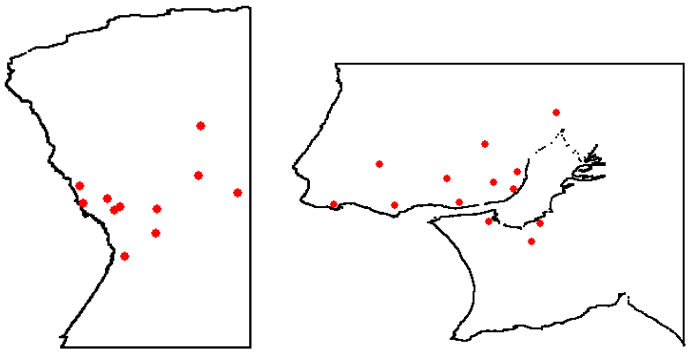


Figure 2: Location of selected monitoring stations in the northern (left) and Lisbon (right) region.

4.2 Data base

Due to the existence of monitoring stations with no data recorded, the data sets to be used to make predictions over each of the regions showed in Figure 2 contains measurements of the monthly mean O_3 concentration over the eight-hour period, from 10h00 to 17h00, relating to 11 locations in the northern region and 13 locations in the Lisbon region.
All data used are available online at https://qualar.apambiente.pt/ and are of the responsibility of APA.

4.3 Exploratory data analysis

Table 1 presents a descriptive summary of O_3 concentrations in the northern region, throughout each month of 2023. Regarding the maximum and minimum concentration values of O_3, it can be observed that the lowest value is measured in January, while the highest values are reached in May, September and October. Although these values are greater than 100, no alerts were issued because the critical value was not reached.

The months of April and May present the highest median values, and the highest average is reached in the month of May, although it may be interesting to note that the month of April has higher median and mean values than what was observed in the months of September and October, when the highest maximum values were recorded.

Table 1: Descriptive measures of O_3 concentrations in the northern region from January to December.

	Min	1st Qu	Median	Mean	3rd Qu	Max
Jan	14.01	54.76	57.55	54.02	59.65	64.93
Feb	29.06	67.25	69.69	65.77	70.19	73.98
Mar	22.99	73.35	75.45	71.72	78.93	84.23
Apr	24.71	73.78	81.84	75.78	83.36	91.03
May	32.96	83.59	88.34	83.69	90.83	100.18
Jun	27.89	65.15	74.81	70.46	78.29	88.53
Jul	28.89	48.84	59.82	56.58	62.78	79.30
Aug	25.41	54.22	66.19	63.06	74.26	86.28
Sept	23.57	61.11	69.02	67.41	74.79	109.62
Oct	27.52	61.57	70.77	73.82	88.50	107.94
Nov	27.52	46.15	56.33	52.38	60.49	68.29
Dec	30.70	42.97	50.02	48.44	54.04	60.49

Table 2: Descriptive measure of O_3 concentrations in the Lisbon region from January to December.

Month	Min	1st Qu	Median	Mean	3rd Qu	Max
Jan	46.26	53.33	56.96	57.62	60.80	70.73
Feb	56.87	64.24	65.59	66.43	67.93	84.02
Mar	64.84	75.10	78.62	77.91	81.11	85.64
Apr	73.79	84.64	87.04	85.97	89.46	92.52
May	85.22	89.84	94.83	94.69	97.98	104.46
Jun	78.70	88.16	90.91	90.19	94.34	97.64
Jul	61.56	68.78	71.28	72.16	76.48	80.39
Aug	72.10	77.28	79.84	80.06	83.98	86.91
Sept	63.92	72.61	78.75	77.77	82.08	87.77
Oct	56.50	71.41	76.50	76.82	84.94	89.23
Nov	42.35	50.52	54.71	56.21	60.27	75.12
Dec	34.30	39.34	41.31	44.11	47.50	58.65

A descriptive summary of O_3 concentrations in the Lisbon region, throughout each month of 2023, is available in Table 2. Regarding the maximum and minimum concentration values of O_3, it can be observed that the lowest value is measured in January, while the highest values are reached in May, September and October. It can be seen that in the Lisbon region no alert was issued because the concentrations reached are below the critical value.
The months of April, May, and June present the highest median values, and the highest average is reached in the month of May, just as happened in the northern region.
From the global analysis of data from the two regions, we can easily see that the highest O_3 values are observed in the hottest months, between March and October.

4.4 Prediction analysis and discussion

Based on the values measured at each of the monitoring stations, summarized in the previous Tables 1 and 2, maps of the estimated concentration values were constructed over the regions shown in Figure 2. For that, at each point of a grid of 100×200 points for the northern region, and of 150×150 for the Lisbon region, equation (2) was used to find the estimated concentration, while equation (3) allowed to determine the estimation error. Tables 3 and 4 and Figures 3 and 4 represent the results obtained for the northern region while for the Lisbon region the results are present in Tables 5 and 6 and Figures 5 and 6, for the predicted concentration and for the estimation error, respectively.

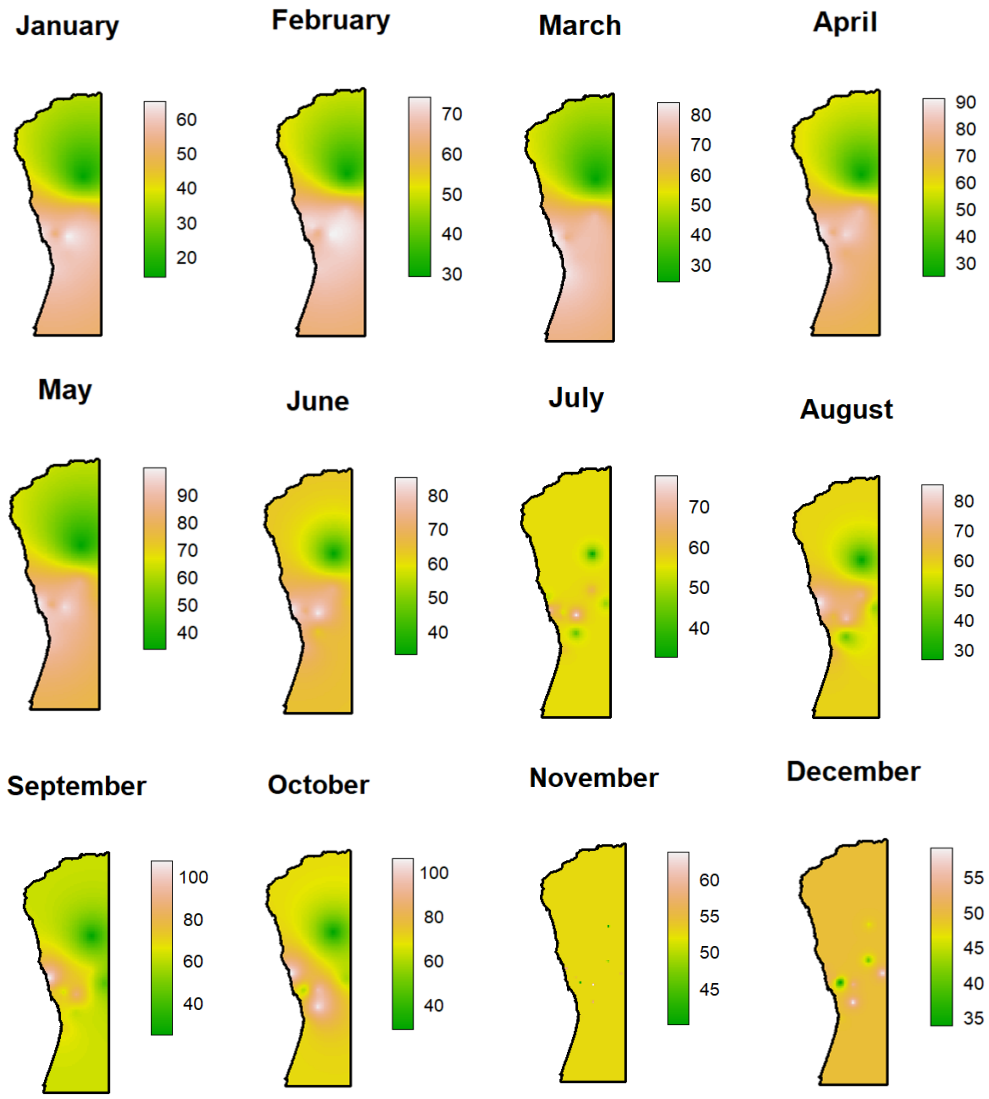


Figure 3: Maps of predicted O_3 concentrations in the northern region from January to December.

Table 3: Summary of O_3 predicted concentration in the northern region

Month	Min	1st Qu	Median	Mean	3rd Qu	Max
Jan	40.12	52.37	52.37	52.37	52.37	63.63
Feb	29.38	50.53	61.17	55.71	64.40	73.82
Mar	24.42	50.84	66.03	61.49	71.81	83.64
Apr	25.32	57.56	66.93	64.62	71.83	90.70
May	33.84	63.04	74.13	71.62	79.84	99.55
Jun	33.40	62.75	64.00	63.92	65.23	84.86
Jul	32.78	55.86	55.86	55.86	55.86	77.39
Aug	26.84	57.54	58.35	58.40	58.77	85.00
Sept	25.03	60.67	61.86	61.98	62.81	106.95
Oct	29.03	68.05	69.41	69.36	70.34	105.87
Nov	40.12	52.37	52.37	52.37	52.37	63.63
Dec	33.97	49.27	49.27	49.27	49.27	59.00

Table 4: Summary of O_3 concentration interpolation error in the northern region

Month	Min	1st Qu	Median	Mean	3rd Qu	Max
Jan	1.85	13.53	16.96	15.94	19.27	21.72
Feb	1.70	12.30	15.31	14.35	17.26	19.25
Mar	4.27	16.58	20.66	19.48	23.43	26.40
Apr	2.79	18.63	22.40	20.92	24.55	26.37
May	3.58	18.30	22.34	20.93	24.82	27.14
Jun	8.50	17.96	19.21	18.31	19.61	19.77
Jul	5.39	14.50	14.51	14.42	14.51	14.51
Aug	3.92	19.97	20.05	19.19	20.22	20.25
Sept	4.60	23.96	25.29	24.04	25.62	25.72
Oct	4.73	25.13	26.67	25.30	27.08	27.21
Nov	9.46	14.19	14.19	14.19	14.19	14.19
Dec	3.69	8.98	8.98	8.94	8.98	8.98

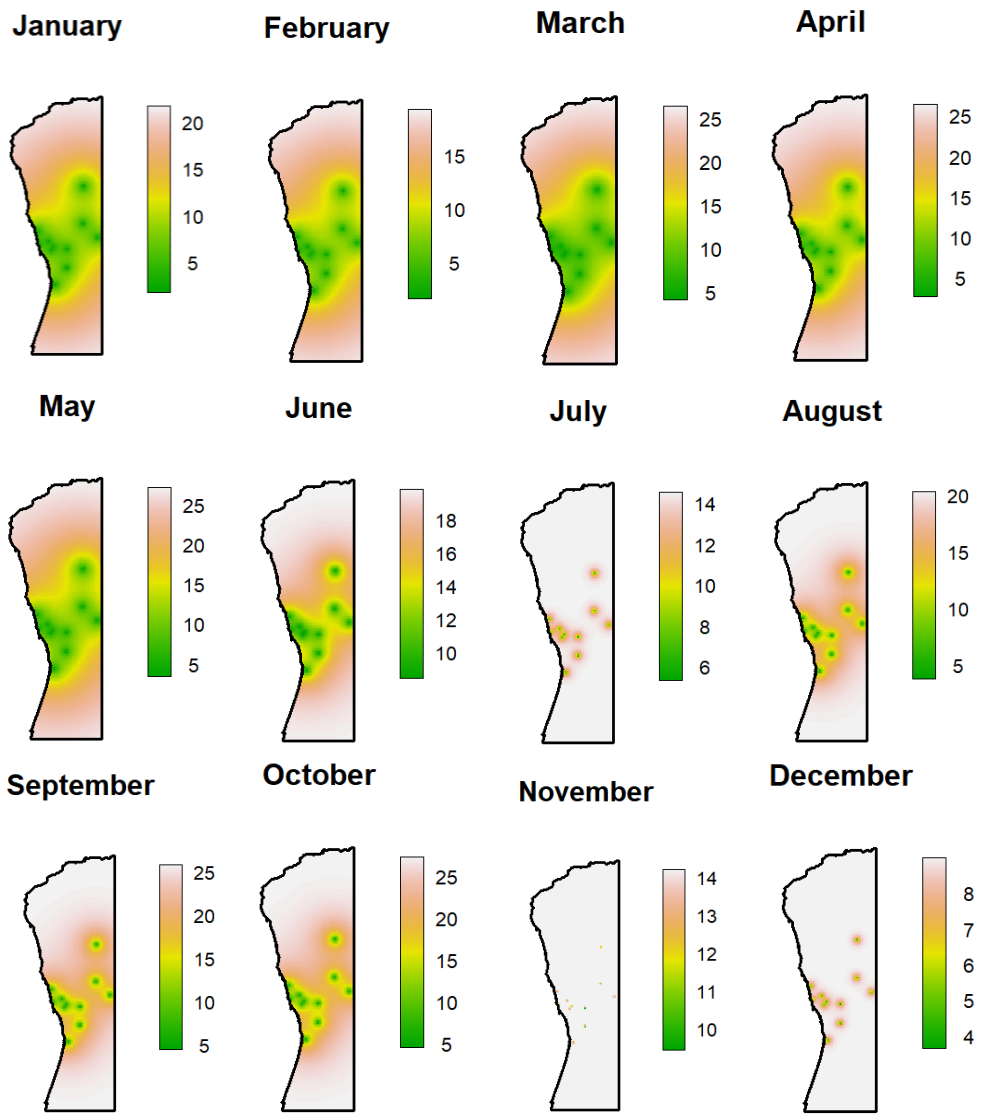


Figure 4: Interpolation error maps of O_3 concentrations in the northern region from January to December.

After analyzing the results presented in Figure 3, it can be seen that the values are now present throughout the entire surface of the considered region and not only in the measurement locations. The differences between winter and summer months are evident, with the maximum values represented in white and the minimum values in green. The same can be observed by analyzing the data in Table 3.
It can be observed that there are areas with almost constant values and, for example, there does not appear to be a variation in the concentration of O_3 in the month of November. It can also be observed that in the northern and interior parts of the region lower values of O_3 are reached, which can be justified by the fact that these areas have more green spaces and less industry.
From the analysis of Figure 4 and Table 4, it can be observed that the variation of the estimation error is smaller in July, November and December. In the other months, the values obtained are close (around 20). As expected, in both cases, higher values of the estimation errors are found in the areas farthest from the measurement stations.
Figure 5 represents the predicted concentrations in the Lisbon region. As expected and in a similar way to the north region, the higher values occur from April to October although in the months of May, August, September and October the concentration values are almost constant throughout the entire region. It is also observed that the highest values are expected to be registered in the area of Lisbon city. A similar comment can be made from the summary of O_3 predicted concentration in the Lisbon region exhibited in Table 5 Regarding the error estimates

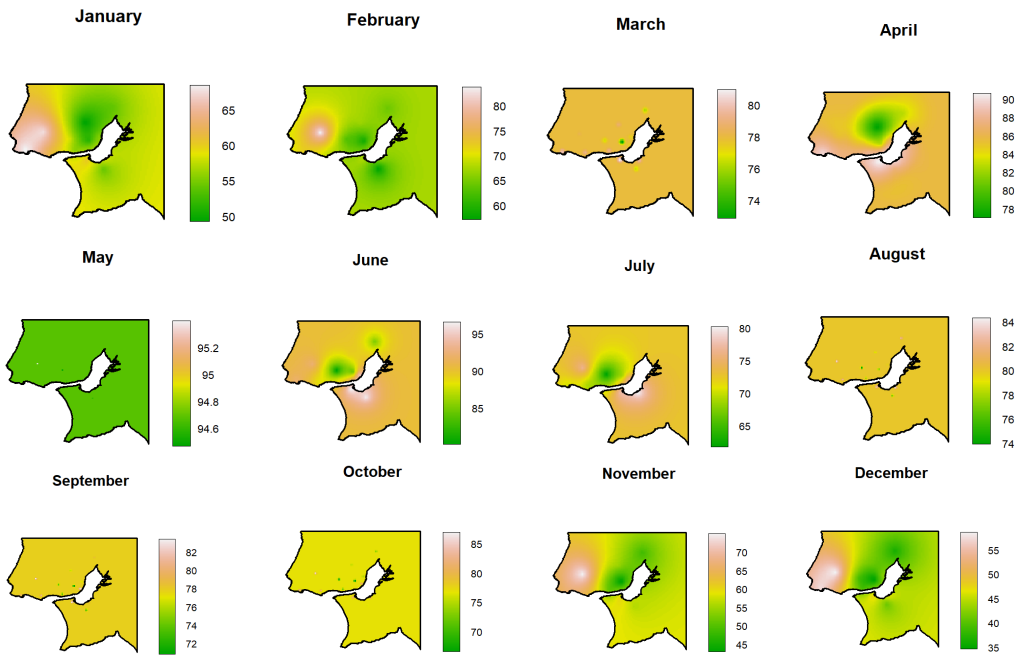


Figure 5: Maps of predicted O_3 concentrations in the Lisbon region from January to December.

for the Lisbon region, given the results of Table 6 and Figure 6, the highest values are observed in the months with the coldest climate. In fact, in these months the error can reach values between 7 and 9, while in the warmer months it is around the value 5.

Table 5: Summary of O_3 predicted concentration in the Lisbon region

Month	Min	1st Qu	Median	Mean	3rd Qu	Max.
Jan	49.33	58.03	58.85	58.98	59.65	68.41
Feb	57.24	67.36	67.65	67.67	67.76	83.70
Mar	72.91	77.91	77.91	77.91	77.91	80.97
Apr	77.10	85.67	85.67	85.67	85.67	90.61
May	94.47	94.69	94.69	94.69	94.69	95.40
Jun	80.16	90.23	90.25	90.25	90.32	96.54
Jul	61.82	72.47	72.53	72.53	72.64	80.16
Aug	73.99	80.06	80.06	80.06	80.06	84.30
Sept	70.88	77.77	77.77	77.77	77.77	83.38
Oct	66.70	76.83	76.82	76.82	76.82	86.88
Nov	43.16	57.61	58.63	58.73	59.72	74.95
Dec	34.79	44.73	45.89	46.19	47.36	58.54

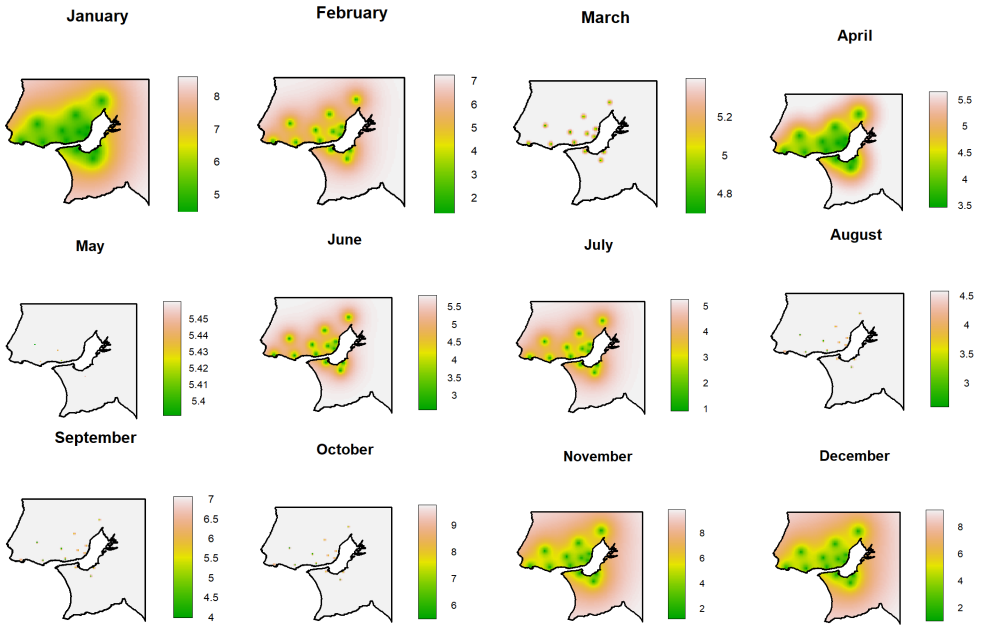


Figure 6: Interpolation error maps of O_3 concentrations in the Lisbon region from January to December.

Table 6: Summary of O_3 concentration interpolation error in the Lisbon region

Month	Min	1st Qu	Median	Mean	3rd Qu	Max.
Jan	4.45	7.37	8.09	7.71	8.38	8.56
Feb	1.29	6.86	7.14	6.80	7.19	7.20
Mar	4.69	5.40	5.40	5.39	5.40	5.30
Apr	3.47	5.49	5.63	5.41	5.63	5.63
May	5.39	5.46	5.46	5.46	5.46	5.46
Jun	2.60	5.69	5.78	5.63	5.79	5.79
Jul	0.91	4.94	5.17	4.92	5.22	5.23
Aug	2.59	4.57	4.57	4.57	4.57	4.57
Sept	3.99	7.04	7.04	7.04	7.04	7.04
Oct	5.50	9.71	9.71	9.71	9.71	9.71
Nov	1.21	8.07	9.15	8.50	9.55	9.79
Dec	1.03	7.16	8.33	7.72	8.82	9.17

5 Conclusion

An analysis of data on O_3 measurements in mainland Portugal throughout 2023 was carried out for the northern region and the Lisbon region. It can be observed that the months with the highest values are the months of May, September and October in the case of the northern region and the months of May, June and October in the case of the Lisbon region. The forecast of the evolution of O_3 throughout each month for each region was obtained using techniques associated with variogram and Kriging. It was observed that for most months the distribution of O_3 is obtained for each region, although almost constant distribution maps are found for some situations, namely in the month of November in the northern region and in the months of March, August, September and October in the Lisbon region. This fact is due to the little variation in the data measured at the stations in each of the regions under study. It is worth noting that, based on the information available and the data used here, we conclude that a critical level alert has never been issued to the population.

While there is a positive aspect in Portugal that there is effective monitoring of pollution levels and a level of safety in terms of air quality, on the other hand, there is a scarce network of monitoring stations distributed throughout the country.

Although this study only involves data from 2023, it would be interesting to analyze the evolution of O_3 concentrations in other years, to study the trend of growth or decrease in ozone levels.

References

[1] Neal Fann, Christopher G. Nolte, Patrick Dolwick, Tanya L. Spero, Amanda Curry Brown, Sharon Phillips, and Susan Anenberg. The geographic distribution and economic value of climate change-related ozone health impacts in the united states in 2030. *Journal of the Air & Waste Management Association*, 65(5):570–580, 2015. PMID: 25947315.

[2] G. Tamás, C. J. Weschler, J. Toftum, and P. O. Fanger. Influence of ozone-limonene reactions on perceived air quality. *Indoor air*, 16(3):168–178, 2006.

[3] Kleopatra Ntourou, Kyriaki-Maria Fameli, Konstantinos Moustris, Antonios Augoustinos, and Christos Tsitsis. The influence of ozone concentrations on public health over the greater athens area, greece. *Environmental Sciences Proceedings*, 26(1), 2023.

[4] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367, 2008. Proceedings of the 4th International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements).

[5] Sumitra Pal Karmakar, Anushree Baul Das, Cyaria Gurung, and Chandra Ghosh. Effects of ozone on plant health and environment: a mini review. *Res. Jr. Agril. Sci*, 13:612–619, 2022.

[6] Gabriele Donzelli and Maria Morales Suarez-Varela. Tropospheric ozone: A critical review of the literature on emissions, exposure, and health effects. *Atmosphere*, 15(7), 2024.

[7] Kari Northeim, Chetan Tiwari, and Joseph Oppong. Surface ozone monitoring and policy: A geospatial decision support tool for suitable location of monitoring stations in urban areas. *Environmental Science & Policy*, 126:48–59, 2021.

[8] Alessandra De Marco, Hector Garcia-Gomez, Alessio Collalti, Yusef Omidi Khaniabadi, Zhaozhong Feng, Chiara Proietti, Pierre Sicard, Marcello Vitale, Alessandro Anav, and Elena Paoletti. Ozone modelling and mapping for risk assessment: an overview of different approaches for human and ecosystems health. *Environmental Research*, 211:113048, 2022.

[9] Suvarna Tikle, Vrinda Anand, and Sandipan Das. Geospatial practices for airpollution and meteorological monitoring, prediction, and forecasting. In *Geospatial Practices in Natural Resources Management*, pages 549–566. Springer, 2024.

[10] L.-J.Sally Liu and A.J. Rossini. Use of kriging models to predict 12-hour mean ozone concentrations in metropolitan toronto—a pilot study. *Environment International*, 22(6):677–692, 1996.

[11] Amit Gorai, Kumar Jain, Neha Shaw, Francis Tuluri, and Paul Tchounwou. Kriging analysis for spatio-temporal variations of ground level ozone concentration. *Asian Journal of Atmospheric Environment*, 9:247–258, 12 2015.

[12] Swatantra R. Kethireddy, Paul B. Tchounwou, Hafiz A. Ahmad, Anjaneyulu Yerramilli, and John H. Young. Geospatial interpolation and mapping of tropospheric ozone pollution using geostatistics. *International Journal of Environmental Research and Public Health*, 11(1):983–1000, 2014.

[13] N. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[14] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

[15] Daniel A Griffith. Spatial autocorrelation. *International encyclopedia of human geography*, 2009:308–316, 2009.

[16] Margaret Armstrong. *Basic linear geostatistics*. Springer Science & Business Media, 1998.

[17] Chunsheng Ma. Spatio-temporal variograms and covariance models. *Advances in Applied Probability*, 37(3):706–725, 2005.

[18] E. Isaacs and R. Srivastava. *Applied Geostatistics*. Oxford University Press, New York, 1989.

[19] Pınar Bostan. Basic kriging methods in geostatistics. *Yuzuncu Yıl University Journal of Agricultural Sciences*, 27(1):10–20, 2017.

[20] Noel Cressie. Spatial prediction and ordinary kriging. *Mathematical Geology*, 20(4):405–421, May 1988.

Chapter 13

A comparative analysis of AR equipment for physical asset management applications

Contribution #33
Cite as:
Francisco Malva, Jose Marinho, Nuno Cid Martins, Ana Malta, Mateus Mendes, Nuno Ferreira. A comparative analysis of AR equipment for physical asset management applications. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-33,
  title ={A comparative analysis of AR equipment for
     physical asset management applications},
  author={Francisco Malva and  José Marinho and  Nuno
     Cid Martins and  Ana Malta and  Mateus Mendes and
     Nuno Ferreira},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

A comparative analysis of AR/MR equipment for physical asset management applications

Francisco Malva[1], José Marinho[1], Nuno Cid Martins[1],
Ana Malta[1], Mateus Mendes[1], Nuno Ferreira[1]

[1] Polytechnic University of Coimbra, 3045-093 Coimbra, Portugal
a2022149969@isec.pt

Abstract

The global market for Augmented Reality (AR) shows that its relevance is growing in a wide range of areas. In this type of reality, people can access and interact with both digital and real-world content, augmenting the information their senses perceive in a normal situation. The integration of AR/MR into Physical Asset Management (PAM) is changing the way operators interact with complex systems, providing real-time access to digital information overlaid on physical equipment. AR/MR equipment offers promising opportunities for maintenance, inspection, training, and remote support. This study presents a comparative analysis of AR/MR equipment currently available on the market and assesses their suitability for PAM applications. The evaluation considers several characteristics including display quality, field of view, ergonomics, connectivity, environmental resistance, and software compatibility. The findings provide practical insights for organizations looking to deploy AR/MR technology in asset-based environments.

Keywords: Augumented and Mixed Reality (AR/MR); Physical Asset Management (PAM); AR/MR Glasses and MR Headsets;

1 Introduction

Augmented and Mixed Reality (AR/MR) technologies superimpose digital content (such as images, audio, or data) onto the real world. Unlike Virtual Reality (VR), which immerses users in a fully simulated digital environment, AR/MR enhances the physical environment by overlaying virtual elements using devices like smartphones, tablets, AR glasses, or headsets.

Because AR/MR enables the real-time delivery of contextual information directly within the user's field of view, it significantly improves the accessibility and visualization of spatial and location-sensitive data. As a result, the adoption of AR/MR devices has grown rapidly across multiple domains, particularly in the field of Physical Asset Management (PAM). For example, as shown by Ahmed [1] and Ojha [2], AR/MR was successfully applied to construction management and asset inspection/training respectively, showcasing their potential for real-world industrial use cases within PAM.

Given this expansion, it is critical that organizations have access to comprehensive and up-to-date resources to support informed decision-making regarding the selection and implementation of AR/MR equipment. Although several detailed papers and resources have explored the usage of AR/MR in select industries, there seems to be no currently existing study comparing AR/MR equipment to gauge its suitability for PAM tasks.

To address this gap, the present paper aims to compile and compare the specifications of various AR-capable headsets and AR glasses currently available. This comparison is intended to support organizations in selecting the most suitable AR/MR solutions tailored to their operational requirements.

2 Methodology

For the purpose of this analysis, AR-capable devices were classified into two primary categories:

- **AR Glasses**: Lightweight, wearable devices designed to overlay digital information onto the user's field of view without significantly obstructing vision. These are best suited for field operations, inspections, remote assistance, and tasks requiring mobility and hands-free access to information;
- **MR Headsets**: Heavier, more immersive devices that primarily serve VR applications but include AR/MR functionality via passthrough cameras. These devices provide enhanced spatial awareness, interaction capabilities, and processing power, making them more appropriate for stationary or semi-mobile use cases such as training simulations, complex maintenance procedures, or asset visualization in controlled environments.

This categorization is necessary due to the inherent differences in ergonomics, application focus, and user interaction. AR glasses are generally intended for real-time support and lightweight visualization, while headsets are used where richer interaction and environmental mapping are essential.

The devices were evaluated based on the following parameters:

- **CPU and RAM**: PAM applications frequently involve 3D visualization, object recognition, and spatial mapping, which are computation-heavy. Evaluating CPU and memory capacity ensures that devices can support these intensive tasks without lag.
- **Internal Storage**: Many PAM environments have limited or no internet access. Local storage of 3D models, technical documentation, and offline databases is crucial for autonomous operation.
- **Operating System**: The OS determines application compatibility and integration with enterprise asset management software. A well-supported OS also affects security and developer accessibility.
- **Ergonomics**: Many PAM tasks require prolonged usage of AR/MR devices, often in physically demanding conditions. Comfort is therefore critical for operator endurance and device usability.
- **Display Resolution and Type**: High resolution is essential for interpreting technical diagrams and overlaying precise information. Display opacity also influences situational awareness and is a decisive factor in choosing between transparent AR and passthrough-based MR.
- **Audio and Voice Interaction**: Industrial environments are often noisy and require hands-free operation. Devices must have robust audio systems and voice recognition to allow efficient interaction without the need for physical input.

- **Camera Capabilities and Spatial Mapping**: Essential for inspection and maintenance, where the ability to document assets and overlay spatially aligned information is crucial. SLAM, anchoring, and depth sensing are critical to this.
- **Connectivity**: Real-time access to cloud-based systems and remote support functionalities depend on robust and versatile connectivity options, especially in distributed asset environments.
- **Artificial Intelligence Support**: Increasingly, AR/MR applications in PAM incorporate Artificial Intelligence (AI) for predictive maintenance, object recognition, and anomaly detection. The ability of a device to support AI-driven applications is therefore an emerging factor of interest.
- **Interaction Methods**: Some PAM scenarios benefit from gesture or controller-based interaction. Evaluating a range of input modalities ensures adaptability to different contexts.
- **Response Latency**: Low latency is vital for immersive AR/MR applications, especially in MR headsets using passthrough video. High latency can result in a disorienting experience and reduced task efficiency.
- **Security Certifications**: In PAM, some devices may be used in harsh or regulated environments. Certifications like IP67 or MIL-STD-810 ensure the hardware's durability and compliance with safety standards.
- **SDK and Engine Support**: For scalable deployment and custom development, support for engines like Unity and Unreal is critical. SDK availability directly affects integration ease and development speed.

In addition to the parameters above, there are several other relevant factors that are important in PAM applications but for which standardized or publicly available data is often lacking. These include:

- **Battery Life**: Many PAM tasks are carried out in the field over extended periods. Evaluating battery life ensures the device can operate through a full work shift without interruption. While this information is not always available across all devices, it remains a critical parameter;
- **Multi-User Collaboration Features**: Important for remote assistance and synchronized work in distributed teams;
- **Battery Charging Options**: Including swappable batteries or rapid charging capabilities;
- **Total Cost of Ownership**: Encompassing the purchase price, cost of required accessories, software licenses (SDKs), and support/maintenance fees;
- **User Authentication and Access Control**: Especially relevant in secure environments or multi-user scenarios;
- **User Feedback and Field Testing Results**: Provides insight into real-world performance, durability, and usability.

Although these parameters are not included in the comparative tables due to data limitations, their importance is recognized, and they should be considered in practical evaluations or pilot studies.

The following AR glasses and MR headsets were selected for analysis:

- Vuzix M400;
- Vuzix M4000;
- RealWear Navigator 500;
- ThirdEye X2;
- MagicLeap 2;
- Meta Quest 3;
- HTC Vive XR Elite;
- Varjo XR-4;
- Pico 4 Pro;
- Apple Vision Pro;
- Microsoft HoloLens 2.

Although the MagicLeap 2 is specifically designed for AR/MR applications, it was categorized with MR headsets due to its specifications and capabilities aligning more closely with that group.

It is important to note that the selection of devices does not reflect a ranking or endorsement. The list is based on prior research, updated to reflect current market offerings, with older or discontinued models replaced by actively supported alternatives.

To gather device data, among other external sources, the following were consulted: the Vuzix M400 product sheet [3], Vuzix M4000 product sheet [4], RealWear Navigator 500 data sheet [5], a presentation on the ThirdEye X2 [6], MagicLeap2 Product Specification Version 6.2 [7], Meta Quest 3 Product Page [8], HTC VIVE XR Elite Specification Page [9], Varjo XR-4 Product Page [10], Pico 4 Product Page [11], Apple Vision Pro Product Page [12], and Microsoft HoloLens 2 Documentation [13].

In cases where additional context was required or specifications were incomplete, further academic and industry sources were consulted.

When certain specification values could not be determined reliably, their corresponding table entries were marked with an em-dash (—).

3 Results and Discussion

This section presents the comparative findings from the evaluation of selected AR/MR and MR devices, attempting to provide a qualitative comparison on key technical specifications relevant to their use in PAM.

3.1 CPU and RAM

The computational capabilities of AR/MR and MR devices are primarily determined by their central processing units (CPUs) and memory (RAM), which directly influence performance, responsiveness, and the ability to run complex applications. This section presents a comparative overview of the CPU and RAM specifications of selected AR glasses and MR headsets, highlighting the variations in processing power across devices designed for different use cases. Table 1 summarizes the key specifications of these systems.

Table 1: CPU and RAM specifications of selected AR glasses.

Device	CPU	RAM
Vuzix M400	Qualcomm XR1	6GB
Vuzix M4000	Qualcomm XR1	6GB
RealWear Navigator 500	Snapdragon 662	4GB
ThirdEye X2	Qualcomm XR1	4GB
MagicLeap 2	AMD Zen 2	16GB
Meta Quest 3	Snapdragon XR2	8GB
HTC Vive VR Elite	Snapdragon XR2	8GB
Varjo XR-4	Host Dependant	Host Dependant
Pico 4 Pro	Snapdragon XR2	8GB
Apple Vision Pro	Apple M2	16GB
Microsoft HoloLens 2	Snapdragon 850 Compute Platform	4GB

The results show a clear division in processing capability between AR glasses and MR headsets. While AR glasses offer modest configurations, MR headsets far surpass them in computing power. As such, whenever the task demands it, MR headsets are generally more suitable for heavier processing workloads.

3.2 Internal Storage

Internal storage capacity and type are critical factors influencing the data handling capabilities and application reliabilty of AR/MR and MR devices. This section presents an overview of the storage specifications of various AR glasses and headsets, highlighting differences in storage size, expandability, and underlying technology.

These specifications can impact the device's suitability for data-intensive applications, such as real-time asset visualization, offline content access, and extended field use without reliance on cloud connectivity (especially in PAM, where some environments may not have a reliable connection to the Internet).

The data gathered on this subject is presented in table 2.

Storage capacities reinforce a very clear performance split. While the discussed AR glasses uniformly feature 64GB flash storage, MR headsets offer significantly higher capacities.

3.3 Ergonomics

The successful deployment of AR/MR devices in PAM hinges on ergonomic design, which directly influences user adoption, operational efficiency, and long-term usability.

Field operators often perform physically demanding tasks over extended an extended period, making comfort, ease of use, and physical adaptability critical in device selection.

Table 2: Storage specifications of the selected AR/MR equipment.

Device	Size	Type
Vuzix M400	64GB	Flash
Vuzix M4000	64GB	Flash
RealWear Navigator 500	64GB	Flash + MicroSD
ThirdEye X2	64GB	Flash
MagicLeap 2	256GB	SSD
Meta Quest 3	512GB	Flash
HTC Vive XR Elite	128GB	Flash
Varjo XR-4	—	Host PC Dependent
Pico 4 Pro	128GB / 256GB	Flash
Apple Vision Pro	256GB / 512GB / 1TB	Flash
Microsoft HoloLens 2	64GB	Flash

This subsection evaluates the weight and Field of View (FoV) of selected AR/MR devices, two key factors which influence ergonomic suitability. The result of this research can be viewed in tables 3 and 4, respectively.

Table 3: Weight specifications of the selected AR/MR equipment.

Device	Weight (grams)
Vuzix M400	190
Vuzix M4000	222
RealWear Navigator 500	274
ThirdEye X2	170
MagicLeap 2	260
Meta Quest 3	515
HTC Vive XR Elite	625
Varjo XR-4	665
Pico 4 Pro	586
Apple Vision Pro	—
Microsoft HoloLens 2	566

The weight analysis highlights the ergonomic advantage of AR/MR AR glasses. In contrast, MR headsets are significantly heavier, often exceeding 500g. Such weight can lead to user fatigue and limit the practical use of headsets for long-duration, mobile tasks.

Complementing this, the FoV comparison highlights a trade-off between form factor and visual immersion: AR glasses, while more comfortable and lightweight, generally offer a narrower field of view. In contrast, headsets provide significantly wider FoVs, offering a more immersive experience that is especially beneficial in wide environments or when managing high information loads.

In summary, lightweight AR glasses are typically better suited for mobile, long-duration field operations, whereas headsets with larger FoV are more appropriate for stationary or semi-mobile scenarios requiring higher levels of visual immersion and interaction.

Table 4: FoV specifications of the selected AR/MR equipment.

Device	FoV (degrees)
Vuzix M400	16.8
Vuzix M4000	28
RealWear Navigator 500	20
ThirdEye X2	42
MagicLeap 2	70
Meta Quest 3	96
HTC Vive XR Elite	110
Varjo XR-4	115
Pico 4 Pro	105
Apple Vision Pro	100–120
Microsoft HoloLens 2	52

3.4 Display Resolution and Type

The display characteristics of AR/MR devices greatly influence their effectiveness in various applications. Display resolution, type, and opacity collectively determine the usability of a device by influencing what their operators can perceive.

Display resolution governs the legibility of technical schematics and component labels during equipment inspections. High-resolution displays aid in more precise and legible visualization of any necessary content.

The display type affects contrast ratios and color fidelity. While this parameter may be unimportant in some situations, use cases that require greater clarity may benefit from choosing a display type with better color optics.

Opacity is perhaps the most operationally significant parameter, creating a fundamental divide in AR/MR device capabilities. Transparent displays come with the ability to maintain environmental awareness even while operating AR/MR equipment and allow the passthrough process to be simpler than opaque displays (as there is no need to capture and project the environment image onto a screen). However, they come at a tradeoff of being more expensive to produce, leading to more expensive equipment and, overall, smaller resolutions. The data gathered on this subject can be consulted in Table 5.

Table 5: Display specifications of selected AR/MR devices.

Device	Resolution	Opacity
Vuzix M400	—	Opaque
Vuzix M4000	—	Transparent
RealWear Navigator 500	720p	Opaque
ThirdEye X2	720p	Transparent
MagicLeap 2	1440p	Transparent
Meta Quest 3	4K	Opaque
HTC Vive VR Elite	720p	Opaque
Varjo XR-4	—	Opaque
Pico 4 Pro	4K	Opaque
Apple Vision Pro	4K	Opaque
Microsoft HoloLens 2	2K	Transparent

Display technology showcases a trade-off between resolution and operational transparency. AR glasses typically offer lower resolutions and simpler display types, though some models like the Vuzix M4000 and ThirdEye X2 support transparent optics, which aid in preserving real-world visibility. MR headsets deliver higher resolutions, providing highly detailed visual feedback ideal for asset visualization or intricate work procedures.

Notably, transparent headsets such as the MagicLeap 2 and HoloLens 2 attempt to bridge this gap by offering both clarity and environmental awareness.

3.5 Sound and Voice Capabilities

Audio input and output features are crucial for effective operation and user interaction in AR/MR systems. Research such as [14] demonstrates that audio-based interactions significantly improve user attention and task efficiency compared to traditional interaction methods like touchscreens. Consequently, it is essential for AR/MR devices to possess robust and versatile audio capabilities to fully support these interaction modalities. This section outlines the sound and voice-related capabilites of selected AR glasses and MR headsets, focusing on the number and type of built-in speakers and microphones, support for external audio devices, and the presence of voice recognition capabilities.

Table 6: Sound and Voice capabilities of selected AR/MR devices.

Device	Speakers	Microphone	External Audio	Voice Recognition
Vuzix M400	1	1	Bluetooth	Yes
Vuzix M4000	1	1	Bluetooth	Yes
RealWear Navigator 500	1	4	Headphone Jack	Yes
ThirdEye X2	2	1	Headphone Jack	Yes
MagicLeap 2	2	4	Bluetooth	Yes
Meta Quest 3	2	3	Bluetooth	Yes
HTC Vive VR Elite	2	2	USB-C	Yes
Varjo XR-4	2	2	Headphone Jack	No
Pico 4 Pro	2	2	Bluetooth	Yes
Apple Vision Pro	2	6	Bluetooth	Yes
Microsoft HoloLens 2	2	5	Bluetooth	Yes

Voice interaction is uniformly supported across all models, yet headsets excel in audio fidelity, mostly due to supporting stereo audio more readily, which allows for richer audio environments and for technologies like spatial audio, which can enhance the user's experience and allow for more in-depth immersion in the AR/MR setting.

3.6 Camera Capabilities and Spatial Mapping

Cameras and spatial mapping technologies are fundamental to AR/MR devices, as they enable accurate overlay of digital content onto the real-world environment. Cameras capture visual data that serves both as media for documentation and as input for spatial mapping. Spatial mapping refers to the process by which an AR/MR device scans and digitally reconstructs its surroundings, allowing virtual objects to be placed and anchored realistically within the physical space.

A key enabler of spatial mapping is Simultaneous Localization and Mapping (SLAM), which allows the device to track its position and orientation while concurrently building a map of the

environment. This capability ensures that virtual content remains stable and correctly aligned even as the user moves. Anchoring further supports this by fixing virtual objects to specific points or surfaces in the real world, preventing unwanted drift.

Depth sensing technologies, such as Light Detection and Ranging (LiDAR) or infrared sensors, capture distance information to generate detailed 3D meshes of the environment. These meshes improve spatial awareness by enabling realistic interactions between physical and virtual elements, including occlusion and collision detection.

Table 7 summarizes the relevant specifications of AR/MR devices. The key parameters are:

- **Camera Resolution (MP):** The megapixel count of the camera, indicating image quality and detail for media capture and spatial analysis.
- **SLAM:** Indicates whether the device supports Simultaneous Localization and Mapping for real-time environment mapping and localization.
- **Anchoring:** Denotes if the device can attach virtual objects persistently to real-world locations.
- **Depth Sensing:** Indicates the presence of sensors and algorithms that capture depth data and generate 3D meshes of the surroundings to enhance spatial mapping and realistic interaction.

Table 7: Camera and Spatial Mapping Capabilities of AR/MR devices.

Brand	Camera Specs	MP	SLAM	Anchoring	Depth Sensing
Vuzix M400	RGB	13	No	No	No
Vuzix M4000	RGB	12.8	No	No	No
RealWear Navigator 500	RGB	48	No	No	No
ThirdEye X2	RGB + depth sensors	13	Yes	Yes	No
MagicLeap 2	RGB + depth sensors	12.6	Yes	Yes	Yes
Meta Quest 3	2 RGB + 1 IR depth	—	Yes	Yes	Yes
HTC Vive XR Elite	RGB + depth sensors	16	Yes	Yes	Yes
Varjo XR-4	2 RGB + LiDAR	20	Yes	Yes	Yes
Pico 4 Pro	2 RGB + 1 RGB depth	16	Yes	Yes	Yes
Apple Vision Pro	Stereo RGB + LiDAR	6.5	Yes	Yes	Yes
Microsoft HoloLens 2	RGB + depth + IR	8	Yes	Yes	Yes

Camera quality and spatial capabilities are key differentiators. While AR glasses often lack SLAM, anchoring, or depth sensing (with the exception of the ThirdEye X2), all MR headsets support these features. Varjo XR-4, with dual 20MP cameras and LiDAR, offers unmatched environmental mapping. Devices like the Meta Quest 3 and Apple Vision Pro also integrate depth sensors and SLAM for precise spatial awareness. This makes headsets significantly more capable for tasks involving AR/MR overlays in dynamic or complex environments.

Another differentiator lies in geolocation capabilities. All analyzed AR glasses are equipped with built-in GPS, enabling accurate outdoor tracking of assets and personnel, an important feature for field operations across large geographic areas. In contrast, the evaluated MR headsets do not include GPS. Instead, they rely on aforementioned techniques such as SLAM for spatial

awareness, making them more suitable for indoor environments where precise positioning is needed.

3.7 Connectivity

In PAM, reliable connectivity is essential for ensuring operational efficiency. It enables seamless data integration with asset management systems, facilitates remote collaboration, and supports a range of other critical functions.

Connectivity features are generally standardized across both AR glasses and mixed reality AR/MR headsets. Most devices support Wi-Fi, Bluetooth, and USB-C connections, providing robust wireless and wired communication options.

3.8 Artificial Intelligence Capabilities

As industries advance and PAM workflows become increasingly more complex, the integration of AI within AR/MR devices presents a significant opportunity to enhance productivity. Built-in AI capabilities enable real-time scene understanding, natural language processing, object recognition, predictive maintenance suggestions, among other tools that help enhance worflows.

This section evaluates the presence and extent of native AI capabilities in the selected AR/MR devices.

The data gathered on this subject can be viewed in table 8.

Table 8: AI Capabilities of AR/MR Devices

Device	AI Capabilities
Vuzix M400	—
Vuzix M4000	—
RealWear Navigator 500	—
ThirdEye X2	—
MagicLeap 2	2 dedicated ML cores/14 image processing cores
Meta Quest 3	Meta AI
HTC Vive XR Elite	—
Varjo XR-4	—
Pico 4 Pro	—
Apple Vision Pro	Apple Intelligence
Microsoft HoloLens 2	Cortana (deprecated)

From this overview, it is evident that AI capabilities remain unevenly distributed across the AR/MR hardware landscape. Most current AR glasses and headsets lack meaningful onboard AI functionality, relying instead on remote processing. This can introduce latency and limit real-time interaction—particularly problematic in bandwidth-constrained industrial environments.

Only a select few devices, such as the MagicLeap 2 and Apple Vision Pro, currently integrate dedicated AI hardware for local inference. These are better suited for sophisticated PAM applications that require on-device intelligence, such as autonomous diagnostics, voice-command control, and advanced environmental awareness.

3.9 Device Interactibility

Efficient interaction is essential for AR/MR hardware, not only for casual environments, but also for industrial environments where hands-free or intuitive input methods improve workflow

efficiency.

This section will present a feature-wise comparison of interaction methods available across AR/MR AR glasses and headsets. The data gathered for this section is present in table 9.

Table 9: Interactibility table for AR/MR devices.

Device	Touch Input	Voice	Hand Tracking	Controller Support
Vuzix M400	Yes	Yes	No	No
Vuzix M4000	Yes	Yes	No	No
RealWear Navigator 500	Buttons only	Yes	No	No
ThirdEye X2	Yes	Yes	No	No
MagicLeap 2	No	Yes	Yes	Yes
Meta Quest 3	No	Yes	Yes	Yes
HTC Vive XR Elite	No	Yes	Yes	Yes
Varjo XR-4	No	Yes	Yes	Yes
Pico 4 Pro	No	Yes	Yes	Yes
Apple Vision Pro	No	Yes	Yes	No
Microsoft HoloLens 2	No	Yes	Yes	Yes

Interaction methods vary widely. Smart glasses are largely limited to voice and touch, with no hand tracking or controller support. Conversely, MR headsets universally support advanced interaction modes, including gesture control and physical controllers. The HoloLens 2 and MagicLeap 2 are particularly robust, offering seamless multimodal interfaces.

This enables more intuitive and efficient workflows in PAM tasks requiring complex interaction, such as virtual equipment manipulation or remote training.

3.10 Response Latency

Response latency, shown in Table 10, is the time elapsed from when an image is captured by a headset's camera until it is displayed on the headset's screen. Essentially, it measures the delay involved in "passing through" the real-world view, processing it, and projecting it onto the display.

As shown by Fuvattanasilp et al. [15], although the study's focus was on mobile devices rather than AR/MR headsets, it revealed that users prefer devices with latency of 132ms or lower, and that latency differences become noticeable when the gap exceeds 69ms.

Low response latency is critical for effective usability in AR devices, as high latency creates a noticeable lag between the user's movements and the corresponding visual update. This lag can cause the system to feel unresponsive and severely impact its usability.

This is particularly relevant to AR/MR headsets that do not utilize transparent displays but instead rely on video passthrough or similar mechanisms to render the real world digitally, enabling AR/MR functionality.

Among headsets, the Apple Vision Pro and Varjo XR-4 show a significant advantage over the other MR devices. Notably, devices with transparent displays mostly bypass this concern, presenting real-world visuals directly, making them more latency-tolerant.

3.11 Security Certifications

In some scenarios, it is important for the device in use to be capable of enduring various unpredictable conditions. Therefore, having a guarantee of the device's resilience and toughness being

Table 10: Latency of select MR devices.

Device	Latency (ms, estimated)
MagicLeap 2	—
Meta Quest 3	39 [16]
Htc Vive XR Elite	40 [16]
Varjo XR-4	22 [10]
Pico 4 Pro	—
Apple Vision Pro	11 [16]
Microsoft Hololens 2	—

paramount. Even though a few headsets have some security certifications, AR glasses usually take the edge on this front. The data on this aspect can be observed in table 11.

Table 11: Security Certifications of AR glasses.

Device	Certifications
Vuzix M400	IP67
Vuzix M4000	IP67
RealWear Navigator 500	IP66, MIL-STD-810
ThirdEye X2	ANSI/ISEA Z87.1, IP40, EN 166*
HoloLens 2 (Industrial Edition)	ISO 14644-1 Class 5-8 UL Class I, Division 2 (Groups A, B, C, D)

For clarity, these certifications will be explained in the following itemized list:

- **IP67**[17]: Fully dust-tight and protected against immersion in water up to 1 meter for 30 minutes. Suitable for rugged environments with high dust and moisture exposure.
- **IP66**[17]: Fully dust-tight and protected against powerful water jets from any direction. Ideal for industrial or outdoor settings where devices may be sprayed or hosed down.
- **IP40**[17]: Limited protection against solid objects larger than 1 mm (e.g., tools or wires); no protection against liquids. Suitable for clean, indoor environments.
- **MIL-STD-810**[18]: U.S. military standard for environmental durability, including resistance to shock, vibration, extreme temperatures, humidity, and altitude. Ensures suitability for field and industrial use.
- **ANSI/ISEA Z87.1**[19]: American national standard for eye protection. Ensures resistance to high-velocity impacts, splash protection, and optical clarity. Commonly required in industrial and construction sites.
- **EN 166**: European standard for personal eye protection. Covers mechanical strength, dust and liquid resistance, and optical quality. Widely used in workplaces across the EU. Was replaced in November 2024 by EN ISO 16321:2022.
- **ISO 14644-1 Class 5 to 8**[20]: Certifies that the device meets the cleanliness standards for use in clean rooms. It complies with particle emission limits required in environments such as pharmaceutical manufacturing and semiconductor fabrication.

- **UL Class I, Division 2 (Groups A, B, C, D)**[21]: Ensures that the device is safe for use in hazardous locations where flammable gases, vapors, or liquids may be present. This is a North American standard for equipment used in potentially explosive atmospheres.

AR glasses have a noticeable advantage over MR headsets. Conversely, MR headsets generally lack such certifications, making them less appropriate for exposure to more abrasive conditions.

3.12 SDK and Engine Support

Engine and SDK support is crucial for good programmer experience and fast turnaround time, as support for graphical engines like Unity[22] and Unreal[23] can be a tremendous help to speed up development. The gathered information can be seen in table 12.

Table 12: Engine and SDK support of select AR/MR devices.

Device	Unity	Unreal	SDK Info
Vuzix M400	Yes	No	No
Vuzix M4000	Yes	No	No
RealWear Navigator 500	Yes	No	No
ThirdEye X2	Yes	Yes	VisionEye SDK
Meta Quest 3	Yes	Yes	Meta SDK / OpenXR
HTC Vive XR Elite	Yes	Yes	VIVE OpenXR/Wave
Varjo XR-4	Yes	Yes	Plugin Support
Pico 4 Pro	Yes	Yes	Pico SDK
Apple Vision Pro	Yes	No	No
Microsoft HoloLens 2	Yes	Yes	Native / OpenXR

Development support is comprehensive across most MR headsets, with near-universal compatibility with Unity and Unreal engines. OpenXR and proprietary SDKs further simplify integration into existing PAM platforms. AR glasses lack in this area, with only the ThirdEye X2 offering broader engine and SDK support.

4 Conclusion

The goal of this paper was to provide a comparative overview of various AR/MR devices currently available on the market, with a focus on assessing their applicability within a PAM context.

While the scope of the study is limited, several key observations can be drawn:

- Smart Glasses tend to offer more affordable and ergonomic solutions. Although they are typically more limited in terms of hardware capabilities, their lightweight design and resilience make them well-suited for use in environments that may be challenging for more delicate electronic equipment;
- AR Headsets, on the other hand, generally provide greater computational power and a broader range of functionalities. However, these advantages come with trade-offs, including higher costs, shorter battery life, increased bulk, and latency issues that are less pronounced in AR glasses.

In summary, no single category of device emerges as definitively superior. Rather, each is better suited to specific use cases and operational environments, underscoring the importance of context when selecting AR/MR equipment for PAM applications.

References

[1] Shakil Ahmed. A review on using opportunities of augmented reality and virtual reality in construction project management. *Organization, Technology and Management in Construction: an International Journal*, 11(1):1839–1852, 2019.

[2] Rajesh Ojha. Integrating digital twin and augmented reality for asset inspection and training introduction. *INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS*, 11:10, 11 2024.

[3] Vuzix. Vuzix m400 product sheet. https://vuzix-website.s3.amazonaws.com/files/Content/product-sheets/Vuzix-M400-Smart-Glasses.pdf. [Online; accessed 2025-05-10].

[4] Vuzix. Vuzix m4000 product sheet. https://vuzix-website.s3.amazonaws.com/files/Content/product-sheets/Vuzix-M4000-Smart-Glasses.pdf. [Online; accessed 2025-05-10].

[5] RealWear. Realwear navigator 500 product sheet. https://realwear.at/wp-content/uploads/2021/12/Datasheet_RealWear_HMT-2_Navigator500.pdf. [Online; accessed 2025-05-10].

[6] BorealTech. Thirdeye x2 presentation. https://borealtech.com/wp-content/uploads/2020/09/Third-Eye-X2-MR.pdf. [Online; accessed 2025-05-10].

[7] MagicLeap. Product specification version 6.2. https://www.magicleap.care/hc/en-us/article_attachments/33734272340493. [Online; accessed 2025-05-10].

[8] Meta. Meta quest 3 product page. https://www.meta.com/quest/quest-3/. [Online; accessed 2025-05-10].

[9] HTC. Htc vive elite xr product specification page. https://www.vive.com/us/product/vive-xr-elite/specs/. [Online; accessed 2025-05-10].

[10] Varjo. Varjo xr-4 product page. https://b2b-store.varjo.com/product/xr-4. [Online; accessed 2025-05-10].

[11] Pico 4. Pico 4 product specification page. https://www.picoxr.com/global/products/pico4/specs. [Online; accessed 2025-05-10].

[12] Apple. Apple vision pro product specification page. https://www.apple.com/apple-vision-pro/specs/. [Online; accessed 2025-05-10].

[13] Microsoft. Microsoft hololens 2 documentation. https://learn.microsoft.com/pt-pt/hololens/. [Online; accessed 2025-05-10].

[14] Jingya Li. Beyond sight: Enhancing augmented reality interactivity with audio-based and non-visual interfaces. *Applied Sciences*, 14(11), 2024.

[15] Varunyu Fuvattanasilp, Matjaž Kljun, Hirokazu Kato, and Klen Čopič Pucihar. The effect of latency on high precision micro instructions in mobile ar. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Tommi Björk. Apple vision pro benchmark test 1: See-through latency, photon-to-photon. https://www.optofidelity.com/insights/blogs/apple-vision-pro-benchmark-test-1-see-through-latency-photon-to-photon, 2024. [Online; accessed 2025-05-17].

[17] International Electrotechnical Commission. Ingress protection (ip) ratings. https://www.iec.ch/ip-ratings. [Online; accessed 2025-05-18].

[18] U.S. Department of Defense. MIL-STD-810G: Department of Defense Test Method Standard for Environmental Engineering Considerations and Laboratory Tests. Military Standard MIL-STD-810G, Department of Defense, United States of America, October 2008. Includes Change Notice 1 (31 July 2014).

[19] International Safety Equipment Association. ANSI/ISEA Z87.1-2020: American National Standard for Occupational and Educational Personal Eye and Face Protection Devices, 2020.

[20] International Organization for Standardization. Iso 14644-1:2015 - cleanrooms and associated controlled environments — part 1: Classification of air cleanliness by particle concentration, 2015.

[21] UL Solutions. Ul and c-ul hazardous areas certification for north america. https://www.ul.com/services/ul-and-c-ul-hazardous-areas-certification-north-america, 2025. [Online; accessed 2025-05-20].

[22] Unity Technologies. Unity Real-Time Development Platform. https://unity.com/, 2025. [Online; accessed 2025-05-17].

[23] Epic Games. Unreal engine. https://www.unrealengine.com, 2025. [Online; accessed 2025-05-17].

Chapter 14

Raw material waste reduction in a food industry manufacturing unit: a case study on equipment changes and process optimization

Contribution #34
Cite as:
Mariana Clemente, Ana C V Vieira. Raw material waste reduction in a food industry manufacturing unit: a case study on equipment changes and process optimization. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-34,
  title ={Raw material waste reduction in a food
     industry manufacturing unit: a case study on
     equipment changes and process optimization},
  author={Mariana Clemente and  Ana C V Vieira},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Raw material waste reduction in a food industry manufacturing unit: a case study on equipment changes and process optimization

Mariana Clemente[1*], Ana V Vieira[1,2,3,4]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[2] RCM2+ Research Centre for Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[3] CISE - Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal

[4] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal

*Corresponding author
marianaclemente2121@gmail.com, avieira@isec.pt

Abstract

In an increasingly competitive industrial landscape, operational efficiency, risk reduction, and mitigation are becoming strategic priorities for companies. The principles of continuous improvement emphasize that eliminating waste and improving working conditions are key practices aligned with operational efficiency.

This paper focuses on the problem of manually handling plastic film rolls in a food industry manufacturing unit, which, at the beginning of this study, was resulting in raw material waste, with a negative impact on productivity, significant ergonomic risks, and considerable costs. The study methodology was based on direct observation of the process, operational data collection, and assessment of several risks associated with this process. Based on this analysis, some improvement proposals are explored, and an investment project for new equipment is detailed.

The results presented in the full version of this paper show that the proposed project ensures employees' physical effort reduction, less film waste, and a process efficiency improvement. In addition, the solution studied contributes to reducing waste in the final product resulting from the use of non-compliant film. The technical and economic feasibility analysis confirms the short-term return on investment, reinforcing the importance of risk management integrated with operational efficiency.

Keywords: Process optimization; Quality control; Risk management; Lean manufacturing; Continuous improvement.

1 Introduction

Risk reduction and operational efficiency are gaining relevance as strategic priorities for the food industry, essential to increasing market competitiveness. This article therefore focuses on a specific challenge identified in a production unit: the manual handling of plastic film rolls in the pouch filling area. Over the years, this activity has led to raw materials waste, ergonomic risks and non-compliance with legal requirements, all of

which have a negative impact on operating costs and productivity. Studies emphasise that eliminating the manual handling of heavy loads is an effective strategy for preventing injuries and increasing efficiency in the production process.

In this context, the main research questions are as follows:

- Is it feasible to automate the handling of rollers in the pouch filling area?
- What are the ergonomic, operational and economic impacts of implementing such a system?
- Which solution offers the highest return on investment while providing the lowest risk to the company?

To answer these questions, a study was carried out, based on direct observation of the process, communication with operators, collection and analysis of waste data, identification of operational and ergonomic risks, and market and financial research.

This article is organised as follows: Section 2 presents a review of the current state of affairs; Section 3 outlines the applied methodology; Section 4 presents the main results; Section 5 discusses their implications; and Section 6 concludes with key recommendations and future perspectives.

2 State of the art

As industries strive for greater efficiency and reduced risk, process automation and ergonomic improvements play a crucial role in improving safety, productivity, and waste reduction, especially in repetitive tasks. In the food sector, manual handling of heavy loads is a significant source of both waste and worker injuries. By examining the latest research and innovations in these areas, this section lays the groundwork for the solutions later discussed.

2.1 Manual handling of loads, industrial ergonomics and process automation

Manual handling of loads remains one of the primary contributors to ergonomic risks in many industries, especially in environments that require repetitive, physically demanding tasks. In the context of the food industry, where high productivity is often paired with physically intense operations, the risks to worker health and safety are exacerbated. This section explores the impact of manual labour on both worker health and operational efficiency, emphasising the importance of automation as a solution to reduce physical strain, improve safety, and enhance productivity.

2.1.1 Ergonomic risks and applicable legislation

Manual handling of loads is considered to be one of the main sources of injuries in an industrial environment, injuries that directly affect and have a major impact on the company's productivity and, above all, on the health of workers. In the food sector, where there is a high level of process repeatability, intense rhythms over a long period of time and large loads to handle, the ergonomic risk is considered to be particularly high.

According to the ISO 11228-1 standard, limits are set for the manual transport and handling of heavy loads, and it is recommended to automate and mechanise processes and eliminate physical effort whenever weights exceed 25 kilograms [1]. On the other hand, ISO 45001 reinforces the importance of identifying hazards and continuously controlling risks associated with occupational health [2]. The main aim of this standard is to provide a framework for organisations to effectively manage safety in the workplace, preventing work-related injuries and health problems.

Adellian and Noeroel carried out a study in which they identified high ergonomic risks in tasks involving the manual handling of weights, with a high incidence of musculoskeletal complaints among operators [3]. The analysis concluded that the workstations where these workers operated

presented a high risk of injuries and accidents at work, requiring immediate corrective measures.

An ergonomic risk assessment associated with handling heavy materials in a food industry reveals a high level of exposure to risk factors such as repetitive movements, incorrect postures and excessive effort. It is therefore recommended that manual processes be replaced by automated and more technological solutions [3]. As a result, there is an urgent need for intervention through the implementation of lifting trolleys and mechanical support in critical industrial tasks [4].

2.1.2 Training and monitoring operators

Implementing this type of equipment is not only important for organisations as it acts as a lever for innovation and technological progress, but also for operators as they are the ones who are in contact with the processes on a daily basis. Thus, the opinion of and direct contact with workers is increasingly important, confirming that their participation in validating new equipment makes the investment more sustainable. In this way, the improvements implemented result not only from the elimination of physical effort, but also from the acceptance of change on the part of the workers, who tend to feel more comfortable due to improved working conditions and daily tasks [5]. Therefore, resistance on the part of operators to the introduction of new technologies can be overcome by accompanying them and training them on an ongoing basis, making them feel involved in the process so that they can express their fears and main difficulties. This type of training and commitment is essential for workers to accept change and for the end result to be success.

2.1.3 Automation as a strategic answer

In addition to ergonomic risks, it is also important to assess whether the equipment implemented is suitable for the factory environment in question. The use of Conformité Européenne (CE) labelled materials, washable materials and materials with controlled mobility were essential conditions for reducing risk factors. Comparing the risks and physical effects of using automatic forklift trucks and manual forklift trucks showed clear reductions in muscle fatigue indices and a great improvement in the execution of tasks, with the result that operators feel highly comfortable and safe at their workstation. The automation of industrial processes has established itself as an effective response to ergonomic and efficiency challenges in organisations, with clear benefits in demanding environments such as the food sector. Replacing manual physical force with automated equipment contributes significantly to reducing the risk of injury, increasing productivity, standardisation between shifts and reducing waste [4] [6].

Several studies have addressed the risks associated with manual handling of loads in a factory environment and have questioned the use of technologies such as manipulators to assist processes. It was concluded that implementing this type of technology would also increase the reliability of the process, reducing operating times and the operator's effort load [7].

With technological evolution and innovation in all industrial sectors, it is essential to see automation as a fundamental pillar in the organisation and planning of work. With the introduction of Industry 4.0, a way of working together with more digital production processes was introduced to the market, with real-time process control allowing for greater efficiency and traceability. However, the latest Industry 5.0 approach focuses on reintegrating the human factor into processes, promoting collaboration, work and continuous training between humans and intelligent machines [8]. This perspective focuses on the use of technology as a support tool, taking on heavy, repetitive and dangerous tasks, while the human being assumes all responsibility, critical thinking and control over the machine. This new way of looking at and studying processes values safety, well-being in the workplace and the creation of more sustainable working environments for the company and the operator [8] [9] [10].

2.2 Reducing waste in production lines

With the constant need to guarantee more sustainable and consistent processes, not just for economic reasons, reducing waste has increasingly become a priority at industrial level and is one of the most important principles of Lean Manufacturing. In sectors such as the food industry, controlling waste is a fundamental step in ensuring quality for customers and competitiveness against competitors. With this, repetitive activities that do not add value must be eliminated or minimised. The eight classic wastes identified in the Lean Philosophy are [11]:

- Overproduction;
- Excess inventory;
- Defects;
- Waiting time;
- Excessive processing;
- Unnecessary transport;
- Unnecessary movements;
- Poor utilisation of human potential.

Thus, over the years, various approaches have been studied in order to reduce or even eliminate losses associated with transport, excessive handling, quality defects or even a lack of standardised procedures. With this, the literature shows that the application of Lean methodologies makes it possible to structure this analysis in a critical and logical way in order to find more efficient solutions [12]. In more automated environments, teamwork between Lean practices and advanced support technology, such as automated handling devices, presents great added value for organisations and the efficiency of their production processes [13]. Practical applications of these tools in real-life contexts show that reducing waste doesn't just require technical changes, but also organisational and cultural changes [14]. Recent studies reinforce the importance of integrating Lean with automation in order to efficiently reduce waste while increasing the efficiency of production lines and their processes [15]. With the increased focus on sustainability in industries, they are forced to respond to environmental and resource efficiency demands, trying to optimise the use of their raw materials as much as possible. Thus, the introduction of Lean tools in operations combined with sustainable strategies makes it possible to significantly reduce waste in the factory environment, increasing productivity, process standardisation and environmental awareness among operators [16]. Results like these show that waste reduction must be approached with a view to production efficiency and environmental sustainability.

2.3 Feasibility and risk analysis of equipment investments

When we talk about industrial investments and their viability, in addition to the direct costs and benefits, the associated risks, uncertainties and strategies must be considered. The literature has been consolidating economic and financial indicators and methods to support investment decisions. These include the Net Present Value (NPV), the Rate of Return (IRR), the Payback Period and the Equivalent Annual Result (ERA). These indicators make it possible to analyse the expected return on an investment in a simple and effective way [17] [18].

Risk analysis is playing an increasingly important role in the viability of investments. In contexts of high uncertainty or projects with multiple variables, it is increasingly important to assess the sensitivity of results to different scenarios [19]. The role of cost-benefit analysis as the most important method in evaluating investments emphasises the importance of valuing the benefits, the return horizon and long-term sustainability [20].

From an academic and technical point of view, investments in assets must take into account their

nature as well as the type of cash flow they generate and must be evaluated in a coherent manner according to their risk and time. In short, the evaluation of industrial investments must consider economic indicators with risk analysis techniques, guaranteeing a balance between return and uncertainty.

3 Methodology

The methodological approach used in this study combined direct observation of processes on a production line with the analysis of operational data, risk analysis and also an economic study in order to understand whether or not the investment would be viable. The aim was to gain an in-depth understanding of the impact of manual handling of rolls of plastic film and to support a proposal for more automatic improvement. The methodology stages were as follows: direct observation and characterisation of the process; data collection and analysis; risk analysis; market research; economic and financial evaluation.

The first stage consisted of direct observation of the process in a real production context, with the support of the operators in the pouch filling area. During the three shifts, photographs, movement sequences and critical risk points were recorded. After analysing the current process, four critical points of impact were identified on the roll that cause creases and, consequently, loss of raw material. In a second stage, plastic film waste was quantified based on production records and real samples taken between 2025-02-20 and 2025-05-01. The samples from 9 different days served as the basis for calculating an estimate of the average daily, monthly and annual waste in the production area and also for calculating the direct financial impact associated with the process. The data collected made it possible to establish metrics such as:

- Kilograms [kg] of clean film wasted annually;
- Average number of pouches not produced with this waste;
- Annual direct cost wasted;
- Number of rolls wasted.

The data was analysed using spreadsheet tools, namely Microsoft Excel, where conversion formulas and basic statistical analysis (average, sum, projected annual estimates) were applied. For the conversion of waste values, an actual average yield of 4,371 pouches per roll of film was used, based on internal production records. In the next step, the risks associated with not realising the improvement were identified, such as legal non-compliance due to the use of non-CE marked equipment and ergonomic risks for operators. In parallel, the risks of implementing the new solution were analysed, such as operator resistance to change or even technical failures of the equipment. All this data was organised into the impact matrices assessed in Section 5. With this in mind, several European suppliers with experience in industrial automation and load handling were contacted. The selection of suppliers was based on four criteria: i) geographical location in Europe, in order to avoid logistical and commercial barriers; ii) ability to adapt the solution to the production process; iii) willingness to carry out practical tests; and iv) guarantee of compliance. The process included analysing quotes, technical specifications and face-to-face tests. This approach made it possible to compare technical compatibility, safety and certification, prices and delivery times, as well as flexibility and technical support between suppliers from different countries. Finally, an economic and financial evaluation was carried out to test the viability of the investment. Indicators such as: Net Present Value (NPV); Internal Rate of Return (IRR); Profitability Index (PI); Annual Equivalent Result (AER) and Payback Period (PR) were calculated. These indicators make it possible to clearly and objectively measure the expected return, the efficiency of the investment over time and the associated risk. Table 1 shows the formulae used to calculate the viability of the investment.

The indicators described above were later applied in Section 4, using real data from the plant in order to validate the attractiveness of the investment.

In order to consolidate the steps taken throughout this study, the Table 2 summarises the stages of the methodology applied, from direct observation to financial validation.

The structure of the methodology described above made it possible to ensure a complete approach, combining technical, legal and economic analyses. The link between these three dimensions proved essential to justify the proposal for improvement presented in this article.

Table 1 - Table with financial indicators

Indicator	Description	Objetive
Net Present Value	$\text{NPV} = \sum_{k=m}^{n} \frac{CF_k - I_k}{(1+i)^k} + \frac{RV}{(1+i)^n}$	Value created by the project
Internal Rate of Return	$NPV_0 = \left[CF\ \frac{(1+\text{IRR})^n - 1}{IRR(1+IRR)^n} + \frac{RV}{(1+IRR)^n}\right] - \left[\sum_{k=m}^{k=n} \frac{I_k}{(1+IRR)^k}\right]$	Internal profitability of the project
Profitability Index	$PI = \left[CF\ \frac{(1+i)^n - 1}{i(1+i)^n} + \frac{RV}{(1+i)^n}\right] / \left[\sum_{k=m}^{k=n} \frac{I_k}{(1+i)^k}\right]$	Investment efficiency
Payback period	$k = m-1 \rightarrow S_k = 0$ $m \leq k < n \rightarrow S_k = S_{k-1} + \frac{CF_k - I_k}{(1+i)^k}$ $k = n \rightarrow S_k = S_{k-1} + \frac{CF_k - I_k}{(1+i)^k} + \frac{RV}{(1+i)^n}$	Speed of return
Equivalent Annual Result	$EAR = \text{NPV}\frac{i(1+i)^n}{(1+i)^n - 1}$	Annual equivalent value

Table 2 – Stages, description and purpose of the methodology used

Stages	Description	Objetive
Direct observation	Observation of the process with the operators	Identification of impact moments and ergonomic risks
Data collection	Real samples and internal records	Quantification of waste and its associated cost
Risk analysis	Analysing current and expected risks	Identification of legal, operational and safety consequences
Market research	Contact with three different suppliers	Analysis of technical and automated solutions and respective budgets
Financial evaluation	Application of NPV, IRR, IR, PR models	Checking the feasibility of the investment

4 Results

The results obtained in this study show the negative impact of manually handling rolls of film in an industrial environment, as well as the benefits expected from adopting an automated solution. It was therefore possible to quantify the impact in three dimensions: operational efficiency, safety and ergonomics, and operating costs. This section presents the main indicators associated with each of these areas and compares them with the results expected after implementing the proposed solution.

Manually changing rolls of plastic film in the filling area involves several consecutive steps with direct physical contact, and four critical steps have been identified where the roll is deformed. These impacts cause creases in the film, which compromises the conformity of the final product and can lead to bursts during transport to the customer and, as a consequence, complaints. Through direct observation and dialogue with the operators, it was found that each of these transfers not only results in a significant waste of film, but also involves physical effort, the risk of falling and potential compensation for the operators due to poor working conditions. Despite the experience of the operators who have made the process more agile by preparing the rolls in advance, the process still depends on various physical manipulations that are not standardised. These deviations from the ideal process have a direct impact not only on the conformity of the final product, but also on the fluidity of the process, since at each changeover it is necessary to remove the first few metres of creased film in order to avoid future problems, as shown by the Figure 1.



Figure 1 – Waste of clean film due to creases

This additional activity represents a consumption of time and material that is repeated cyclically and could be reduced or eliminated.

In addition, this process is not standardised between shifts, which leads to variations in execution times, the methods used and, consequently, the results obtained. Differences like these can jeopardise the process and increase the likelihood of human error. By adopting an automated system, it is hoped to improve the efficiency of the process, resulting in benefits such as a reduction in the number of manual steps, standardisation of the roll change process, greater safety for operators and a reduction in film losses associated with creases, which directly impacts productivity and reduces the amount of plastic to waste. The operational results show that the current process, although functional, has room for improvement in terms of consistency, productivity and impact on quality. This equipment provides an effective response not only to reduce waste, but also to standardise and stabilise the process.

The ergonomic analysis of the current roller change process proved to be one of the most critical aspects of the study. The handling of loads with significant weight, combined with the repetition of movements and the lack of adequate mechanical support, represents a high risk to the operators'

health and a clear non-compliance with good safety practices at work. Currently, each roll to be handled weighs approximately 98kg, to which is added a 20kg metal shaft, totalling 119kg per handling. This figure is well above the limits recommended by ISO 11228-1, the international standard governing the handling of loads. According to this standard, loads of more than 25kg should be avoided or mechanised, especially when lifting is done repeatedly and in forced postures. Throughout the process, operators make repetitive movements to lift, rotate and transfer the rollers, in unstable positions and needing to compensate for the weight with the strength of their torso and arms. The absence of an automated and certified system increases the risk of injury, especially to the lower back, physical or mental fatigue or accidents at work, in the case of falling rollers or falls due to exposed rails. In addition to these physical risks, this situation puts the company in a fragile legal position, since the trolley currently in use has no CE marking, instruction manual, technical documentation or maintenance records. In this way, the current process violates the requirements established by ISO 45001, the reference standard for occupational health and safety management systems. Replacing the manual process with an automated one will eliminate the need for physical effort on the part of the operators, reducing the risk of injury. The equipment proposed in Figure 2 is battery-powered, has a movement system with wheels and a brake, and includes intuitive controls to control the height and rotation of the roller without direct contact.



Figure 2 – Example of a budget roller manipulator

The adoption of the new equipment is expected to have a very positive impact on the well-being of operators, by promoting better working conditions and reducing the physical effort required, which highlights an urgent need for intervention.

Several European suppliers specialising in industrial automation were contacted and assessed. Based on the criteria of location, ability to adapt to the production process, legal compliance and post-purchase technical support, it was possible to select three suppliers with the best cost-benefit ratio and the most technical flexibility.

Based on samples taken between 2025-02-20 and 2025-05-01, and considering daily production records, an average daily waste of 24.89kg of clean plastic film was estimated, which represents an annual loss of 6,296.89kg. This loss corresponds to the non-production of 209,896 pouches, as illustrated in Table 3.

Table 3 – Quantifying plastic film waste

Period	Wasted film [kg]	Wasted rolls	Non-produced pouches
Daily	24,89	0,19	829,63
Monthly	547,58	4,16	18 251,85
Annual	6 296,89	48,02	209 896,30

The cost associated with raw material waste was estimated at 50762.81€ per year, based on the unit price of film (8.06€/kg). Based on this figure, the total cost associated with the annual loss of film is 50762.81€. The economic and financial analysis was carried out based on the following assumptions: i) Initial investment of €25,890.00, the highest value of the quotes received; ii) 3-year analysis period; iii) Annual operating cash flow of €50,762.81; iv) 10% discount rate; and v) Residual value of €2,500.

Applying the formulae, the following values were obtained Table 4.

Table 4 – Financial indicators and results obtained

Indicator	Result
Net Present Value (NPV)	102 227,88€
Internal Rate of Return (IRR)	188,67%
Profitability Index (PI)	4,95
Annual Equivalent Result (AER)	41 107,34€/year
Payback Period (PR)	6 months e 22 days

These indicators show that the investment is highly attractive, since a Net Present Value of 102,227.88 euros indicates a significant gain on the investment. An Internal Rate of Return of 188.67% shows an internal return well above the minimum rate required. The payback period of less than 7 months shows a very rapid recovery of the capital invested and, finally, the Equivalent Annual Result and the Profitability Index reinforce that the project is efficient over time. A sensitivity analysis was carried out to test the robustness of the project to variations in the main financial parameters. The impact of the discount rate (i) on the following indicators will be analysed: Net Present Value (NPV), Profitability Index (PI) and Payback Period (PR). To carry out the sensitivity analysis, we will assume a variation to the downside. In other words, the variations allowed will always be in the direction of jeopardising the viability of the investment. We chose to test three scenarios, with rates different from the initial rate of 10 per cent, so Table 5, shows the values of the different indicators with a change in the update rate (i), while maintaining the other parameters.

Table 5 – Sensitivity analysis with update rate variation

Discount rate (i)	Net Present Value (NPV)	Profitability Index (PI)	Payback Period (PR)
5%	114 509,32€	5,42	0,54 years
10%	102 227,88€	4,95	0,56 years
15%	91 656,71€	4,54	0,59 years
20%	82 487,68€	4,19	0,61 years

The sensitivity analysis presented in the table shows that the project remains economically viable even with negative variations in the discount rate. As the rate increases from 5% to 20%, the Net Present Value (NPV) and the Profitability Index (PI) decrease. However, even in the most adverse scenario (i=20%), the NPV remains high and the IR remains well above 1, and the Payback Period is still less than 1 year, which proves the viability of the proposal.

These results show that the investment decision is not very sensitive to variations in the discount rate, reinforcing the security and attractiveness of the project. As a result, even under more conservative financial scenarios, the solution continues to show a rapid return on invested. With this financial analysis, it is once again possible to reinforce the viability of implementing the automated system, showing a direct impact on current losses, a significant reduction in costs and a high return potential for the company.

In short, the combination of operational, ergonomic and financial gains reinforces the strategic impact of this automated implementation as a strategic, safe and economically viable measure, with positive and sustainable impacts in the short term.

5 Discussion

The results obtained in this study confirm the trends identified in the state of the art regarding the impacts and consequences of manual handling of loads in industrial environments. The data obtained shows that the manual handling of heavy rolls has a direct impact not only on operator safety, but also on the amount of plastic wasted, as already evidenced by other studies in the State of the Art. When comparing the data from the plant with the practices recommended by the ISO 11228-1 standard, it was found that the operators were handling rolls with weights much higher than those permitted, reinforcing the urgency of automating the process. Investing in a new piece of equipment with lifting capacity and CE marking is a good solution to this problem and ensures that the company complies with legal and technical requirements. When it comes to efficiency on the production line, this study reveals large losses of plastic film associated with creases due to poor handling of the rolls. This waste exceeds 6,000kg per year, which jeopardises operational resources. The state-of-the-art shows that these types of losses can be combated using Lean methodologies in conjunction with automation. Finally, the economic and financial analysis validates the viability of the investment, with very favourable indicators. With regard to the research questions formulated in the introduction, at the end of this study it is possible to answer that:

- Is it feasible to automate roller handling? - Yes, both technically and economically.
- What are the ergonomic, operational and economic impacts? - Reduced physical effort, improved process standardisation and rapid financial return.
- What is the best solution in terms of return and risk? - The automated solution with mechanical support proved to be more balanced, minimising legal risks and accidents at work, while maximising efficiency.

6 Conclusions

The study for this article arose because of the problem identified in the manual handling of rolls of plastic film in a food sector organisation, which caused waste of raw materials, ergonomic risks for operators and non-compliance with legal regulations. In order to solve this problem, a methodology was applied based on direct observation, data collection and analysis, risk assessment, market research and economic and financial analysis. The results showed that this implementation of the automated system made it possible to mitigate the physical effort of operators, reduce plastic waste by more than 6000kg/year and ensure compliance of the procedures adopted with ISO 11228-1 and ISO 45001 standards. This study contributes to the state of the art by quantifying the operational, ergonomic and economic impact in a real environment, highlighting the benefits of Lean and

automated solutions. Among the advantages of the study are the use of real factory data and the multidimensional approach to the problem; as a limitation, the application is restricted to a single production process and does not cover different products or packaging formats. For future work, it is recommended to accept the automated solution and analyse the way operators work, studying the adaptation of the system to different types of rolls with different weights. It is also recommended to explore integration with technologies associated with Industry 4.0 and 5.0, promoting greater efficiency, traceability and safety.

Acknowledgments: We would like to thank the operators and production team for their essential collaboration in collecting and validating the data. We appreciate the support provided by the RCM2^{+} (Research Centre for Asset Management and Systems Engineering) and by ISEC (Coimbra Institute of Engineering). Our thanks are extended to the company that provided the opportunity to develop and execute the project, fundamental to the development of the work, as well as all the personnel involved.
This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Projects UIDB/04131/2020 and UIDP/04131/2020.

References

[1] I. Standard, "ISO 11228-1-Ergonomics-Manual handling-Lifting, lowering and carrying," 2021.

[2] I. Standard, ISO 45001-Sistema de gestão de saúde e segurança ocupacional-Requisitos com orientação para uso, 2018.

[3] A. Nugraha e N. Widajati, "Ergonomic risks in manual material handling activities and musculoskeletal disorders complaints in the animal feed industry production area in East Java, Indonesia," *World Journal of Advanced Research and Reviews,* 2024.

[4] R. Umar, N. Ahmad, I. Halim, P. Lee e M. Hamid, "Design and Development of an Ergonomic Trolley-Lifter for Sheet Meatl Handling Task: A Preliminary Study," *Safety and Helth at Work,* 2019.

[5] N. Dewantari e L. Herlina , "LITERATURE REVIEW: HUMAN FACTORS AND ERGONOMICS (HFE)," October 2022.

[6] F. Sarmiento-Ortiz, B. Topón-Visarrea, C. Iza, J. Morales e V. León, "Design of a load elevator as a measure to reduce ergonomic risks," 2024.

[7] L. Isaza e K. Cepa, "Atomation and augmentation: A process study of how robotization shapes tasks of operational employees," *European Management Journal,* 2024.

[8] T. Callari, Y. Curzi e N. Lohse, "Realising human-robot collaboration in manufacturing? A journey towards industry 5.0 amid organisational paradoxical tensions," *Technological Forecasting & Social Change,* 2025.

[9] J. Ribeiro, R. Lima, T. Eckhardt e S. Paiva, "Robotic Process Automation and Artificial Intelligence in Industry 4.0 - A Literature review," 2020.

[10] Y. Dmytriyev, M. Carnevale e H. Giberti, "Enhancingflexibilityandsafety:collaborativeroboticsformaterial handling in end-of-line industrial operations," *Procedia Computer Science,* 2024.

[11] P. Rewers, J. Trojanowska e P. Chabowski, “Tools and methods of Lean Manufacturing - a literature review,” June 2016.

[12] I. Widiwati, S. Liman e F. Nurprihatin, “The implementation of Lean Six Sigma approach to minimize waste at a food manufacturing industry,” *Journal of Engineering Research,* 2025.

[13] L. Junior, C. Santos, R. Chamie, R. Pierezan, R. Loures, P. Santos, E. Costa e P. Lima, “Automated System Gains in Lean Manufacturing Improvement Projects,” *Procedia Manufacturing,* 2020.

[14] A. Zavala, “Lean Tools implementation for scarp reduction in a production line - A case study,” May 2022.

[15] S. Setiawan, I. Setiawan, C. Jaqin, H. Prabowo e H. Purba, “Integration of Waste Assessment Model and Lean Automation to Improve Process Cycle Efficiency in the Automotive Industry,” 2021.

[16] Y. Kazancoglu, E. Ekinci, Y. Ozen e M. Pala, “REDUCING FOOD WASTE THROUGH LEAN AND SUSTAINABLE OPERATIONS: A CASE STUDY FROM THE POULTRY INDUSTRY,” *Journal of Business Management,* 2020.

[17] A. Konstantin, “Feasibility Study for an investment of an industrial plant evaluating Red Mud by an innovating method. A case study,” 2015.

[18] P. Lucas, T. Queiroz, O. Sabbag e D. Santos, “Analysis of economic viability for replacement of equipment in sugarcane production: a case study in cultural treatment operations,” *Revista Capital Científico - Eletrônica,* 2022.

[19] V. Platon e A. Constantinescu, “Monte Carlo Method in risk analysis for investment projects,” *Procedia Economics and Finance,* 2014.

[20] T. Weerasinghe, . D. Amaratunga e R. Haigh, “Appraising investments in Disaster Risk Reduction (DRR): A systematic literature review,” *Progress in Disaster Science,* 2025.

Chapter 15

3D Modeling of Railway Components for Augmented Reality-Based Maintenance Training

Contribution #36
Cite as:
Hugo Pereira, Ana Malta, Nuno Cid Martins, Mateus Mendes. 3D Modeling of Railway Components for Augmented Reality-Based Maintenance Training. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-36,
  title ={3D Modeling of Railway Components for
     Augmented Reality-Based Maintenance Training},
  author={Hugo Pereira and  Ana Malta and  Nuno Cid
     Martins and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

3D Modeling of Railway Components for Augmented Reality-Based Maintenance Training

Hugo Pereira[1], Ana Malta [1,2], Nuno Cid Martins [1], Mateus Mendes [1,2,3]

[1]Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[2]RCM^{2+} Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[3]Institute of Systems and Robotics,
Department of Electrical and Computer Engineering,
University of Coimbra, 3030-290 Coimbra, Portugal
a21220079, a21230211, ncmartin, mmendes@isec.pt

Abstract

This paper addresses the challenge of improving technical training for railway maintenance personnel by leveraging the use of 3D modeling and Augmented Reality (AR) technologies. With the modernization of railway infrastructure, the increasing push towards digitalization, and the shortage of highly skilled technicians, this work proposes the development of accurate virtual representations of critical railway components to support immersive, AR-based training scenarios. The methodology involved a comparative evaluation of leading 3D modeling software, including Blender, Cinema 4D, and some Autodesk products, ultimately selecting Blender for its accessibility, open-source nature, and community-driven support. Key components of the pantograph system, such as brushes and structural elements, were modeled in accordance with EU standards EN 50119 and EN 50367. The resulting 3D models are also designed for compatibility with Building Information Modeling (BIM) systems, enhancing their applicability in integrated infrastructure management. The findings underline the potential of combining the use of 3D modeling with AR to modernize maintenance training and contribute to the digital transformation of the railway sector.

Keywords: 3D Modeling; Augmented Reality; Railway Maintenance.

1 Introduction

Constant technological evolution has caused several impacts on industrial sectors, where the railway industry is no exception. The modernization of railway infrastructures and the constant training of highly specialized technicians are essential to ensure the safety of railway transport systems. In this sense, the use of innovative technologies, such as Augmented Reality (AR) with the aid of three-dimensional (3D) modeling, appears as a promising solution to assist technicians in their tasks. Training technicians using immersive environments allows professionals to simulate real interventions on complex components, such as the pantograph, without the need to have physical access to the equipment. It can improve cost reduction, increase productivity, optimize learning processes, in addition to other advantages [1].

This paper aims to research how 3D modeling of railway components, integrated with AR technologies, can contribute to the modernization of technical training in the railway. The focus is on modeling the pantograph components, namely the brushes and the base structure, according to European standards EN 50119 and EN 50367, ensuring technical compliance and realism in the models developed.

The methodology adopted involved a comparative analysis between several 3D modeling applications, namely Blender, Cinema 4D, Autodesk 3ds Max, Maya and Fusion 360, evaluating criteria such as versatility, cost and learning curve. After this analysis, Blender was selected because it is free, open-source, and because of its strong integration with the BlenderBIM plugin, which allows converting models to IFC format, facilitating interoperability with other Building Information Modeling (BIM) platforms.

This paper is structured as follows: Chapter 1 presents the introduction to this paper; Chapter 2 presents the State of the Art related to the use of AR and 3D in maintenance; Chapter 3 presents an analysis and selection of modeling tools, the railway components that will be modeled and the BIM integration; Chapter 4 presents the 3D models with a brief explanation on how we got to the final object, an explanation of how to export Blender objects to IFC format that can be used in BIM applications and how we can use the Blender objects to incorporate in AR; And finally Chapter 5 presents the conclusion of the work that was done.

2 State of the art

One of the most common problems in maintenance teams is the lack of young technicians with specialized knowledge. With the entry of new employees into companies, these technicians, even if they have extensive knowledge about machine maintenance, may not have enough experience to perform medium to difficult maintenance tasks. Machine and component manufacturers are collaborating closely with maintenance teams, asking for guidance and directly taking on some of the tasks. Companies often have to transport employees from machine and parts manufacturers from all over the world to make corrections, which are often quick and could have been done locally. Unfortunately, some of the failures are difficult to predict and, in these cases, maintenance requires a detailed analysis of the problem in question, followed by the intervention of highly specialized technicians to carry out the repair of the machine. This type of assistance can be quite expensive, both for the machine owner and for the company providing the maintenance service, and sometimes involves a long waiting time until the specialist arrives at the location where the intervention is carried out [2].

In order to deal with these difficulties, several technologies have been developed with the aim of improving training and maintenance support. AR and three-dimensional (3D) modeling stand out as particularly promising tools, offering immersive, interactive learning experiences without the need to be at the training site or have physical access to the equipment being maintained. These technologies allow the simulation of technical interventions in virtual environments that accurately replicate real maintenance scenarios.

The use of development kits such as ARKit (Apple) and ARCore (Google) has been instrumental in creating AR mobile applications. Mao Lin [3], for example, proposed an architecture that combines AR with BIM models, allowing technicians to visualize Heating, Ventilation and Air Conditioning (HVAC) systems in buildings through mobile devices (figure 1). The developed application, integrated with Unity and based on ARKit/ARCore, allows aligning virtual content with real space and offers intelligent navigation capabilities within buildings, facilitating inspection and maintenance tasks.

Another relevant approach was developed using the Metaio tool, where the authors of the



Figure 1: HVAC view through the app [3].

article [4] created a prototype that allows the monitoring of maintenance tasks through AR. The application consists of a system that presents step-by-step instructions in the technician's field of vision (figure 2), using CAD models and sensors to guide actions safely and accurately. This methodology has proven effective in reducing errors and increasing the autonomy of less experienced operators.



Figure 2: Example using Metaio [4].

In turn, the Vuforia system has been used to superimpose informative elements, such as texts, 3D models or videos, directly onto physical components (figure 3). Masoni *et al.* [2] propose the use of a communication logic, applied between experienced and beginner operators, where the first could send visual instructions that would be presented in AR on the second's device, facilitating remote interventions with a high degree of effectiveness.

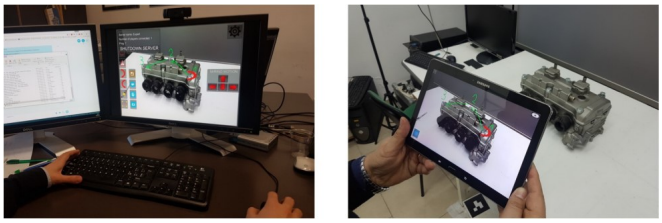


Figure 3: Example using Vuforia [2].

Additionally, the YOLOv5 Convolutional Neural Network (CNN) has been applied for automatic real-time component recognition based on images captured by AR glasses. This technology allows parts to be identified and actions to be suggested to the technician, offering contextual-



Figure 4: Example when detecting objects [5].

ized and precise assistance. Malta *et al.* [5],performed experiments where YOLOv5 was able to identify multiple automotive engine components with high accuracy (figure 4), and this approach is easily adaptable to other industrial domains such as railway industry.

Despite the potential of these solutions, it appears that most existing studies focus on generic industrial contexts, such as building maintenance, HVAC or automobiles, and there is still a significant gap in the application of these technologies to the railway sector. The integration between 3D modeling of specific railway components, such as the pantograph, with BIM and AR systems, is still little explored.

This document seeks to fill this gap, proposing the creation of 3D models of railway components compatible with BIM and optimized for AR training scenarios. In doing so, we contribute to the modernization of technical training processes and the digital transition of railway maintenance.

3 Methodology

The study methodology adopted in this paper was structured in three main phases: Analysis and Selection of Tools; Railway Components Modeling based on real components and integration of models with Building Information Modeling (BIM) systems, with a view to their future application in training contexts with AR.

3.1 Tool Analysis and Selection

At this stage, there are several software programs on the market that can be used in the 3D modeling of pantograph components. Among the most used in the industry, we chose Blender, Cinema 4D, Autodesk 3ds Max, Autodesk Maya, Autodesk Fusion 360 and SolidWorks for the study. This study was based on:

- Cost (licensing vs. free);
- Animation;
- Rendering
- Simulation
- CAM (Computer Aided Manufacturing)
- Support for exporting interoperable formats (such as IFC).

Table 1: Software Comparisons.

Software	Pricing	Animation	Rendering	Simulation	CAM (Manufacturing)	IFC Support
Blender	Free	✓	✓	✓	×	✓
Cinema 4D	€71.64/month	✓	✓	✓	×	×
Autodesk 3ds Max	285€/month	✓	✓	✓	×	×
Autodesk Maya	285€/month	✓	✓	✓	×	×
SolidWorks	2682€/year	×	✓ (basic)	✓	✓	×
Autodesk Fusion 360	97€/month	×	✓ (basic)	✓	✓	×

Based on this comparative analysis (summarized in Table 1), Blender was selected because it is a free, open-source tool with active community support. Its compatibility with the BlenderBIM add-on was also a decisive factor, as it allows direct conversion of Blender models to IFC format, essential for integration with BIM platforms.

3.2 Railway Components Modeling

In the second phase, three-dimensional modeling of various components of a pantograph was carried out, based on real images. The lack of detailed technical drawings was overcome through the use of the fSpy tool, which allowed the perspective of photographs to be calibrated (figure 5 and objects to be reconstructed.

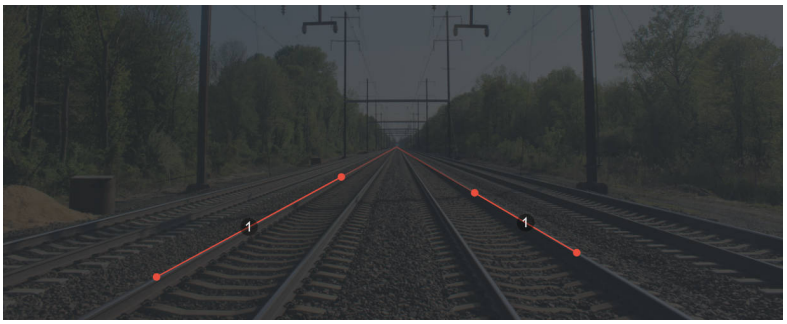


Figure 5: Example of using two points to define the X axis [6].

The following elements were modeled for this paper:

- Basic structure of the pantograph;
- Brushes;
- Insulators;
- Disconnector;
- Pantograph head.

During modeling, in Blender, techniques such as extrusion and scaling of profiles, application of modifiers (mirror, boolean, array) and conversion of Benzier and/or Nurbs curves to meshes were applied. The brushes were modeled based on the European standard EN 50367, ensuring that the virtual model complies with the technical requirements demanded in the railway sector.

3.3 BIM Integration

The third phase involved preparing the Blender models for use in BIM contexts, using the BlenderBIM add-on to associate technical information (material, function, geometric properties) with 3D objects. These models, exported in IFC format, are designed so that they can be integrated into BIM platforms and also into AR-supported training scenarios.

This process aims to improve the visualization of pantograph components, both in BIM environments and in immersive scenarios, promoting more interactive, well-founded and safe training for maintenance technicians in training.

4 Results

Based on the methodology described in subchapter 3.2, several 3D models of pantograph components were developed. These models were created using Blender software, based on real images and technical standards in the sector, such as EN 50119 and EN 50367.

4.1 3D Modeling Developed

The following subsections will present the models created within the scope of this paper, in which each model presents the modeling process (techniques and modifiers used) and, finally, an image of the model in 3D will be presented.

4.1.1 Pantograph Structure

Based on visual references taken from real images, the modeling of this object used basic modeling techniques, such as profile extrusion and for modifiers, the mirror modifier to make a copy of half of the object, so that it is symmetrical, and the boolean modifier to unite all the objects into one.

Below is an image of the three-dimensional model (figure 6) developed for the basic structure of the pantograph, highlighting the main volumes that make up the support base and the respective fitting areas for the upper and lower components of the pantograph.

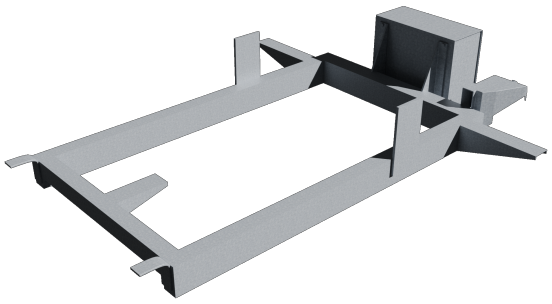


Figure 6: Basic structure of the Pantograph.

4.1.2 Brushes

Like all other modeling carried out in this paper, it was carried out with the aid of reference images of pantographs used on the railway. For the brushes, the modeling process began with the creation of Bézier lines in order to capture the geometry of the brushes from the EN 50367 standard, then an extrusion was made along the curve in order to give volume to the curve. Modifiers were applied, such as the mirror to make the curve symmetrical and the matrix modifier to create a copy of the brush, thus having two brushes. The modeling was done with a focus on correspondence with real proportions, so that it could later be passed to the IFC format, used in BIM.

The following image (figure 7) shows the 3D representation of the brushes used in the pantograph. It is possible to observe their curvature and the way in which they are fixed to the main support structure, thus allowing continuous contact with the catenary during the train's movement.



Figure 7: Basic structure of the Pantograph.

4.1.3 Insulators

Different types of insulators were created, with different dimensions and shapes, using reference photographs. The techniques used in modeling these components were 2D profile extrusion. Some parts include additional details, such as metal fittings or fixing bases, which were modeled with the help of modifiers, such as boolean and array. This diversity of insulators was important to represent, to indicate the different isolation points of the complete system.

The following image (figure 8) shows the variety of modeled insulators, with different colors, sizes and functions. These components are strategically distributed throughout the pantograph structure and wire connections, thus ensuring insulation at all critical electrical contact points.

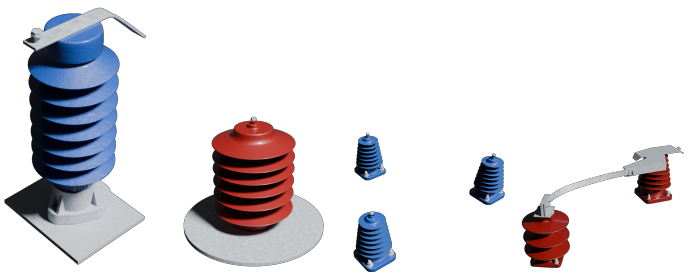


Figure 8: Various insulators present in the Pantograph.

4.1.4 Pantograph

The pantograph was modeled in order to integrate the various components previously developed, such as the base structure, brushes and insulators. The construction of the object was done using basic profile extrusion operations. In this object, only the mirror modifier was needed to create symmetry between the object's parts. The articulated arms were modeled with attention to the rotation points and tensioners, representing the lifting mechanism.

The following figure (figure 9) shows the complete 3D model of the pantograph, with all its associated elements. In this model, it is possible to observe the connections between the components and the articulated system that allows vertical movement for contact with the catenary.



Figure 9: 3D Pantograph.

4.1.5 Disconnector

The modeling of the disconnector, like all previous ones, was based on reference images. The construction of the model was done in stages, starting with the insulating base, followed by the main body and, finally, the mobile and fixed contacts. The main modifier in the construction of this model was boolean, to be able to unite all the constituent elements of the object.

Figure 10 represents the 3D model of the disconnector, where its main components stand out: insulating base, contacts and central body. The 3D representation of this element illustrates its basic functioning and its context in the railway infrastructure.

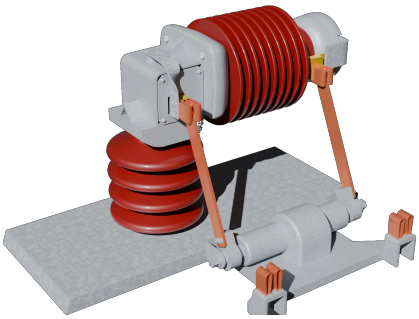


Figure 10: Disconnector.

4.1.6 Real picture VS 3D image

With the modeling of the various components of the pantograph, the comparison between a real image and the 3D model aims to evaluate the degree of precision of the developed model in relation to reality. This analysis allows to verify whether the structural and functional elements of the pantograph were correctly represented, as well as the visual quality and level of detail achieved in the modeling. Thus, it is possible to conclude that the modeled objects are very identical to the real elements of a pantograph.

Figure 11 shows a real photograph of a pantograph, followed by the developed 3D model. Looking at the two images, it is possible to observe the similarity between the main elements and confirm the success of the 3D recreation of the pantograph.

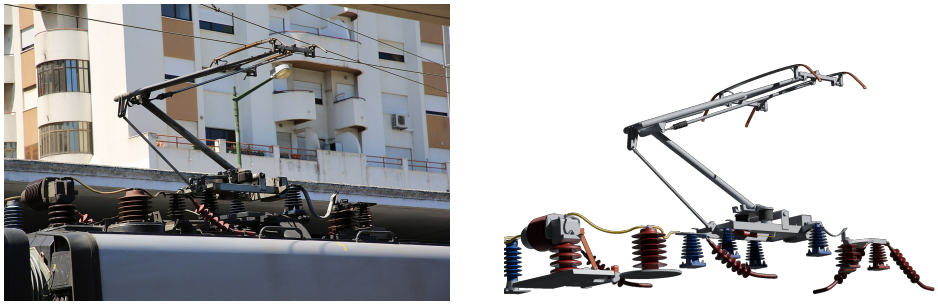


Figure 11: Photograph versus rendered image.

4.2 Blender to IFC

The export of previously developed 3D models to IFC format was carried out with the support of the BlenderBIM add-on, which, as already discussed, allows converting 3D objects into IFC objects that can be recognized in the BIM environment. Figure 12 illustrates the complete process, from creating the IFC project to exporting it. The figure shows how the workflow is done, from the beginning with the blend file until the IFC file is obtained. As a demonstration example, only one cube was used. The process begins with the creation of an IFC project, with the Blender working project already open with the models we want to transfer to IFC, where we define the IFC version and the different units. After creating the IFC project, we can consult several tabs in the scene, one of which is quite important: "Object Information". This is where the Object Metadata (information that will be associated with the model in IFC) is classified. In the same tab, there is also "Property Sets", which is used to manually insert, under "Custom Pset", the different properties of the object in question. Moving on to the "Geometry and Materials" tab, that is where materials are assigned to objects.

To model other objects, it is a repetitive process for all models. This process thus ensures that the developed models can be used in compatible BIM platforms, preserving the geometry and Metadata necessary for integration into maintenance workflows.

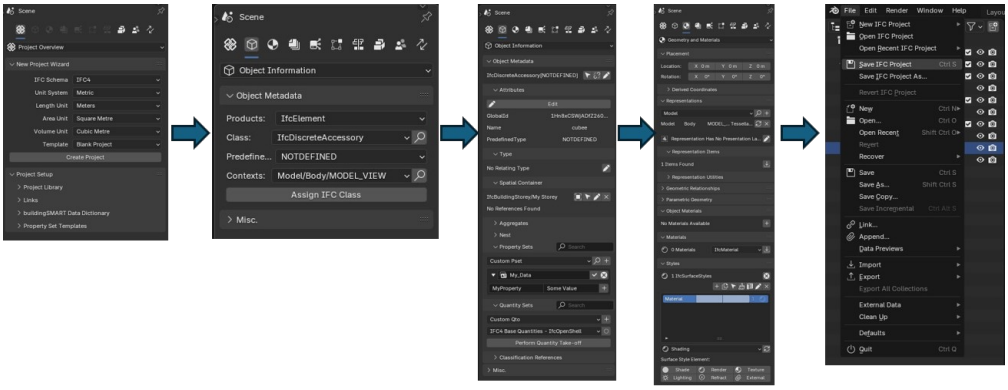


Figure 12: Process to transform blender objects into IFC..

4.3 Augmented Reality Compatibility

Although the final integration with AR applications has not been implemented, the models developed in Blender are technically prepared for this. The geometry can be exported in blend format which is compatible with graphics engines like Unity, often used in AR applications. Unity allows the following items to be imported [7]:

- Meshes;
- Bones;
- Animations.

All the items described above are the most important for using models for AR, with Mesh containing all the necessary geometry, Bones helping to create realistic animations and Animations to use animations made in Blender to simulate some maintenance operation on a pantograph component.

5 Conclusion

The increasing complexity of railway systems and the resources required to train technicians motivated this work, which sought to explore solutions based on 3D modeling, BIM and AR to support technical training in railway maintenance.

The methodology used in this paper involved the comparative analysis of 3D modeling tools, the modeling of pantograph components based on real images and European standards (EN 50119 and EN 50367), and the export of the models in IFC format for later integration into BIM flows.

The results presented include the creation of 3D models, compatible with BIM systems and ready for use in AR application deployments, using Unity. The results support the state of the art by applying open-source technologies to a digital domain that is still little explored, the railway, thus promoting an improvement in the digitalization, training and maintenance of systems.

The main advantages of using this proposal include cost reduction in technician training, as well as flexibility in using the models developed and their compliance with European standards. As a limitation, the implementation and the absence of practical tests in real environments using a system developed for AR stand out.

Future work includes the implementation of an AR application with support for smartphones, tablets and/or AR glasses, where the developed models will be implemented, as well as tests carried out with specialized technicians/trainers to carry out a pedagogical and functional validation of the entire developed system.

References

[1] Siembra Automacao. Realidade aumentada na automação industrial, 2021. Available at: https://www.siembra.com.br/noticias/realidade-aumentada-na-automacao-industrial/ [Accessed 13 February 2025].

[2] Riccardo Masoni, Francesco Ferrise, Monica Bordegoni, Michele Gattullo, Antonio E. Uva, Michele Fiorentino, Ernesto Carrabba, and Michele Di Donato. Supporting remote maintenance in industry 4.0 through augmented reality. *Procedia Manufacturing*, 11:1296–1302, 2017. 27th International Conference on Flexible Automation and Intelligent Manufacturing, FAIM2017, 27-30 June 2017, Modena, Italy.

[3] Mao Lin. Augmented reality solution for facility maintenance with bim. *MsC thesis in Human Computer Interaction and Design , School of Science - Aalto University*, 2019.

[4] Fabrizio Lamberti, Federico Manuri, Andrea Sanna, Gianluca Paravati, Pietro Pezzolla, and Paolo Montuschi. Challenges, opportunities, and future trends of emerging techniques for augmented reality-based maintenance. *IEEE Transactions on Emerging Topics in Computing*, 2(4):411–421, 2014.

[5] Ana Malta, Mateus Mendes, and Torres Farinha. Augmented reality maintenance assistant using yolov5. *Applied Sciences*, 11(11):4758, 2021.

[6] fSpy. The basics, 2025. Available at: https://fspy.io/assets/vp2.jpg [Accessed 09 May 2025].

[7] Unity. Importing objects from blender, 2016. Available at: https://docs.unity3d.com/530/Documentation/Manual/HOWTO-ImportObjectBlender.html [Accessed 23 May 2025].

Chapter 16

Optimizing the Requisition System for the Purchasing Process in the Industrial Sector: A Case Study on Digital Transformation and Continuous Improvement

Contribution #37
Cite as:
Maria Inês Pinho, Ana C V Vieira. Optimizing the Requisition System for the Purchasing Process in the Industrial Sector: A Case Study on Digital Transformation and Continuous Improvement. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-37,
  title ={Optimizing the Requisition System for the
     Purchasing Process in the Industrial Sector: A
     Case Study on Digital Transformation and
     Continuous Improvement},
  author={Maria Inês Pinho and  Ana C V Vieira},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Optimizing the Requisition System for the Purchasing Process in the Industrial Sector: A Case Study on Digital Transformation and Continuous Improvement

Maria Inês Pinho[1*], Ana V Vieira[1,2,3,4]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[2] RCM2+ Research Centre for Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[3] CISE - Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal

[4] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal

*Corresponding author

pinhomariaines@gmail.com, avieira@isec.pt

Abstract

In an increasingly competitive industrial environment, digital transformation and operational efficiency are strategic priorities. The principles of continuous improvement highlight the importance of eliminating waste, improving information flow, and ensuring that all activities add value to the organization.

This article addresses the challenges in the indirect purchasing requisition process in a company within the metalworking sector, where manual procedures result in information loss, non-value-added activities, and delays in the process. The initial phase involved direct observation of the existing system, mapping the flow from requisition to receipt of materials, and identifying critical points. Based on this analysis, a requisition module was developed and integrated into the company's existing software, tailored to its specific needs. This digital solution helped reduce errors and increase the reliability of information.

The results presented in the full version of this article show that the project contributed to waste reduction and improved process efficiency. Thus, reinforcing the role of information systems in supporting continuous improvement and the sustainable modernization of industrial operations.

Keywords: Digital Transformation; Indirect Purchasing; Continuous Improvement; Operational Efficiency; Procurement Management.

1 Introduction

With the continuous growth of organisations and the increasing complexity of their operations, the need for well-structured, digital and traceable processes is becoming increasingly evident. One of the departments most impacted by this transformation is the supply chain, especially the purchasing sector, whose role is essential to maintaining business operations. However, many companies still face significant challenges related to informality, inefficient communication and a lack of standardised procedures.

This study looks at the reality of an organisation that, due to its growth and the consequent increase in the volume of requisitions, was faced with operational problems in the purchasing process. Initially, any employee

could contact suppliers directly, which resulted in disorganisation, loss of information and a lack of visibility over the status of requisitions. The initial attempt at organisation, using an Excel file shared via Microsoft teams, proved to be insufficient, given the lack of access control, flaws in the data entered and difficulty in following up orders.

In order to respond to these problems, a digital system based on PHC CS WEB was implemented, allowing for the formalisation of processes, efficient control of requisitions and detailed monitoring of all purchasing phases. This transformation was carried out using a participatory methodology. Thus, roles were taken on in the purchasing department with the aim of acting as a link between the end users within the company and the external organisation responsible for developing the digital solution, which meant gaining an in-depth understanding of the process from requisition to receipt of the material.

The importance of this research lies in the way it addresses a real case of digitising purchasing processes, demonstrating how technology can be adapted to the specific needs of each organisation.

The methodology adopted includes observation, document analysis, meetings with internal and external stakeholders, and monitoring the customisation of the PHC system. This approach made it possible to obtain qualitative data directly from the work process and to carry out iterative interventions to continuously improve the system.

This article is structured as follows:

- Section 2: Presents a review of the state of the art on digital requisition processes and ERP systems.
- Section 3 Describes the methodology used to conduct the study.
- Section 4: Outlines the results obtained with the implementation of the new system.
- Section 5: Discusses the impacts, limitations and implications of the results.
- Section 6: Presents the conclusions of the study and proposes directions for future work.

2 State of the art

Digital transformation is taking centre stage in the modernisation of business processes, particularly in critical areas such as purchasing management and order control. As organisations grow, it becomes clear that informal methods, such as the use of shared Excel files, reveal serious limitations in terms of internal control, traceability and operational efficiency. In this context, the literature has shown that the digitalisation and automation of processes are not just technological trends, but strategic imperatives to guarantee competitiveness and long-term sustainability [1], [2].

Firstly, it establishes a direct relationship between maturity in Business Process Management (BPM) and the benefits achieved with Robotic Process Automation (RPA). Organisations that already have mapped and standardised processes are able to extract greater gains from digitalisation, namely in terms of time reduction, mitigation of manual errors and improved control over operational activities.[3].

In addition, it introduces the concept of the digital twin as a tool to support decision making in real time [4]. Although applied to the selection of machines, the principle of digitally modelling a process before carrying it out allows for more informed optimisation in line with real needs. This concept is transversal to any operational area, including the purchasing department.

The relevance of digitalisation also extends to the supply chain, as shown by Shuleski and Mãitã [5]. These authors emphasise that adopting digital approaches helps to increase the resilience and flexibility of the logistics system. Traceability and data centralisation are key factors in ensuring operational continuity, even in contexts of instability or high growth.

In addition, it shows that by modelling processes with intelligent technologies, it is possible to achieve not only operational gains, but also economic ones, eliminating redundant tasks and improving

the allocation of resources [6] This is especially visible in activities such as manual status updates or dispersed document management, as was the case in the case study analysed before the intervention.

The automation of administrative processes using RPA is explored in depth in works that identify clear improvements in the ordering process, with emphasis on reducing execution time, eliminating repetitive tasks and strengthening the credibility of information.[5] [7].

From a human and organisational point of view, it is argued that the clarity of digital processes increases employee motivation by offering intuitive, simple tools adapted to their daily reality[8]. This user-centred approach is also present in the solution developed in this case study, where different profiles were considered, applicants, approvers and buyers, with specific functionalities developed for each of the profiles.

Finally, it should be noted that digital integration between areas of the company promotes a global and cohesive vision of processes, facilitating communication, transparency and efficiency [9]. The centralisation of information, as implemented in this project, reduces the time it takes to search for data and eliminates dependencies on paper records or manual files.

In summary, the literature review shows that digitalisation, automation and standardisation of processes not only increase efficiency and reduce errors, but also promote greater transparency, traceability and internal motivation. The studies analysed converge on the idea that a well-implemented digital system must be built based on the real needs of the organisation and its employees [1-10].

3 Methodology

The methodology adopted in this article was qualitative and participatory in nature, based on direct observation and practical involvement in the routines of the purchasing department, with the aim of understanding the existing operational difficulties and proposing a technological solution in line with the organisation's reality.

Initially, the role of buyer was temporarily taken on. This approach enabled an in-depth understanding of the purchasing process and made it easier to identify various operational flaws.

The workflow identified in the purchasing process, in its initial state, developed as follows:

- The requester filled in a requisition in a shared Excel file.
- The purchasing department consulted the market and requested quotations.
- The purchase order was created in the PHC system.
- The order was placed for approval, with multiple requisitions depending on a single approver.
- The status of the order was updated manually (e.g. ‘under approval’).
- After approval, the order was sent to the supplier;
- The order status was updated manually again (e.g. ‘order dispatched’).

The tool used, an Excel file shared on Microsoft Teams, proved to be insufficient as the organisation grew. The lack of access control meant that any employee could access and change requisition data, leading to loss of information, registration errors, inconsistencies and difficulty in formalising orders for the purchasing department.

From the purchasing department's point of view, in addition to the length of the process and the vulnerability of the shared file, the use of Excel made it significantly more difficult to monitor open orders. It was necessary to use manual paper records and constantly update files, which proved unsustainable in a growing organisation. This reality was confirmed by actually timing operations in two different areas (design and maintenance). It was found that passing data from Excel to the order form in PHC, as well as analysing the associated budget, took around 10 to 20 minutes per request, depending on complexity. Updating the status of orders (such as ‘request for quotation’, ‘under approval’, “sent”,

‘received’) took an average of 12 minutes and was repeated between two and three times a day. In addition to the time consumed, these tasks were considered operational waste, as they added no value to the process and limited the department's ability to act strategically.

In addition, the mapping revealed the existence of a large number of orders awaiting approval, all under the responsibility of a single approver. This centralisation generated significant delays (between one and three days), compromising the fluidity of the process and the planning of the other departments.

Given these limitations and the urgent need to reformulate the process, meetings were held with the main internal players, namely the purchasing, accounting and IT departments, as well as with the entity responsible for customising the PHC CS Web platform. These sessions made it possible to present the problems identified, clarify the intended objectives and define the requirements of the technological solution to be implemented.

It was agreed to segment the requirements of the new solution according to the different user profiles: requesters, approvers and buyers.

For applicants, the requirements defined included:

- Make the cost centre and analytical item fields compulsory, ensuring correct categorisation of requisitions.
- Allow the statistical order to be recorded, where applicable, according to the organisation's internal parameters.
- Guarantee individual access to the history of their own requisitions, with the possibility of monitoring the status (in approval, approved, converted, sent, completed) and consulting the associated documents.
- Providing a simple, clear interface with adapted nomenclature in order to reduce filling errors and make it easier for all employees to use.

For approvers, it was essential to introduce an approval system ordered by value, given the growth in the volume of requisitions and the need for greater control. Thus, requisitions up to a certain monetary value would be approved by a level 3 manager, while higher values would go to higher hierarchical levels (level 2 and level 1). This measure made it possible to avoid duplicate purchases and promote greater visibility of each department's real needs.

In the case of the purchasing department (buyers), the requirements were geared towards strengthening control and the ability to analyse the process. Among the most important aspects were:

- The introduction of filters by area, supplier and order status, to facilitate the segmentation and analysis of requisitions made via PHC CS WEB.
- The implementation of dashboards and specific analyses to identify orders that have been approved but not yet sent to the supplier.
- The effective monitoring of open orders, to prevent forgetfulness and delays.
- The digital centralisation of all relevant documentation, from the requisition to the purchase order and invoice, accessible with a single click. This measure proved to be fundamental, since previously the documents were scattered across different platforms, making it difficult to consult, follow the process and trace the documents. With the new solution, all documents are automatically associated with the respective requisition, eliminating loss of information and facilitating complete monitoring of the purchasing cycle.
- The possibility of exporting data from specific analyses to Excel, an essential feature for drawing up reports and external monitoring, which was previously unavailable.

During this digitalization project, work instructions were developed to guide users in adopting the new platform, promoting the standardization of procedures. Training sessions were also held for the first users and further training sessions are planned as the system is rolled out to other areas of the company that do not yet use the new requisition system.

It should be emphasized that the implementation phase of the PHC CS WEB system required several adjustments to the basic solution, with additional meetings being held with the external company responsible for developing the digital solution. Continuous and direct involvement with the process was essential to ensure that the final solution was truly aligned with the specific needs and operational reality of the organization.

4 Results

The implementation of the new solution in PHC CS Web enabled measurable operational gains and a significant improvement in the control and efficiency of the purchasing process. Tabel 1 shows the main results obtained, based on the comparison between the initial scenario and the implemented system.

Tabel 1: Time evaluation results

Task	Initial System (Excel)	Integrated system (PHC CS WEB)	Estimated Gain
Transfer of data from the requisition to the purchase order	10-20 min	≈30 sec	95% reduction
Manual update of order status	12 min per update 2-3x/day	Automatic	Deletion of task
Document history control and order tracking	Manual and dispersed (paper, folders)	Centralised and accessible at 1 click	Reduction of consultation time from several minutes to seconds

By analysing the results, we can see a significant reduction in the time spent on manual and repetitive tasks. In the previous system, simply transferring data from the requisition filled out in Excel to the order form in PHC could take between 10 and 20 minutes. With the new solution, this process is now carried out directly in the system, requiring only around 30 seconds. This change represented a time saving of over 95 per cent.

In addition, the manual updating of order status, which previously required approximately 12 minutes per operation and was carried out two or three times a day, has been fully automated. This elimination of redundant tasks has freed up resources for activities with greater added value, such as analysing suppliers or negotiating commercial conditions.

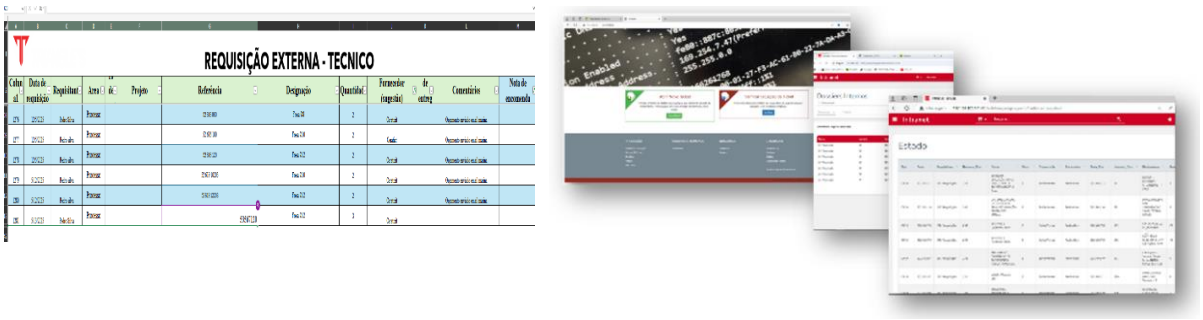
From the purchasing department's point of view, the improvements were not limited to time efficiency. The digital centralisation of all the documentation associated with the process (requisition, quotation, purchase order and invoice), accessible with a single click, solved one of the main problems previously identified: the dispersion of information across different media (email, paper, shared folders), which made it difficult to monitor and audit processes. This reorganisation of information made a decisive contribution to improving traceability and transparency.

On the other hand, it was possible to strengthen internal control, and the quality of the data entered. Making it compulsory to fill in essential fields, such as the cost centre and the analytical item, drastically reduced errors in categorizing requisitions. It is also important to mention the gains made in terms of the

autonomy of those involved. Requesters now have access to the history of their own requisitions, as well as up-to-date information on the status of each request (in approval, approved, converted, sent, delivered), which has reduced the need to contact the purchasing department and improved communication between areas.

As far as approval management is concerned, the creation of a hierarchical circuit by value, with approval by different levels depending on the amount of the requisition, has made it possible to better control spending by the department and avoid duplicate orders. This new approval model has also helped speed up response times and facilitate internal planning.

Finally, Figure 1 shows a general overview of the old and new requisition systems.



a) **b)**

Figure 1: Si Requisitions System: a) previous, through shared Excel; b) developed collaboratively and through PHC CS WEB

5 Discussion

The starting point for the transformation of the purchasing process presented in this study was the realization, made during the practical immersion phase, that the previous system based on a shared Excel file had critical operational weaknesses: lack of control over access, high probability of errors, lack of traceability and time wasted on repetitive tasks with no added value. On the basis of this diagnosis, and in the light of good practices in literature, a digital solution was designed based on the customization of PHC CS Web.

The results obtained show significant operational gains, such as a 95 per cent reduction in the time needed to transfer data from the requisition to the order form and the total elimination of the task of manually updating statuses. These advances validate what is advocated by Stravinskiene and Serafinas when they state that the standardization of processes is an essential condition for effective automation [3]. The case under study confirms that, in the absence of formalized processes, the benefits of digitalisation are compromised, which is exactly what we sought to resolve with the segmentation of profiles and approval rules in the new system

In addition, the centralization of documents and real-time visibility of active requests has eliminated the dispersion of information that previously existed in paper, physical folders or informal sharing. This type of consolidation is in line with the benefits pointed out by Najat and Eddine, who emphasize the importance of digitally integrating the value chain to improve communication and responsiveness between departments [9].

From the point of view of efficiency, the gains seen in this study, both in terms of time and error reduction, reflect the effects described by Shuleski and Mãitã, whose work showed that adopting RPA can drastically reduce the administrative burden, allowing employees to focus on activities with greater added value [5]. The elimination of redundant manual activities, such as updating order status or physically checking the history, reinforces this trend.

User motivation has also visibly improved with the adoption of the new solution, as the clear interface adapted to their needs has reduced frustration and errors when filling in forms. This human component, which is often underestimated in digital transformation projects, is highlighted by Behrendt et al. who emphasize the

importance of clarity in digital flows for employee engagement and learning. [8]. By offering a more intuitive and structured experience, the new system contributed to greater acceptance of the change.

Another critical aspect was the ability to monitor and trace orders. Previously non-existent in automated form, this functionality is now integrated into the PHC CS WEB system. It enables rigorous, real-time control, something advocated as essential by Shamsuzzoha and Pelkonen, who showed that automating ordering processes in the supply chain reduces operational failures and speeds up decision-making [7].

The connection between this change and the concepts of resilience and organizational adaptation also deserves to be highlighted. As highlighted by Mauss et al., digitalization makes it easier to respond to unforeseen events, improves processes and makes companies more prepared to deal with uncertain contexts [10]. The case presented here demonstrates that even an apparently simple process, such as internal purchase requisitions, when well-structured and digitalized, can bring measurable gains and generate a positive impact on the organization.

In summary, the results obtained in this case study validate the premises presented in the introduction and confirm the trends identified in the literature. Digitizing the purchasing process has not only improved operational efficiency and reduced waste, it has also increased transparency and internal control. Experience reinforces the importance of aligning technology with well-defined processes, involving users from the design phase. Furthermore, it demonstrates that even traditionally less prioritized areas such as request management can be relevant sources of strategic gain, internal motivation and preparation for future challenges. This approach represents a solid step towards the organization's digital maturity, with room for continuous improvement.

6 Conclusion

The present study aimed to analyze the impact of implementing a digital requisition system on the purchasing process of a growing organization. Through a participatory approach, it was possible to diagnose critical flaws in an informal system and closely monitor the adaptation of the PHC CS WEB computerized solution to the organization's real needs.

The main results demonstrate: i) greater efficiency and reduced time in processing requests; ii) increasing the quality of information and process control; iii) improving internal communication and transparency; and iv) reduction of operational failures, such as loss of data or unshipped orders.

The importance of this work lies in the way it highlights the active role of users in adapting digital systems, making the transformation process more effective and adjusted to reality.

The improvements achieved are evident, meaning the system continues to be subject to improvements to resolve some limitations that are still identified, and the necessary procedures are being implemented so that it can also be used in other areas of the company. The limitations must be addressed in future development phases, such as the inclusion of additional features, integration with other platforms, and improved data export. For future work, it is recommended that more robust quantitative analyzes be carried out, which will allow us to more accurately measure the economic and operational impact of the implemented digitalization.

Acknowledgments: We appreciate the support provided by the RCM2^{+} (Research Centre for Asset Management and Systems Engineering) and by ISEC (Coimbra Institute of Engineering). Our thanks are extended to the company that provided the opportunity to develop and execute the project, fundamental to the development of the work, as well as all the personnel involved.
This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Projects UIDB/04131/2020 and UIDP/04131/2020.

References

[1] José Dinis Carvalho, *Melhoria Contínua nas Organizações*, 1ª Edição. Biblioteca Lean, 2021.

[2] José Crespo de Carvalho, *Logística e Gestão da Cadeia de Abastecimento*, 3ª Edição. 2020.

[3] I. Stravinskiene and D. Serafinas, "Process Automation in an Organisation," *TRANSFORMATIONS IN BUSINESS & ECONOMICS*, vol. 22, no. 60A, pp. 1060–1073, 2023.

[4] K. Skianis, A. Giannopoulos, A. Kalafatelis, and P. Trakadas, "Digital Twin for Automated Industrial Optimization: Intelligent Machine Selection via Process Modelling," in *Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 84–90. doi: 10.1109/IAICT59002.2023.10205949.

[5] D. SHULESKI and D. N. MÃITÃ, "Automation and Optimization of Management Processes through RPA," *Review of International Comparative Management*, no. Vol. 25 No. 4 / 2024, Oct. 2024, doi: 10.24818/RMCI.2024.4.775.

[6] A. G. Tanase, A. F. Preda, R. F. Cretu, and V. C. Banta, "Modelling of an Economic Process Using Smart Technologies in Manufacturing in the Context of Industry 4.0: A Case Study," *Accounting and Finance*, vol. 4, no. 106, pp. 169–176, 2024, doi: 10.33146/2307-9878-2024-4(106)-169-176.

[7] A. Shamsuzzoha and S. Pelkonen, "A robotic process automation model for order-handling optimization in supply chain management," *Supply Chain Analytics*, vol. 9, Mar. 2025, doi: 10.1016/j.sca.2025.100102.

[8] F. Behrendt, O. Lehner, A. Rettmann, N. Schmidtke, and T. Wollert, "Process analysis of a teaching and learning factory environment to demonstrate Industry 4.0 solutions by using the Smart Logistics Zone approach," in *2022 IEEE 6th International Conference on Logistics Operations Management, GOL 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/GOL53975.2022.9820054.

[9] T. Najat and E. M. Alaa Eddine, "Digitalization and Business Automation for an Effective Supply Chain Integration: A literature review," in *2024 IEEE 15th International Colloquium of Logistics and Supply Chain Management, LOGISTIQUA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/LOGISTIQUA61063.2024.10571518.

[10] N. A. Mauss, M. Ruehs, and J. Fottner, "Potentials of Circular Economy Approaches for Supply Chain Resilience," in *2022 IEEE 6th International Conference on Logistics Operations Management, GOL 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/GOL53975.2022.9820143.

Chapter 17

Pareto-Optimized Model Predictive Control for Real-Time 3D Trajectory Planning of Collaborative Robots

Contribution #40
Cite as:
Zeinel Momynkulov, Azhar Tursynova, Olzhas Olzhayev, Akhanseri Ikramov, Sayat Ibrayev, Amandyk Tuleshov, Batyrkhan Omarov. Pareto-Optimized Model Predictive Control for Real-Time 3D Trajectory Planning of Collaborative Robots. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-40,
  title ={Pareto-Optimized Model Predictive Control for
     Real-Time 3D Trajectory Planning of Collaborative
     Robots},
  author={Zeinel Momynkulov and  Azhar Tursynova and
     Olzhas Olzhayev and  Akhanseri Ikramov and  Sayat
     Ibrayev and  Amandyk Tuleshov and  Batyrkhan
     Omarov},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Pareto-Optimized Model Predictive Control for Real-Time 3D Trajectory Planning of Collaborative Robots

Zeinel Momynkulov[1,2], Azhar Tursynova[1,2], Olzhas Olzhayev[1,2], Akhanseri Ikramov[1], Sayat Ibrayev[1], Amandyk Tuleshov[1], Batyrkhan Omarov[1,2,*]

[1] Institute of Mechanics and Engineering Science, Almaty, Kazakhstan
[2] International Information Technology University, Almaty, Kazakhstan

*Corresponding author
batyahan@gmail.com

Abstract

This study proposes a Pareto-optimized Model Predictive Control (MPC) framework for real-time 3D trajectory planning of collaborative robots operating in human-shared environments. Unlike classical interpolation techniques such as cubic spline, B-spline, and piecewise linear methods—which emphasize geometric smoothness but neglect actuator constraints—the proposed approach integrates dynamic feasibility directly into the planning process. The system is modeled using a discrete second-order kinematic representation, and trajectory generation is formulated as a constrained optimization problem minimizing both tracking error and control effort. To address the trade-off between positional accuracy and control smoothness, a structured grid search over prediction horizon lengths and control weights is performed, followed by Pareto-front analysis to identify non-dominated configurations. Experimental comparisons demonstrate that, while MPC incurs a slight increase in positional error relative to spline-based methods, it significantly reduces peak acceleration and velocity fluctuations—key factors for ensuring safety and compliance in collaborative tasks. The results confirm that Pareto-optimized MPC offers an effective solution for trajectory generation in real-time applications, balancing precision and physical feasibility in dynamic environments where human-robot interaction demands both responsiveness and reliability.

Keywords: Model Predictive Control; 3D trajectory planning; spline interpolation; Pareto optimization; waypoint navigation; real-time control; motion planning; robotic systems

1 Introduction

Collaborative robots (cobots) have become integral to modern industrial and service applications, offering the flexibility to operate in shared workspaces with humans. A fundamental requirement for safe and efficient cobot operation is the generation of smooth, real-time trajectories in three-dimensional (3D) space. These trajectories must pass through predefined waypoints while satisfying strict physical constraints such as limits on velocity, acceleration, and control input [1]. Traditional geometric interpolation methods—including cubic splines, B-splines, and linear interpolation—are widely used due to their simplicity and smoothness properties [2]. However, these methods do not account for dynamic feasibility, often producing trajectories with discontinuous or excessive accelerations that exceed actuator limits or compromise safety in collaborative environments [3].

To overcome these limitations, Model Predictive Control (MPC) has been adopted for trajectory planning in dynamic systems. MPC optimizes control inputs over a finite prediction horizon, explicitly incorporating kinematic constraints and system dynamics to generate feasible trajectories [4]. By minimizing a cost function that includes tracking error and control effort, MPC ensures smoother motion profiles suited for real-time applications [5]. However, the performance of MPC depends heavily on the choice of parameters such as prediction horizon length and control penalization weight. Poorly chosen parameters can lead to high computational cost or poor tracking performance, especially in time-sensitive robotic systems [6].

In this study, we introduce a Pareto-optimized MPC framework for real-time 3D trajectory planning of collaborative robots. The method employs a grid search over candidate MPC configurations and applies Pareto-front analysis to identify optimal trade-offs between positional accuracy and control smoothness [7]. Comparative experiments against spline-based methods demonstrate that the proposed approach produces dynamically feasible trajectories with significantly reduced acceleration peaks, making it suitable for safe and effective cobot deployment in human-shared environments.

2 Related Works

Trajectory planning has traditionally relied on geometric interpolation techniques to generate continuous paths through predefined waypoints. Methods such as cubic spline interpolation offer second-order continuity, enabling smooth transitions in position and velocity across trajectory segments [8]. B-spline interpolation extends this approach by using basis functions of adjustable degree, offering greater control over smoothness but often at the cost of deviating from exact waypoint passage [9]. Piecewise linear interpolation, while computationally inexpensive, introduces discontinuities in velocity and acceleration at segment junctions, posing risks for actuators and system stability [10]. Although these methods are effective for offline planning, they lack the capacity to enforce dynamic feasibility, which is critical in collaborative robotics where motion must comply with strict physical constraints to ensure safe human-robot interaction [11].

Model Predictive Control (MPC) addresses these limitations by framing trajectory generation as a finite-horizon optimization problem that incorporates system dynamics and constraint enforcement [12]. MPC has shown promise in a variety of domains, including automotive systems where it ensures passenger comfort and collision avoidance [13], and robotic manipulators where it helps maintain joint limits and actuator bounds during motion execution [14]. Despite its advantages, MPC's practical deployment is often challenged by its computational demands, especially for real-time applications involving long horizons or complex constraints [15]. This has led to significant research on efficient solvers, horizon reduction strategies, and adaptive parameter tuning to make MPC viable for high-speed robotic systems.

In response to the need for balancing multiple performance criteria, researchers have integrated multi-objective optimization strategies into MPC frameworks. Genetic algorithms have been employed to tune control weights for objectives such as energy efficiency and accuracy [16], while Pareto front analysis has been used to identify non-dominated solutions that balance trade-offs between control effort, precision, and responsiveness [17]. In the context of UAV path planning, such strategies help optimize metrics like flight time, safety margins, and energy use simultaneously [18]. Grid-based parameter exploration coupled with Pareto filtering has also been applied to process control [19], humanoid gait generation [20], and adaptive planning systems [21]. The present study builds upon these works by applying Pareto-optimized MPC to real-time 3D trajectory planning for collaborative robots, offering a structured method to manage the competing demands of accuracy and feasibility in shared human-robot environments.

3 Problem Statement

Trajectory planning in three-dimensional environments requires balancing precise waypoint tracking with adherence to dynamic and actuator limitations. The following formulation defines the optimization objectives and constraints underpinning our Pareto-optimized MPC approach. Consider a set of ordered 3D waypoints

$$\left\{p_{ref,0}, p_{ref,1}, \ldots, p_{ref,T}\right\} \subset R^3 \quad (1)$$

to be visited sequentially over a discrete time horizon T with sampling interval Δt. Define the system state at step k as

$$x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix} \in R^6 \tag{2}$$

Where $p_k \in R^3$ is position and $v_k \in R^3$ is velocity. The control input is the acceleration $a_k \in R^3$. The discrete-time dynamics satisfy

$$\begin{aligned} p_{k+1} &= p_k + \Delta t v_k + \frac{1}{2}\Delta t^2 a_k, \\ v_{k+1} &= v_k + \Delta t a_k \\ k &= 0,...,T-1 \end{aligned} \tag{3}$$

We seek the acceleration sequence $\{a_k\}_{k=0}^{T-1}$ that minimizes both the mean tracking error and the control effort:

$$\begin{aligned} &\min_{\{a_k\}} \{J_1, J_2\}, \\ &J_1 = \frac{1}{T}\sum_{k=1}^{T} \left\| p_k - p_{ref,k} \right\|^2, \\ &J_2 = \frac{1}{T}\sum_{k=0}^{T-1} \left\| a_k \right\|^2 \end{aligned} \tag{4}$$

Subject to actuator and safety constraints

$$v_{\min} \le v_k \le v_{\max}, \quad a_{\min} \le a_k \le a_{\max} \tag{5}$$

Because reducing positional error J_1 often increases acceleration peaks J_2 we employ a Pareto-optimization strategy over MPC parameters—prediction horizon N and weight λ – to identify non-dominated configurations that optimally trade off between these objectives.

4 Materials and Methods

Kinematic Model of Motion. In this study, the motion of a point-mass system in three-dimensional (3D) space is described by a second-order discrete-time kinematic model. This formulation is suitable for trajectory planning tasks where smooth transitions between waypoints and compliance with physical constraints—such as limits on acceleration and velocity—are essential.

At each discrete time step k, the system state is defined as:

$$x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix} \in R^6 \tag{6}$$

Where $p_k \in R^3$ denotes the 3D position vector, and $v_k \in R^3$ is the corresponding velocity vector. The control input is represented by the acceleration vector $a_k \in R^3$. Assuming a fixed sampling interval Δt, the discrete-time kinematic equations governing the system's evolution are given by:

$$\begin{aligned} p_{k+1} &= p_k + \Delta t v_k + \frac{1}{2}\Delta t^2 a_k, \\ v_{k+1} &= v_k + \Delta t a_k \end{aligned} \tag{7}$$

Equation (7) captures the position update based on the current velocity and acceleration, while equation (5) updates the velocity using the applied acceleration. These equations assume constant acceleration over each time interval Δt, which is consistent with the assumptions used in standard discrete-time motion models.

To ensure feasibility and safety during trajectory execution, the system must comply with predefined physical constraints on both velocity and acceleration. These are defined as component-wise bounds:

$$v_{\min} \le v_k \le v_{\max}, \quad a_{\min} \le a_k \le a_{\max}, \quad \forall k \in \{0,1,...,T-1\} \tag{8}$$

where the inequalities are understood to hold element-wise.

This kinematic model serves as the predictive dynamic framework within the Model Predictive Control (MPC) formulation. It allows for accurate forward simulation of system behavior and enables constraint-aware optimization for real-time trajectory planning in 3D environments.

Comparative Trajectory Models. To benchmark the MPC method, three classical interpolation techniques were used for trajectory generation through predefined 3D waypoints: cubic spline interpolation, B-spline interpolation, and piecewise linear interpolation.

Cubic Spline Interpolation constructs piecewise third-degree polynomials for each segment between successive waypoints $\{p_i, p_{i+1}\}$, ensuring C^2-continuity of the resulting trajectory:

$$S_i(t) = a_i t^3 + b_i t^2 + c_i t + d_i, \quad u \in [0,1] \tag{9}$$

which results in a smooth, globally continuous path without necessarily passing through all waypoints.

Piecewise Linear Interpolation generates a segmented linear trajectory between waypoints:

$$L_i(\tau) = (1-\tau)p_i + \tau p_{i+1}, \quad \tau \in [0,1] \tag{10}$$

where the inequalities are understood to hold element-wise.

which is simple to compute but introduces discontinuities in velocity and acceleration.

The MPC-generated trajectories are compared against these baseline methods using a common set of performance metrics.

Search for Optimal MPC Parameters. The proposed MPC method depends on two critical hyperparameters:

- Prediction Horizon Length $N \in \{40,60,80,100\}$
- Sampling Interval $\Delta t \in \{0.05,0.1,0.2\}$ seconds

Where Each candidate parameter pair $(N, \Delta t)$ defines a specific MPC configuration. The controller minimizes the objective function:

$$J = \sum_{i=0}^{N-1} \left\| p_{k+i+1} - p_{ref,i} \right\|^2 + \lambda \sum_{i=0}^{N-1} \left\| a_{k+i}{}^2 \right\| \tag{11}$$

Where λ is weighting factor controlling the trade-off between tracking accuracy and smoothness of control.

For evaluation, the following performance indicators were computed over a full trajectory of length T:

Pointwise Error:

$$e_i = \left\| p_i^{MPC} - p_i^{ref} \right\|, \quad i = 1,...,T \tag{12}$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{T} \sum_{i=1}^{T} e_i \tag{13}$$

Maximum Error:

$$e_{\max} = \max_{1 \le i \le T} e_i \tag{14}$$

Velocity and Acceleration Norms:

$$\|v_i\|_2 = \sqrt{v_i^T v_i}, \quad \|a_i\|_2 = \sqrt{a_i^T a_i} \tag{15}$$

Pareto-Based Parameter Selection. To identify optimal configurations, each candidate MPC parameter set was evaluated and plotted in the objective space defined by (MAE,mean acceleration norm). A solution is considered Pareto-optimal if no other solution achieves both a lower MAE and a lower acceleration norm

simultaneously. The Pareto front is defined as:

$$\mathrm{P}=\left\{(N,\Delta t,\lambda)\,|\,\neg\exists(N',\Delta t',\lambda')\ such\ that\ J_1^{'}<J_1\ and\ J_2^{'}<J_2\right\} \tag{16}$$

Where J_1 and J_2 are the MAE and average acceleration norm, respectively.

Pareto-optimal configurations were then selected for further analysis, enabling systematic exploration of the trade-off between tracking precision and smoothness in real-time trajectory planning.

MPC Algorithm. The proposed MPC algorithm generates dynamically feasible trajectories by solving a finite-horizon optimal control problem at each time step, while enforcing system dynamics and input constraints. The algorithm iteratively computes acceleration inputs that optimize a cost function balancing tracking accuracy and control smoothness.

Algorithm: Model Predictive Control for 3D Trajectory Tracking

Given:

Initial state $x_0=\begin{bmatrix}p_0\\v_0\end{bmatrix}\in R^6$

Prediction horizon N, time step Δt, weighting parameter λ

Reference trajectory: $\{p_{ref,i}\}_{i=1}^{T}$

Constraints: $a_{\min}$, $a_{\max}$, $v_{\min}$, $v_{\max}$

Initialize: $k\leftarrow 0$, $x_k=\begin{bmatrix}p_k\\v_k\end{bmatrix}$

While $k<T$ do:

1. Prediction:

Use the kinematic model to simulate state evolution over horizon N given candidate acceleration sequence $\{a_{k+i}\}_{i=0}^{N-1}$ using:

$$\begin{aligned}p_{k+i+1}&=p_{k+i}+\frac{1}{2}\Delta t v_{k+i}+\frac{1}{2}\Delta t^2 a_{k+i},\\ v_{k+i+1}&=v_{k+i}+\Delta t a_{k+i}\end{aligned} \tag{17}$$

2. Optimization:

Solve the following constrained minimization problem:

$$\min_{\{a_{k+i}\}}\quad J=\sum_{i=0}^{N-1}\left\|p_{k+i+1}-p_{rej,K=I+1}\right\|^2+\lambda\sum_{i=0}^{N-1}\left\|a_{k+i}\right\|^2 \tag{18}$$

Subject to:

$$\begin{aligned}&v_{\min}\le v_{k+i}\le v_{\max},\\ &a_{\min}\le a_{k+i}\le a_{\max},\quad \forall i=0,\ldots,N-1\end{aligned} \tag{19}$$

Apply First Control Input:

Apply only the first control action from the optimal sequence:

$$\begin{aligned}p_{k+1}&=p_k+\Delta t v_k+\frac{1}{2}\Delta t^2 a_k^*,\\ v_{k+1}&=v_k+\Delta t a_k^*\end{aligned} \tag{20}$$

Advance Time Step:

$$k\leftarrow k+1 \tag{21}$$

End While

This receding horizon approach ensures that only the first optimal control input is executed at each iteration, while the optimization is repeated at every subsequent time step using updated system states. As a result, the MPC algorithm adapts to disturbances and modeling inaccuracies while respecting real-time feasibility.

5 Results

This section presents the results of comparative experiments conducted to evaluate the performance of the proposed MPC-based trajectory planning method against classical interpolation techniques. The analysis focuses on trajectory accuracy, velocity smoothness, and acceleration feasibility across different MPC configurations. Quantitative metrics, including mean absolute error, velocity norms, and acceleration magnitudes, are reported alongside visualizations to illustrate the trade-offs between geometric fidelity and control effort.

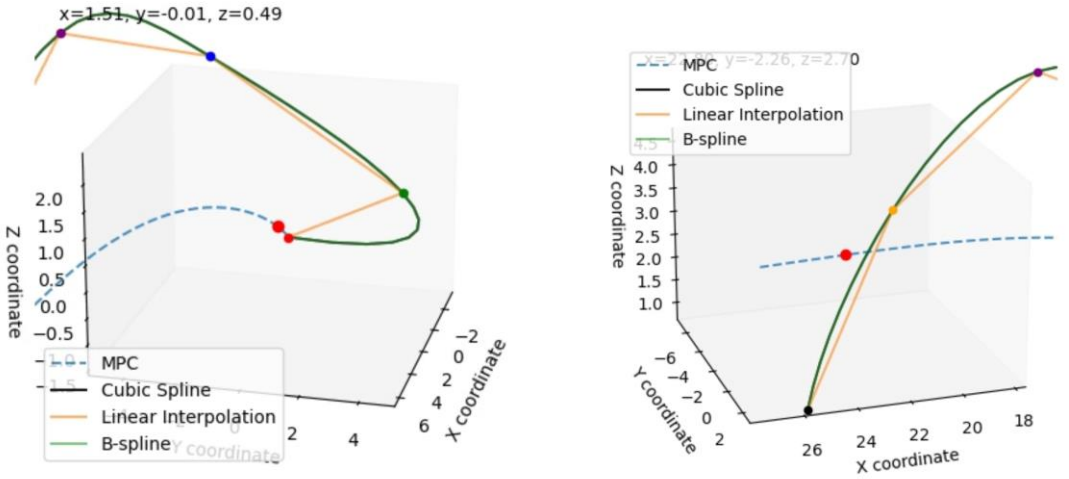


Figure 1: Comparison of MPC and Spline-Based Trajectories in 3D Space for Experiment

Figure 1 illustrates the trajectory comparison from Experiment №0 conducted with MPC parameters set to a sampling interval of $\Delta t=0.05$ seconds and a prediction horizon of $N=40$. The two subplots present distinct trajectory segments in three-dimensional space, highlighting the qualitative behavior of the MPC-generated path relative to traditional interpolation methods—cubic spline, linear interpolation, and B-spline. The MPC trajectory (dashed blue line) distinctly diverges from the purely geometric curves, opting for a dynamically feasible arc that adheres to velocity and acceleration constraints. While the spline-based methods tightly conform to the reference waypoints, they exhibit sharp curvature and non-uniform transitions, particularly visible in the linear segments and high curvature zones of the B-spline. In contrast, the MPC trajectory demonstrates a smoother transition between waypoints, avoiding abrupt directional changes, albeit at the cost of positional deviation. For instance, the trajectory under MPC control takes a more gradual turn in the left subplot and a gentler ascent in the right subplot. These differences underscore the MPC's tendency to favor actuator-friendly motion profiles over geometric precision. Despite the larger deviation in position, particularly around high-curvature regions, the resulting trajectory remains within tolerable error margins, validating the trade-off strategy embedded in the cost function. Overall, Figure 1 provides a compelling visual confirmation that the MPC approach, while not passing exactly through all waypoints, successfully prioritizes dynamic feasibility and smooth control action—key requirements for real-time execution in physical systems.

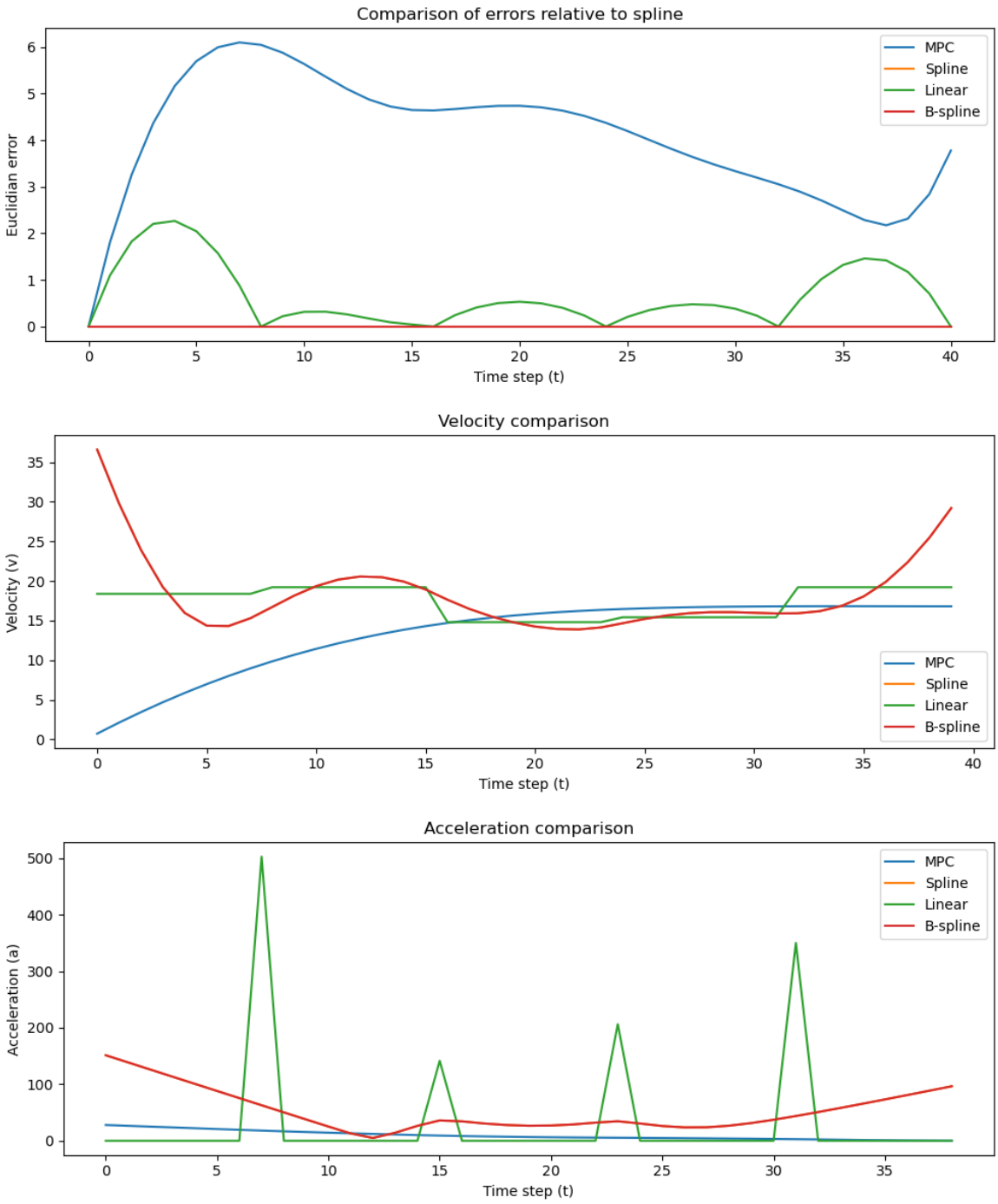


Figure 2: Temporal Comparison of Error, Velocity, and Acceleration Profiles for MPC and Spline-Based Methods in Experiment

Figure 2 presents a comparative analysis of the temporal behavior of different trajectory generation methods—MPC, cubic spline, piecewise linear, and B-spline—over the full horizon of Experiment №0. The top subplot illustrates the Euclidean error at each time step between the generated trajectories and the reference spline trajectory. As shown, the Model Predictive Control

(MPC) approach introduces higher tracking error, especially during the initial and mid-trajectory segments, reflecting its prioritization of dynamic feasibility over strict path adherence. In contrast, the B-spline maintains a zero error baseline, as it is used as the reference trajectory in this experiment, while linear interpolation exhibits periodic fluctuations due to abrupt segment transitions. The middle subplot compares the velocity profiles. MPC yields a gradual and physically consistent velocity increase that stabilizes over time, indicating a controlled motion strategy. In contrast, linear interpolation shows piecewise-constant velocities with abrupt jumps, and the B-spline trajectory exhibits significant variability, including sharp velocity spikes at the boundaries and curvature transitions. This inconsistency may compromise actuator performance in real-world systems. Most strikingly, the bottom subplot reveals the acceleration characteristics: the MPC trajectory maintains a smooth and bounded acceleration profile, indicative of its optimization-based regularization. In comparison, the linear interpolation method produces large acceleration spikes—peaking near 500 units—which are physically unrealistic and would be impractical for most robotic actuators. Even the B-spline, while smoother than the linear method, generates higher and less consistent acceleration magnitudes than MPC. These findings confirm that although MPC sacrifices some positional fidelity, it substantially reduces high-frequency oscillations and extreme control efforts, making it more suitable for deployment in systems with stringent actuation constraints. Therefore, the experiment supports the hypothesis that MPC offers a superior balance between motion realism and computational tractability, especially in scenarios requiring both dynamic feasibility and safety.

In summary, the results demonstrate that the proposed MPC framework effectively generates smooth and dynamically feasible trajectories while maintaining acceptable positional accuracy. Through systematic parameter tuning and Pareto optimization, the approach balances competing objectives, outperforming classical interpolation methods in terms of control smoothness and physical realism. These findings reinforce the suitability of MPC for real-time trajectory planning in robotics and autonomous systems.

6 Discussion

The experimental results presented in Figures 1 and 2 highlight the practical advantages of the proposed MPC-based trajectory planning method when evaluated against classical spline-based interpolation techniques. While spline and B-spline methods excel in minimizing positional error—often passing precisely through or near designated waypoints—they do so without accounting for physical feasibility, particularly regarding actuator limitations. In contrast, MPC explicitly integrates system dynamics and input constraints, resulting in smoother control actions and more realistic motion profiles, albeit with a marginal increase in tracking error.

One of the most notable findings is the substantial reduction in acceleration magnitude achieved by the MPC strategy. As shown in the acceleration comparison plot, spline-based and especially piecewise linear interpolations tend to produce discontinuous or sharply fluctuating acceleration profiles. These high-frequency variations may seem trivial in simulation, but in real-world robotic or vehicular systems, they can lead to actuator wear, instability, or even physical infeasibility. The MPC trajectories, although not always geometrically optimal, demonstrate a consistent smoothness in acceleration, maintaining low variance and bounded peak values across different time steps.

Another insight emerges when analyzing velocity behavior. While splines offer smooth transitions in theory, the empirical results reveal that both B-splines and linear interpolation can yield unpredictable or jagged velocity profiles, especially in regions of high curvature or abrupt directional changes. MPC-generated velocities, by contrast, evolve gradually and remain within practical bounds, which is crucial for maintaining safe operation in robotic arms, drones, or autonomous vehicles.

Interestingly, the MPC framework introduces a trade-off between tracking accuracy and actuation smoothness, governed by the weight parameter λ in the cost function. A low λ yields more accurate tracking but at the expense of sharper control inputs, whereas a high λ emphasizes control smoothness with slight degradation in trajectory fidelity. The grid-based Pareto optimization employed in this study

effectively identifies parameter sets that balance these competing objectives. Notably, the configuration with $\Delta t=0.05$ and $N=40$ lies on the Pareto frontier, providing a good compromise between control effort and mean absolute error.

Ultimately, the findings validate the viability of MPC as a dynamically feasible trajectory planner for real-time applications. While spline methods may still be preferred for offline path planning or purely geometric applications, MPC offers a robust alternative where real-time constraint satisfaction, actuator safety, and motion smoothness are essential.

7 Conclusion

This study proposed and evaluated a Model Predictive Control (MPC) framework for generating dynamically feasible 3D trajectories through multiple waypoints, with a focus on achieving a balance between positional accuracy and control smoothness. By formulating the trajectory generation task as a constrained optimization problem and tuning parameters through grid-based Pareto analysis, the approach demonstrated flexibility and robustness in managing the inherent trade-offs between tracking fidelity and actuation effort. The experimental comparison against traditional interpolation techniques—cubic spline, B-spline, and linear interpolation—revealed that while MPC trajectories may exhibit slightly higher positional errors, they significantly outperform in terms of smoothness and physical feasibility. The reduction in acceleration spikes and consistent velocity transitions observed in the MPC outputs are particularly valuable for real-time applications where safety, energy efficiency, and hardware longevity are critical. Additionally, the use of Pareto-optimal configurations provides a principled method for controller parameter selection, ensuring performance can be tuned according to task-specific priorities. Overall, the findings establish MPC as a compelling alternative to purely geometric planners, especially in scenarios involving strict dynamic constraints. Future work will focus on extending this framework through adaptive learning strategies, such as reinforcement learning, to enable online tuning and adaptation to uncertain or changing system dynamics.

Acknowledgments: This study was funded by the research project "BR24992947—Development of Robots, Scientific, Technical, and Software for Flexible Robotization and Industrial Automation (RPA) in Automotive Industrial Enterprises in Kazakhstan Using Artificial Intelligence".

References

[1] Zhenxiang, L., Alnfiai, M. M., Jung, L. T., Alotaibi, N. N., Bajuri, M. Y., Alnefaie, S. M. M., & Ghoneim, M. E. (2025). Mobile Robot Control Using Pursuit–Evasion Differential Game Strategy for Double Integrator Dynamic Control with Deep Reinforcement Learning. Dynamic Games and Applications, 1-29.

[2] Cai, M., Qi, Z., Cao, Y., Wu, X., Xu, T., & Zhang, L. (2024). Development of reconfigurable electromagnetic actuation system with large workspaces: Design, optimization, and validation. IEEE Transactions on Automation Science and Engineering.

[3] Ahn, H., Park, J., Bang, H., & Kim, Y. (2022). Model predictive control-based multirotor three-dimensional motion planning with point cloud obstacle. Journal of Aerospace Information Systems, 19(3), 179-193.

[4] Li, Z., Zhao, P., Jiang, C., Huang, W., & Liang, H. (2022). A learning-based model predictive trajectory planning controller for automated driving in unstructured dynamic environments. IEEE Transactions on Vehicular Technology, 71(6), 5944-5959.

[5] Baselizadeh, A., Khaksar, W., & Torresen, J. (2022, October). Motion planning and obstacle avoidance for robot manipulators using model predictive control-based reinforcement learning. In 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1584-1591). IEEE.

[6] Toumieh, C., & Lambert, A. (2022). Decentralized multi-agent planning using model predictive

control and time-aware safe corridors. IEEE Robotics and Automation Letters, 7(4), 11110-11117.

[7] Xiao, A., Luan, H., Zhao, Z., Hong, Y., Zhao, J., Chen, W., ... & Meng, M. Q. H. (2022, May). Robotic autonomous trolley collection with progressive perception and nonlinear model predictive control. In 2022 International Conference on Robotics and Automation (ICRA) (pp. 4480-4486). IEEE.

[8] White, J., Jay, D., Wang, T., & Hubicki, C. (2022, October). Avoiding dynamic obstacles with real-time motion planning using quadratic programming for varied locomotion modes. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 13626-13633). IEEE.

[9] Lin, P., & Tsukada, M. (2022). Model predictive path-planning controller with potential function for emergency collision avoidance on highway driving. IEEE Robotics and Automation Letters, 7(2), 4662-4669.

[10] Wu, H., Long, X., & Lu, D. (2023). Trajectory planning and tracking for four-wheel independent drive intelligent vehicle based on model predictive control. SAE International Journal of Advances and Current Practices in Mobility, 5(2023-01-0752), 2471-2485.

[11] Şenbaşlar, B., Hönig, W., & Ayanian, N. (2023). RLSS: real-time, decentralized, cooperative, networkless multi-robot trajectory planning using linear spatial separations. Autonomous Robots, 47(7), 921-946.

[12] Niu, Y., Yan, X., Wang, Y., & Niu, Y. (2024). 3D real-time dynamic path planning for UAV based on improved interfered fluid dynamical system and artificial neural network. Advanced Engineering Informatics, 59, 102306.

[13] Yavari, M., Gupta, K., & Mehrandezh, M. (2022). Interleaved predictive control and planning for an unmanned aerial manipulator with on-the-fly rapid re-planning in unknown environments. IEEE Transactions on Automation Science and Engineering, 20(3), 1690-1705.

[14] Chu, H., Meng, D., Huang, S., Tian, M., Zhang, J., Gao, B., & Chen, H. (2023). Autonomous high-speed overtaking of intelligent chassis using fast iterative model predictive control. IEEE Transactions on Transportation Electrification, 10(1), 1244-1256.

[15] Chow, S., Chang, D., & Hollinger, G. A. (2023). Parallelized control-aware motion planning with learned controller proxies. IEEE Robotics and Automation Letters, 8(4), 2237-2244.

[16] Vivekanandan, R., Chang, D., & Hollinger, G. A. (2023, May). Autonomous underwater docking using flow state estimation and model predictive control. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1062-1068). IEEE.

[17] Pan, Y. J., Buchanan, S., Chen, Q., Wan, L., Chen, N., Forbrigger, S., & Smith, S. (2025). Survey on recent advances in planning and control for collaborative robotics. IEEJ Journal of Industry Applications, 14(2), 139-151.

[18] Song, J., Tao, G., Zang, Z., Dong, H., Wang, B., & Gong, J. (2023). Isolating trajectory tracking from motion control: A model predictive control and robust control framework for unmanned ground vehicles. IEEE Robotics and Automation Letters, 8(3), 1699-1706.

[19] Khan, S., Guivant, J., Li, Y., Liu, W., & Li, X. (2023). Hybrid model predictive control for unmanned ground vehicles. IEEE Transactions on Intelligent Vehicles, 9(1), 1537-1546.

[20] Daza, I. G., Izquierdo, R., Martínez, L. M., Benderius, O., & Llorca, D. F. (2023). Sim-to-real transfer and reality gap modeling in model predictive control for autonomous driving. Applied Intelligence, 53(10), 12719-12735.

[21] Katayama, S., Murooka, M., & Tazaki, Y. (2023). Model predictive control of legged and humanoid robots: models and algorithms. Advanced Robotics, 37(5), 298-315.

Chapter 18

Prediction and Mitigation of Wild Forest Fires using AI - Literature review

Contribution #41
Cite as:
Luís Travassos, João Santos, Diogo Pinho, Paulo Fernandes, Fernanda Brito Correia. Prediction and Mitigation of Wild Forest Fires using AI - Literature review. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-41,
  title ={Prediction and Mitigation of Wild Forest
     Fires using AI - Literature review},
  author={Luís Travassos and  João Santos and  Diogo
     Pinho and  Paulo Fernandes and  Fernanda Brito
     Correia},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Prediction and Mitigation of Wild Forest Fires Using AI - Literature review

Luís Travassos[1,*], João Santos[1], Diogo Pinho[1], Paulo Fernandes[1], Fernanda Brito Correia[1]
[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093, Coimbra, Portugal
*Corresponding author
a2021136600@isec.pt

Abstract
Wildfires pose an escalating global threat exacerbated by climate change and human activity, necessitating advanced predictive systems to mitigate ecological and societal impacts. This literature review evaluated 28 studies, from 2020 to 2025, to assess the integration of Business Intelligence & Analytics including Machine Learning and Deep Learning in wildfires prediction, based on PRISMA guidelines to ensure methodological rigor. Key findings highlighted the dominance of Random Forest and Convolutional Neural Networks in spatial-temporal modelling, persistent challenges in data heterogeneity and model interoperability, and a misalignment between academic metrics and operational priorities like early detection. The synthesis underscored the urgent need for interdisciplinary collaboration, lightweight architectures, and the stakeholder-eccentric frameworks to bridge theoretical advancements with employable solutions in wildfires management.

Keywords: Wildfire prediction, Business Intelligence and Analytics, Machine Learning, Deep Learning, Literature Review.

1 Introduction

Wildfires have emerged as a growing global threat, exacerbated by climate change, land mismanagement, and human expansion into fire-prone regions. These disasters cause irreversible ecological damage, economic losses, and pose significant risks to human safety. Traditional wildfire management approaches, often reactive and resource-intensive, are increasingly inadequate in addressing these challenges [28]. The integration of Business Intelligence & Analytics (BI&A) including traditional Machine Learning (ML) and Deep Learning (DL) techniques offers a transformative opportunity for proactive wildfire prediction and mitigation.

This literature review, from 2020 to 2025, in English, evaluated 28 studies to assess the effectiveness of BI&A and ML/DL in wildfire forecasting, focusing on model performance, data quality, and real-world implementation barriers. By synthesizing cutting-edge research, this study aimed to bridge the gap between theoretical advancements and operational deployment, providing actionable insights for researchers, policymakers, and disaster management agencies.

In recent years, the intersection of BI&A with traditional ML and DL has gained increasing traction as a promising avenue for enhancing wildfire prediction and response capabilities [6], [3]. The literature from 2020 to 2025 publication years revealed a significant shift from traditional rule-based and statistical approaches toward data-driven, Artificial Intelligence (AI) enhanced forecasting models. This transition was largely motivated by the growing availability of heterogeneous data sources—such as satellite imagery, meteorological data, and topographic information—and the advancement of computational resources capable of handling large-scale, high-dimensional datasets.

Furthermore, hybrid and ensemble models combining BI&A with ML/DL frameworks had been explored to enhance predictive reliability and decision-making support. These approaches leveraged the strengths of multiple algorithms and data fusion techniques, offering improved adaptability to diverse geospatial contexts. Several studies had also incorporated real-time data pipelines and geospatial dashboards, facilitating near-instantaneous risk assessments and visualizations that can aid emergency responders and policymakers [4], [12], [27].

Despite these advancements, literature identified persistent challenges. Data quality remained a fundamental issue, especially regarding inconsistencies across spatial and temporal resolutions. Model interpretability, particularly for DL models, was another critical barrier to real-world adoption, as stakeholders often require transparent and explainable outputs. Scalability and generalizability across regions with varying ecological and socio-economic conditions also continued to limit the operational

deployment of AI-driven wildfire management systems. Overall, the current body of research demonstrated considerable potential for the integration of BI&A with ML/DL in wildfire forecasting. However, achieving practical impact necessitates further work on model transparency, cross-domain interoperability, and the incorporation of domain knowledge into algorithmic design.
This literature review addressed 4 research questions (Table 1) to evaluate the BI&A including traditional ML and DL methods in wildfire prediction systems. These questions were used to analyse technical advancements, data ecosystems, performance metrics, and implementation barriers, aiming to bridge gaps between theoretical innovation and operational deployment.

Table 1 - RQ, motivations and connection to fire mitigation

RQ	Motivation	Connection to Fire Mitigation
RQ1: What are the current state-of-the-art BI&A techniques including traditional ML and DL models in wildfire prediction systems?	Identify effective models and their comparative advantages in wildfire contexts	High-accuracy predictions enable optimized resource allocation (e.g., firefighting teams, evacuations)
RQ2: What types of datasets and preprocessing methods are most employed?	Evaluate data quality, sources, and preprocessing challenges	Reliable integration of heterogeneous data improves model generalizability across regions
RQ3: What are the key performance metrics used for evaluating wildfire prediction models?	Assess the alignment of metrics with operational needs	Interpretable metrics foster trust and enable rapid decision-making by agencies
RQ4: What are the major implementation challenges in operational systems?	Highlight barriers like computational costs or cross-regional validation	Scalable, explainable solutions are critical for adoption in diverse ecosystems

The first research question (RQ1) focused on identifying state-of-the-art BI&A techniques including traditional ML and DL models that demonstrate high accuracy in wildfire contexts. Their efficacy in real-time prediction directly supported optimized resource allocation for firefighting and evacuation efforts. RQ2 answer examined datasets and preprocessing methods, analysing the integration of heterogeneous data sources like satellite imagery, IoT sensors, and climate records. Inconsistencies in resolution, format, and quality may hindered model generalizability, underscoring the need for standardized fusion techniques. RQ3 pretended to evaluate performance metrics, investigating if traditional indicators like accuracy failed to align with operational priorities such as early detection rates and false alarm minimization. Fire agencies require interpretable metrics to prioritize actionable insights during emergencies. Finally, RQ4 answer pretended to highlight implementation challenges, including the opacity of "black box" models, computational costs, and validation disparities across ecosystems. Scalability and explainability emerged as critical factors for adoption, particularly in regions with limited resources or diverse environmental conditions. Addressing these challenges requires balancing technical sophistication with practical usability to ensure solutions meet the evolving demands of wildfire mitigation.
In this paper, section 2 describes the methodology used section 3 presents the results obtained from the analysis of the papers, section 4 discusses the results and section 5 gives some conclusions. The final section lists the references used along the paper.

2 Methodology

This section outlines the systematic approach adopted to review recent advancements in BI&A applications for wildfire prediction and mitigation. The methodology was based on Preferred Reporting Items for Systematic Reviews and Meta-Analyzes (PRISMA) guidelines, ensuring rigor and transparency in article selection, quality assessment, and synthesis. Each phase of the review process, planning, execution, and analysis was structured to address the defined research questions (RQs) and validate the practical relevance of the findings.
This review employed a multi-phase screening process (Table 2) to refine an extensive pool of academic

literature into a focused corpus of studies relevant to BI&A-driven wildfire prediction. This structured approach balanced breadth and specificity, ensuring methodological rigor while addressing the rapid evolution of machine learning applications in fire management.

Table 2 - Search string refinement and results

Phase	Search String	Results
Initial Query	("Business Intelligence" OR "BI") AND ("Forest Fire" OR "Wildfire") AND ("Predict" OR "Forecast") AND ("Machine Learning" OR "Deep Learning") AND (Since 2020)	4,680
Title Screening	(intitle:"wildfire" OR intitle:"fire" OR intitle:"forest") AND (intitle:"analytics" OR intitle:"analyses" OR intitle:"analysis" OR intitle:"review" OR intitle:"case study") AND ("Business Intelligence" OR "BI") AND ("Forest Fire" OR "Wildfire") AND ("Predict" OR "Forecast") AND ("Machine Learning" OR "Deep Learning") AND (Since 2020)	77
Abstract Screening	[Abstract review to ensure the quality of inclusion criteria]	28

The initial query combined key terms across four domains, BI&A, wildfire terminology, predictive functions, and technical approaches, yielding 4,680 records. This broad scope captured recent advancements but included substantial noise from unrelated theoretical frameworks or non-technical studies. Title screening applied stringent criteria to the title, requiring explicit mention of analytical methods and wildfire contexts, which reduced results to 77 articles. This phase eliminated studies lacking practical validation, such as general fire science research or policy analyses without BI&A integration. Abstract screening prioritized technical depth and alignment with predefined quality criteria, further narrowing the selection to 28 studies. This final phase excluded articles with insufficient methodological detail, such as those omitting preprocessing steps or validation protocols. The iterative refinement ensured that the retained studies addressed operational challenges, such as data heterogeneity and model scalability, while maintaining relevance to real-world fire mitigation. The process underscored the necessity of systematic filtering to isolate actionable insights within a rapidly expanding research landscape.

The literature review applied predefined quality inclusion criteria (Table 3) to evaluate the methodological rigor and practical relevance of studies integrating BI&A and traditional ML and DL in wildfire prediction. These criteria focused on clarity of research objectives, methodological transparency, data documentation, metric alignment, and acknowledgment of implementation barriers, ensuring studies balanced technical innovation with operational applicability.

Table 3 – Quality Inclusion criteria

Quality Inclusion Criteria
Are the research objectives clearly stated and relevant to wildfire prediction?
Is the proposed methodology well-described?
Are the datasets and preprocessing methods adequately documented?
Are the evaluation metrics appropriate and comprehensively reported?
Are the limitations and practical implementation challenges discussed?

The first criterion assessed whether studies clearly defined objectives tied to wildfire prediction, such as early detection or spread modelling, ensuring alignment with real-world fire management needs. The second evaluated methodological descriptions, emphasizing the need for detailed workflows of BI&A methods including traditional ML algorithms and DL architectures to enable reproducibility. Studies lacking explicit documentation of preprocessing steps or model configurations were flagged, as incomplete methodologies hinder validation and scalability. The remaining criteria addressed data transparency and practical limitations. Adequate dataset documentation, including sources like satellite imagery or climate records, was critical to assess model generalizability across ecosystems. Similarly, comprehensive reporting of evaluation metrics ensured alignment with operational priorities, such as minimizing false alarms or latency in detection systems. Finally, discussions of limitations, such as

computational costs or regional validation gaps, were essential to contextualize findings and guide future improvements. Together, these criteria ensured the reviewed studies provided actionable insights while highlighting persistent challenges in translating theoretical models into deployable solutions.
The literature review followed a structured selection process based on PRISMA guidelines, as illustrated in the flow diagram (Figure 1).

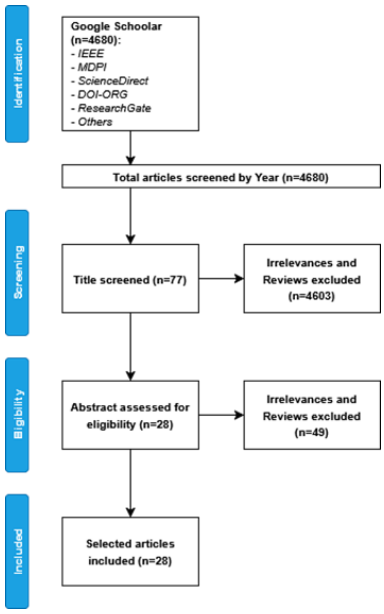


Figure 1 - PRISMA diagram

This methodology ensured transparency and rigor in identifying relevant studies at the intersection of BI&A for wildfire prediction, balancing comprehensive coverage with methodological relevance. The initial search across academic databases aggregated via Google Scholar yielded 4,680 records, reflecting broad interest in predictive analytics for fire management. Sources included peer-reviewed platforms such as IEEE, MDPI, and ScienceDirect, alongside repositories like ResearchGate. Title screening applied strict criteria, requiring explicit mention of analytical methods and wildfire contexts, reducing the pool to 77 studies. This phase eliminated 4,603 records, primarily theoretical ML papers, non-technical policy analyses, or studies lacking BI&A integration. Abstract screening further refined the corpus to 29 articles by prioritizing technical validation, dataset documentation, and alignment with predefined quality criteria. The final selection retained studies that addressed operational challenges such as data heterogeneity, model interpretability, and scalability. The annual distribution of publications highlighted a growing emphasis on BI&A-driven wildfire research, with a notable concentration of studies from 2023 onward. This trend underscores the accelerating adoption of advanced analytics to address escalating environmental threats.
The annual distribution of publications (Figure 2) revealed evolving research priorities in BI&A with machine learning applications for wildfire prediction, with notable fluctuations in output between 2020 and 2025.

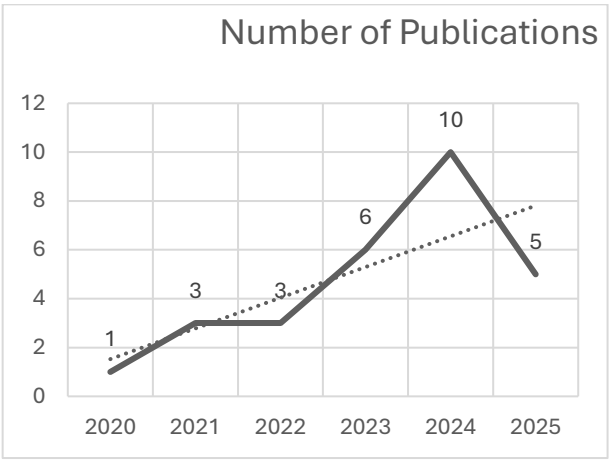


Figure 2 - Year-wise distribution of publications (2020–2023)

The distribution of articles across digital libraries highlights the diverse academic sources contributing to wildfire prediction research, reflecting varying publication practices and accessibility within the field. Peer-reviewed platforms such as IEEE and ScienceDirect collectively account for six articles, representing foundational technical frameworks and validation studies. MDPI contribute with five articles, emphasizing open-access trends and interdisciplinary collaboration in environmental analytics.

3 Results

This section synthesizes perspectives from the literature review, addressing state-of-the-art techniques, design considerations, and evaluation frameworks for BI&A driven wildfire prediction models.
The following analysis synthesizes the capabilities and limitations of prominent wildfire prediction models (Table 4), emphasizing their technical strengths, operational constraints, and alignment with real-world fire mitigation needs.
This evaluation focused on six computational approaches widely applied in wildfire research, highlighting their adaptability to environmental challenges and scalability in diverse ecosystems.

Table 4 - Analysis of wildfire prediction models

Model	Strengths	Weaknesses	References
Binary Logistic Regression (BLR)	Interpretable risk probability; low computational cost.	Limited to linear relationships; requires feature engineering.	[5], [12]
Convolutional Neural Networks (CNN)	High precision in fire/smoke detection via satellite imagery.	Performance degrades under fog/smoke; requires labelled image data.	[4], [6], [8], [21], [20]
Digital Twin Systems	Integrates real-time satellite/sensor data; enables scenario testing.	High computational/resource demands; require calibration.	[11], [12]
Hidden Markov Models (HMM)	Model's probabilistic transitions; interpretable.	Limited to annual predictions; ignores spatial factors.	[14], [12]
Long Short-Term Memory (LSTM)	Captures temporal dependencies in weather/fire data; ideal for seasonal risk.	Requires large datasets; computationally intensive for real-time deployment.	[13], [16], [17]
Random Forest (RF)	Handles heterogeneous data; quantifies feature importance.	Less effective for temporal predictions; risk of overfitting.	[3], [5], [19], [26]

BLR was valued for its interpretability and computational efficiency, making it suitable for initial risk assessments in resource-limited regions. However, its reliance on linear relationships and manual feature engineering restricts its utility in modelling complex, non-linear fire dynamics. CNNs exceled in detecting fires and smoke through high-resolution satellite imagery, achieving precision in spatial analysis. Challenges emerged under adverse conditions such as fog or smoke, where image quality deteriorates, and the dependency on labelled datasets complicate scalability. Digital Twin Systems integrated real-time data streams from satellites and sensors, enabling dynamic fire spread simulations. Their operational adoption was hindered by significant computational costs and the need for frequent calibration, particularly in regions with limited technical infrastructure. HMMs offered probabilistic insights into temporal fire transitions but overlooked spatial factors, limiting their application to annual risk forecasts rather than real-time predictions. LSTM networks effectively captured temporal dependencies in weather and fire data, ideal for seasonal risk modelling. Their reliance on extensive datasets and computational resources posed barriers to real-time deployment in areas with sparse historical records. RF handled heterogeneous data effectively, providing feature importance metrics for actionable insights. While robust for static predictions, it struggled with temporal sequences and risks overfitting without rigorous parameter tuning. The models' strengths and weaknesses reflected broader tensions between technical sophistication and practical usability. For instance, while CNNs and LSTMs advanced predictive accuracy, their resource demanded to highlight the need for lightweight alternatives in operational settings. Similarly, Digital Twin Systems and RF models demonstrated the potential of integrating diverse data sources but underscored the importance of balancing complexity with deplorability. These insights aligned with current research priorities, which emphasized adaptable frameworks capable of addressing escalating wildfire threats while accommodating ecological and operational diversity.

The literature review identified three primary data categories critical to wildfire prediction systems (Table 5), each with distinct sources, limitations, and implications for model reliability. These datasets, satellite imagery, weather data, and historical fire records formed the foundation of predictive analytics but faced challenges that affected their utility in diverse operational contexts.

Table 5 - Wildfire dataset landscape

Dataset Type	Sources	Limitations	References
Satellite Imagery	NASA FIRMS, MODIS	Low resolution, cloud obstruction	[3], [5], [8], [11], [18], [19], [20], [26], [27]
Weather Data	ERA5, NOAA	Temporal gaps, coarse spatial scales	[16], [23]
Historical Fire Records	Global Fire Atlas, CALFIRE	Biased toward populated regions	[3], [5], [6], [8], [11], [15], [18], [19], [23], [27]

Satellite imagery from NASA FIRMS and MODIS enabled broad spatial coverage for fire detection, yet low resolution and frequent cloud obstruction compromise precision in dynamic environments. Efforts to address these limitations focused on integrating auxiliary data sources, such as UAV imagery, to enhance resolution and filled observational gaps during adverse conditions. Weather data from ERA5 and NOAA provided essential climatic variables but suffered from temporal discontinuities and coarse spatial granularity, complicating localized fire spread modelling. Advanced downscaling techniques and ML interpolation were increasingly employed to refine spatial resolution and mitigate temporal inconsistencies. Historical fire records from Global Fire Atlas and CALFIRE offered valuable insights into fire behaviour but exhibited geographic biases, overrepresenting populated or monitored regions, while neglecting remote ecosystems. This skew distorted risk assessments and hindered model generalizability across diverse landscapes. Recent methodologies emphasized transfer learning and synthetic data generation to counteract biases, improving predictive accuracy in underrepresented areas. These advancements highlighted the ongoing need to balance data availability with ecological representativeness, ensuring wildfire models remain adaptable to the escalating variability of fire-prone environments.

The analysis of the frequency of the use of preprocessing techniques revealed their varying adoption across wildfire prediction studies, reflecting priorities in addressing data quality challenges. Feature selection was

utilized in 18 articles to reduce dimensionality and enhance model efficiency, while train-test splitting appeared in 20 studies to ensure robust validation. Outlier handling, though less prevalent, remained essential for minimizing noise in heterogeneous datasets. The predominance of missing data handling aligned with the challenges of integrating incomplete satellite or weather records, where imputation or exclusion strategies were vital for maintaining prediction reliability. Feature selection's prominence highlighted the need to balance computational complexity with actionable insights, particularly when processing multi-source data like satellite imagery and climate variables. Train-test splitting's widespread use emphasized the field's focus on reproducibility, though its application often overlooked temporal or spatial dependencies unique to wildfire dynamics. Outlier handling, while less prioritized, addressed anomalies arising from sensor errors or extreme weather events, which could distort risk assessments. The distribution of these techniques illustrated the iterative nature of data refinement in wildfire prediction, where preprocessing choices directly influenced model accuracy and operational trust. As data heterogeneity grows, standardized protocols for these methods will become increasingly critical to ensure consistency across diverse ecosystems and fire regimes.

The evaluation metrics employed in wildfire prediction research demonstrated a strong emphasis on comprehensive performance assessment, with Receiver Operating Characteristic Area Under the Curve (ROC-AUC) emerging as the most widely reported indicator, utilized in 24 studies. This preference reflected its ability to balance sensitivity and specificity, critical for models operating in imbalanced fire/no-fire scenarios. Precision, Recall, and F1-score followed closely, appearing in 22, 21, and 23 studies respectively, highlighting the field's dual focus on minimizing false alarms and maximizing early detection rates. ROC-AUC's dominance underscored its suitability for evaluating models tasked with prioritizing high-risk areas amid data imbalances, a common challenge in wildfire datasets. Precision's prevalence aligned with operational demands to reduce false positives, ensuring resource allocation is not diverted by erroneous alerts. Recall's slightly lower adoption might indicate gaps in addressing early detection needs, as maximizing true positive rates remained vital for timely evacuations and pre-emptive containment. The F1-score's intermediate position bridged these priorities, offering a balanced measure for stakeholders requiring both accuracy and reliability. These trends revealed a consensus on the importance of multi-metric evaluation but also highlighted potential mismatches between academic benchmarks and operational realities. For instance, while ROC-AUC provided theoretical robustness, fire agencies often prioritized latency-adjusted metrics absent from standardized reporting. The distribution underscored the need for domain-specific evaluation frameworks that aligned technical performance with actionable outcomes, ensuring predictive models met the urgent, evolving demands of wildfire management.

4 Discussion

The literature review revealed persistent barriers to deploying robust wildfire prediction systems, while identifying opportunities for advancing BI&A applications in fire management. This review identified persistent challenges and future priorities for advancing wildfire prediction systems and were categorized as seen in Table 6. These categories reflected the multifaceted nature of developing reliable, scalable solutions capable of addressing escalating fire risks amid ecological and operational complexities.

Dataset characteristics faced limitations in geographical and temporal coverage, often focusing on specific regions or short-term records, which restricted model generalizability. Future efforts emphasize collecting multi-year, multi-region data to capture diverse fire regimes and integrating socioeconomic variables like land use patterns, which were frequently excluded despite their influence on fire susceptibility. Preprocessing challenges centred on noisy or incomplete sensor data, driving the need for automated cleaning pipelines to streamline data refinement. Severe class imbalances between fire and non-fire instances further complicated model training, prompting exploration of synthetic data generation to enhance representation of rare but critical fire events. Model selection grappled with accuracy compute trade-offs, particularly for hybrid architectures like CNN-LSTM networks, which struggled with real-time deployment due to high resource demands. Regularization techniques and ensemble methods were prioritized to improve robustness in data-scarce environments. Interpretability remained a critical barrier, as opaque "black box" models hinder trust among firefighters and policymakers. Hybrid systems combining rule-based logic with AI, alongside model distillation for transparency, were proposed to bridge this gap. The lack of widespread adoption of explainability tools like SHAP or LIME underscored the urgency of developing intuitive interfaces that aligned technical outputs with operational decision-making. These insights charted a path toward adaptive frameworks that balanced

innovation with practicality, ensuring predictive systems evolved in tandem with ecological dynamics and stakeholder needs. By addressing data biases, computational constraints, and transparency gaps, the field can transition from theoretical excellence to operational resilience, empowering communities to anticipate and mitigate wildfire threats proactively.

Table 6 - Challenges and future directions in wildfire prediction

Category	Challenges	Future Directions
Dataset Characteristics	Limited geographical/temporal coverage	Collect multi-region, multi-year data
	Exclusion of human activity/land use factors	Integrate socioeconomic variables
Pre-Processing	Noisy/missing sensor data	Develop automated cleaning pipelines
	Severe class imbalance (fire vs. non-fire)	Synthetic data generation
Model Selection	Accuracy-compute trade-offs (e.g., LSTM/CNN hybrids)	Regularization techniques for small datasets
	No consensus on optimal architecture	Ensemble methods for robustness
Interpretability	Black-box neural networks hinder stakeholder trust	Hybrid rule-based + AI systems
	Lack of SHAP/LIME adoption in reviewed studies	Model distillation for transparency

5 Conclusion

The integration of BI&A with traditional ML and DL techniques had undeniably revolutionized wildfire prediction capabilities, offering unprecedented accuracy in risk assessment and early detection.
The review synthesized key findings across four research questions (Table 7), offering a comprehensive assessment of advancements and persistent gaps in wildfire prediction systems.

Table 7 - Key findings across Research Questions

RQ	Answer
RQ1: What are the current state-of-the-art BI&A techniques including traditional ML and DL models in wildfire prediction systems?	RF and CNNs dominated due to their efficacy in handling spatial-temporal data and heterogeneous inputs.
RQ2: What types of datasets and preprocessing methods are most employed?	Satellite imagery (NASA FIRMS, MODIS), weather data (ERA5, NOAA), and historical fire records were primary sources.
RQ3: What are the key performance metrics for evaluating wildfire prediction models?	ROC-AUC and F1-score were most prevalent, balancing sensitivity and specificity.
RQ4: What are the major implementation challenges in operational systems?	Key barriers included "black box" model opacity, high computational costs for real-time deployment, and cross-regional validation failures.

The dominance of RF and CNNs in wildfire prediction underscored their adaptability to spatial-temporal data and heterogeneous inputs, such as satellite imagery and multi-source environmental variables. Hybrid architecture, including CNN-LSTM models, demonstrated potential for capturing both spatial and temporal fire dynamics but faced scalability challenges due to computational intensity. While deep learning models achieved high accuracy, their "black box" nature and resource demands hindered real-world adoption, particularly in resource-constrained regions.
Satellite imagery from NASA FIRMS and MODIS, alongside weather data from ERA5 and NOAA, formed the backbone of wildfire datasets, supplemented by historical fire records. However,

preprocessing these datasets remained a critical hurdle, with missing data imputation and feature selection prioritized to address gaps and noise. Spatial-temporal splitting methods were increasingly adopted to validate models across diverse ecosystems, though resolution mismatches and geographic biases persisted, limiting generalizability.

Performance metrics like ROC-AUC and F1-score dominated academic evaluations, balancing Sensitivity and Specificity in fire/no-fire classification. However, operational priorities such as early detection rates (Recall) and false alarm minimization remained underrepresented, revealing a disconnection between technical benchmarks and frontline needs. This gap emphasized the necessity for domain-specific metrics that align with evacuation timelines and resource allocation protocols.

Implementation challenges were centred on translating theoretical models into deployable tools. The opacity of complex neural networks eroded stakeholder trust, while high computational costs restricted real-time deployment of advanced architectures. Cross-regional validation failures further exposed limitations in model adaptability, as algorithms trained on specific ecosystems often underperformed in geographically distinct areas. Addressing these barriers requires integrating socioeconomic variables, such as land use patterns and human activity, into datasets, alongside hybrid systems that combine AI with rule-based logic for interpretability. Scalable cloud-edge architectures and model distillation techniques emerged as critical enablers for balancing accuracy with operational feasibility.

These conclusions collectively advocate for interdisciplinary frameworks that prioritize not only technological innovation but also ecological representativeness, computational efficiency, and stakeholder collaboration. By harmonizing data quality, model transparency, and evaluation rigor, the field can advance toward predictive systems capable of mitigating wildfire risks in an era of escalating environmental uncertainty.

References

[1] S. Sonwani, S. Hussain, and P. Saxena, “Air pollution and climate change impact on forest ecosystems in Asian region – a review,” *Ecosystem Health and Sustainability*, vol. 8, no. 1, Dec. 2022, doi: 10.1080/20964129.2022.2090448.

[2] M. Lombardi, D. Berardi, and M. Galuppi, “A Critical Review of Fire Tests and Safety Systems in Road Tunnels: Limitations and Open Points,” *Fire 2023, Vol. 6, Page 213*, vol. 6, no. 5, p. 213, May 2023, doi: 10.3390/FIRE6050213.

[3] A. Usta and O. Akbulut, “Exploring the Potential of Machine Learning Approaches in Fire Detection: A Case Study in Marmara and Kocaeli Region,” in *HORA 2024 - 6th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/HORA61326.2024.10550458.

[4] C. J. Liang, S. H. Lu, J. J. Liang, F. C. Lin, and P. R. Yu, “Integrated Air Quality Monitoring and Alert System Based on Two Image Analysis Techniques for Reportable Fire Events,” *Atmosphere 2021, Vol. 12, Page 117*, vol. 12, no. 1, p. 117, Jan. 2021, doi: 10.3390/ATMOS12010117.

[5] S. Saha, M. Saha, K. Mukherjee, A. Arabameri, P. T. T. Ngo, and G. C. Paul, “Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, REPTree: A case study at the Gumani River Basin, India,” *Science of The Total Environment*, vol. 730, p. 139197, Aug. 2020, doi: 10.1016/J.SCITOTENV.2020.139197.

[6] D. Gragnaniello, A. Greco, C. Sansone, and B. Vento, “FLAME: fire detection in videos combining a deep neural network with a model-based motion analysis,” *Neural Comput Appl*, vol. 37, no. 8, pp. 6181–6197, Jan. 2025, doi: 10.1007/S00521-024-10963-Z/TABLES/6.

[7] X. Jiang, Y. Eum, and E. H. Yoo, “The impact of fire-specific PM2.5 calibration on health effect analyses,” *Science of The Total Environment*, vol. 857, p. 159548, Jan. 2023, doi: 10.1016/J.SCITOTENV.2022.159548.

[8] A. M. Goncalves, T. Brandao, and J. C. Ferreira, “Wildfire Detection with Deep Learning - A Case Study for the CICLOPE Project,” *IEEE Access*, vol. 12, pp. 82095–82110, 2024, doi: 10.1109/ACCESS.2024.3406215.

[9] D. Gragnaniello, A. Greco, C. Sansone, and B. Vento, “Fire and smoke detection from videos: A literature review under a novel taxonomy,” *Expert Syst Appl*, vol. 255, p. 124783, Dec. 2024, doi: 10.1016/J.ESWA.2024.124783.

[10] S. Yang, Q. Huang, and M. Yu, “Advancements in remote sensing for active fire detection: A review of datasets and methods,” Sep. 15, 2024, *Elsevier B.V.* doi: 10.1016/j.scitotenv.2024.173273.

[11] Y. Li, T. Zhang, Y. Ding, … R. W.-J. of F., and undefined 2025, “Review and perspectives of digital twin systems for wildland fire management,” *Springer*, vol. 36, no. 1, p. 14, Dec. 2025, doi: 10.1007/s11676-024-01810-x.

[12] C. Di and L. Magistrale, “A proposal for a Wildfire Digital Twin Framework through automatic extraction of remotely sensed data: the Italian case study of the Susa Valley,” 2023, Accessed: Mar. 11, 2025. [Online]. Available: https://webthesis.biblio.polito.it/27289/

[13] R. Tao, M. Zhu, H. Cao, and H. E. Ren, “Parameterization before Meta-Analysis: Cross-Modal Embedding Clustering for Forest Ecology Question-Answering,” *Forests 2024, Vol. 15, Page 1670*, vol. 15, no. 9, p. 1670, Sep. 2024, doi: 10.3390/F15091670.

[14] “AN ANALYSIS ON HIDDEN MARKOV MODEL PREDICTION FACTORS USING CALIFORNIA WILDFIRE AND METEOROLOGICAL RECORDS,” 2019.

[15] H. Podschwit *et al.*, “Reliability of cross-regional applications of global fire danger models: a Peruvian case study,” *Fire Ecology*, vol. 18, no. 1, Dec. 2022, doi: 10.1186/s42408-022-00150-7.

[16] D. A. Dzulhijjah *et al.*, “Comparative Analysis of Hybrid Long Short-Term Memory Models for Fire Danger Index Forecasting

with Weather Data," in *2023 6th International Conference on Information and Communications Technology, ICOIACT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 165–170. doi: 10.1109/ICOIACT59844.2023.10455879.
[17] X. Hu, "Fire frequency analysis and prediction based on back propagation neural network model and Long-Short Term Memory model," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1755-1315/675/1/012169.
[18] Y. Su, L. Zhao, H. Li, X. Li, J. Chen, and Y. Ge, "An Efficient Task Implementation Modeling Framework with Multi-Stage Feature Selection and AutoML: A Case Study in Forest Fire Risk Prediction," *Remote Sens (Basel)*, vol. 16, no. 17, p. 3190, Sep. 2024, doi: 10.3390/RS16173190/S1.
[19] Y. Hao, M. Li, J. Wang, X. Li, and J. Chen, "A High-Resolution Spatial Distribution-Based Integration Machine Learning Algorithm for Urban Fire Risk Assessment: A Case Study in Chengdu, China," *ISPRS International Journal of Geo-Information 2023, Vol. 12, Page 404*, vol. 12, no. 10, p. 404, Oct. 2023, doi: 10.3390/IJGI12100404.
[20] Maskouni, F. H., & Seydi, S. T. (2021). Forest Burned Area Mapping Using Bi-Temporal Sentinel-2 Imagery Based on a Convolutional Neural Network: Case Study in Golestan Forest †. *Engineering Proceedings*, *10*(1). https://doi.org/10.3390/ecsa-8-11291
[21] J. L. Wilk-Jakubowski *et al.*, "Acoustic Waves and Their Application in Modern Fire Detection Using Artificial Vision Systems: A Review," *Sensors (Basel)*, vol. 25, no. 3, p. 935, Feb. 2025, doi: 10.3390/S25030935.
[22] "SYMPOSIA ON SYSTEMS ANALYSIS IN FOREST RESOURCES: A 50TH YEAR ANNIVERSARY (1975-2024)." Accessed: Mar. 18, 2025. [Online]. Available: https://www.researchgate.net/profile/B-Bruce-Bare/publication/385910216_SYMPOSIA_ON_SYSTEMS_ANALYSIS_IN_FOREST_RESOURCES_A_50TH_YEAR_ANNIVERSARY_1975-2024/links/673b76eab903016a31c3cfa5/SYMPOSIA-ON-SYSTEMS-ANALYSIS-IN-FOREST-RESOURCES-A-50TH-YEAR-ANNIVERSARY-1975-2024.pdf
[23] R. N. Andarini and M. A. Hasan, "The interplay of community, corporate, and government interests: An analysis of forest and land fires (FLFS) in South Sumatera," *Interaction, Community Engagement, and Social Environment* , vol. 2, no. 2, pp. 107–122, Jan. 2025, doi: 10.61511/ICESE.V2I2.2025.1158.
[24] N. Dalila Ismail, R. Ramli, and M. Nizam Ab Rahman, "A Systematic Literature Review of Vision-Based Fire Detection, Prediction and Forecasting," *Jurnal Kejuruteraan*, vol. 37, no. 1, pp. 191–218, 2024, doi: 10.17576/jkukm-2025-37(1)-13.
[25] J. Qi, J. Z.-R. Analysis, and undefined 2024, "A review of optimization and decision models of prescribed burning for wildfire management," *Wiley Online Library*, 2024, doi: 10.1111/risa.17680.
[26] U. ScholarWorks, U. Graduate Theses, and A. Al Saim, "Machine Learning & Big Data Analyses for Wildfire & Air Pollution Incorporating GIS & Google Earth Engine," 2021, Accessed: Mar. 11, 2025. [Online]. Available: https://search.proquest.com/openview/f2a5891e8d6d318322d5a674a446dbf7/1?pq-origsite=gscholar&cbl=18750&diss=y
[27] M. D.-E. A. E. REVIEW and undefined 2023, "Forest Fire Incident Forecasting System In Permanent Reserved Forest In Peninsular Malaysia Using Big Data Analytics," *academia.edu*, 2023, Accessed: Mar. 11, 2025. [Online]. Available: https://www.academia.edu/download/107204501/3_52455_GP.pdf
[28] M. Yandouzi *et al.*, "REVIEW ON FOREST FIRES DETECTION AND PREDICTION USING DEEP LEARNING AND DRONES," *J Theor Appl Inf Technol*, vol. 30, no. 12, 2022, Accessed: Mar. 28, 2025. [Online]. Available: https://www.researchgate.net/profile/Idriss-Idrissi/publication/361739530_REVIEW_ON_FOREST_FIRES_DETECTION_AND_PREDICTION_USING_DEEP_LEARNING_AND_DRONES/links/62c2df25412e4c2aaeaa91a5/REVIEW-ON-FOREST-FIRES-DETECTION-AND-PREDICTION-USING-DEEP-LEARNING-AND-DRONES.pdf

Chapter 19

Combining Clustering and Predictive Algorithms for Real Estate Market Trends – Literature Review

Contribution #42
Cite as:
Sara Pinto, Nuno Fidalgo, Halima Moukrime, Tiago Dias, Rafaela Carvalho, Fernanda Brito Correia. Combining Clustering and Predictive Algorithms for Real Estate Market Trends – Literature Review. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-42,
  title ={Combining Clustering and Predictive
     Algorithms for Real Estate Market Trends
     Literature Review},
  author={Sara Pinto and  Nuno Fidalgo and  Halima
     Moukrime and  Tiago Dias and  Rafaela Carvalho and
      Fernanda Brito Correia},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Combining Clustering and Predictive Algorithms for Real Estate Market Trends – Literature Review

Sara Oliveira Pinto[1,*], Nuno Rafael Reis Teixeira Fidalgo[1], Halima Moukrime[2], Tiago Alexandre Pais Dias[1], Rafaela Oliveira Carvalho[1], Fernanda Brito Correia[1]

[1] Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093, Coimbra, Portugal

[2] Università di Reggio Emilia e Modena, Via P.Vivarelli 10, 42125 Modena MO, Italy,

[*]Corresponding author

a2021129757@isec.pt

Abstract

The real estate market is a complex and dynamic sector influenced by various economic, social, and environmental factors. Clustering and predictive modeling have emerged as powerful tools for analyzing market trends, estimating property values, and segmenting real estate assets. This paper presents a literature review on the application of predictive models integrated with clustering algorithms within the field of Machine Learning. This combined approach has demonstrated improved prediction accuracy, particularly when using Regression and Decision Tree-Based models. However, clustering techniques alone have not shown a significant impact on enhancing property categorization. The review highlights the advantages of combining clustering techniques with predictive methods for precise market prediction and segmentation, demonstrating their effectiveness through case studies and comparative analyses. Nevertheless, some critical issues remain, especially those related to data complexity. All these findings contribute to the advancement of data-driven strategies in real estate market analysis, supporting informed decision-making for investors, policymakers, and real estate professionals.

Keywords: real estate price prediction; predictive modeling; Clustering; Machine Learning; trend analysis

1 Introduction

The real estate market plays a key role in economic stability, with trends shaped by factors such as economic conditions, population growth, interest rates, and policy changes. Traditional analytical methods often struggle to handle the market's complexity. Recently, data analytics, including Machine Learning (ML) had emerged as a powerful tool for predicting prices, identifying trends, and segmenting markets [1, 2, 3, 4, 5].

Predictive modeling enables accurate property valuation, while clustering groups similar properties for clearer market segmentation [6, 7, 8, 9]. This paper explores the combined use of these techniques to enhance real estate analysis, offering practical insights for investors, developers, policymakers, and homebuyers [3, 8, 10].

Four Research Questions (RQs) were formulated to guide the exploration of ML techniques in the context of clustering and predictive modeling within the real estate market. These RQs aimed to provide a comprehensive understanding of the impact of various ML methods on model performance, as well as the challenges faced in optimizing and selecting appropriate algorithms. The list of the RQs is in Table 1, where N is RQ number.

Overall, this work does an analysis of the existing literature on clustering techniques and predictive modeling applied to the real estate market. After this introduction, where the RQs are presented, the methodology used to identify and select the relevant studies are described. Then, the results that emerged from the analysis are presented followed by the reflections derived from these results offering answers to the RQs highlighting both the main challenges and possible strategies to address them. Finally, the study concludes with a summary of its main contributions, acknowledges its limitations, and proposes directions for future research.

Table 1: List of Research Questions

N	RQ	Description
RQ1	How do different ML approaches impact the efficiency of clustering and predictive models?	To examine how different ML techniques influence prediction accuracy and clustering efficiency.
RQ2	How do supervised and unsupervised ML techniques complement each other in clustering and predictive modeling?	To explore how supervised and unsupervised learning combine to improve clustering and prediction.
RQ3	How can ML algorithms be tailored to enhance both clustering and predictive performance?	To examine the optimization of ML algorithms for clustering and prediction via tuning, feature selection, and hybrid methods.
RQ4	What are the key challenges in selecting and optimizing ML algorithms for clustering and prediction?	To address clustering and prediction methods for real estate, with focus on overfitting, scalability, and interpretability.

2 Methodology

To conduct this review, an analysis of the literature was carried out. The literature review followed a stage process inspired by the methodology proposed by Kitchenham [11]. The search strategy begins by selecting Google Scholar to track scientific papers that address the intersection of clustering, prediction, and the real estate market. The central theme of this review focuses on research articles that explore "Clustering," "Predicting," "Real Estate Market," "Prices," and "machine learning". The search string used to retrieve articles from Google Scholar was: *"Clustering AND Predicting AND 'Real Estate Market' AND Prices AND 'machine learning'"*.

To ensure the relevance of the articles, we applied two search preconditions: the articles must be written in English and published since 2020. These criteria help to narrow the search to the most recent and pertinent studies in the field, focusing on modern approaches to clustering and prediction models using ML applied to real estate pricing. The inclusion and exclusion criteria for this review are explained in Table 2.

Table 2: Inclusion and exclusion criteria

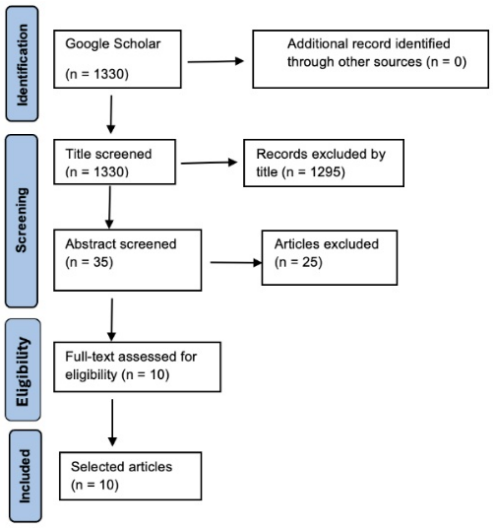
Inclusion criterium	Exclusion criterium
Discuss or apply ML algorithms like K-Means, DBSCAN, Random Forest, Extreme Gradient Boosting (XGBoost), Neural Networks, to predict or cluster real estate price data.	Address other areas of predicting with ML or clustering that do not involve the real estate market.

These inclusion and exclusion criteria were used to narrow the selection of relevant studies during the review process, ensuring that only articles that directly contribute to the intersection of ML techniques and real estate price prediction or clustering were considered.

The identification phase began with a Google Scholar search, using the mentioned search string, which initially retrieved 1330 articles. Applying the exclusion criteria in the analysis of the titles, 1295 articles were excluded and, when applying the exclusion criteria in the analysis of the abstracts, 10 articles remained.

During the eligibility phase, these 10 articles were fully analyzed to ensure they met the inclusion criteria, focusing on the application of ML algorithms for clustering and predictive modeling in the real estate market. Ultimately, these 10 articles were confirmed as the final set for review and included in the study, Figure 1.

Figure 1. PRISMA flow diagram



3 Results

Figure 2 shows the trend of publications over 5 years. Out of the 10 selected articles, the highest number were published in 2024.

The studies identified used different datasets and arrived at varied conclusions, reflecting the diversity of approaches and contexts within the real estate domain. Table 3 summarizes the datasets used in each study.

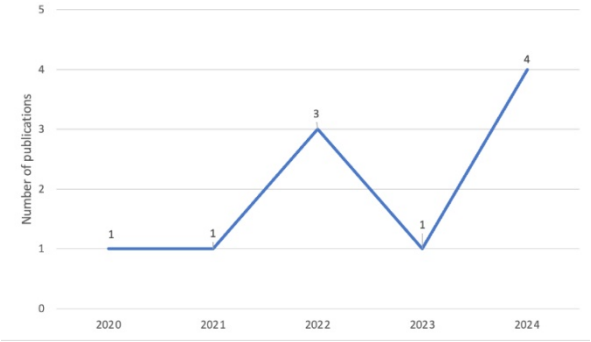


Figure 2: Year-wise distribution of publications selected

Table 3: Datasets from selected articles

Article	Datasets
[1]	Compiled using data obtained from various internet real estate sites
[2]	Web scraping from Kadıköy, Istambul; Kaggle dataset: The Ames Housing"
[3]	Property data in the city of Bangalore, India
[4]	Datasets include UCI's Boston Housing, Zillow (USA), Taiwan Real Estate Transactions, Kaggle, Magicbricks, 99 acres, government sources, REPD-3000 (textual/visual data), geospatial/satellite images, and real data from Taichung, Taiwan (2016–2019).
[5]	Dataset provided by the consulting firm Empirica ag (2016-2019)
[6]	Sourced from the Ministry of the Interior, Taiwan

Article	Datasets
[7]	Data analyzed comes from user interactions on an online real estate portal
[8]	No information
[9]	Historical transaction data from the Swedish market from the cadastral and land registration authority.
[10]	The data used was prepared with 121 criteria, incorporating approximately 200,000 real estate values.

The review revealed a wide range of predictive algorithms used in real estate trend analysis. Multiple Linear Regression (MLR) was the most frequent (60%), followed by Least Absolute Shrinkage and Selection Operator (LASSO) (50%), Random Forest, RIDGE Regression (40%), and Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), XGBoost, and LightGBM (30% each). Boosting methods like Adaptive Boosting (AdaBoost), appeared in 20% of studies. Regularization techniques (Bayesian Ridge Regression, RANSAC Regression, ElasticNet) were also applied in 10–20% of cases. Specialized models such as Multivariate Adaptive Regression Splines (MARS), Least Squares Support Vector Regression (LSSVR), Naive Bayes, Logistic Regression, SVM, and Genetic Algorithms (GA) were used in 10–20% of the articles. Neural network variants like Back-propagation Neural Network (BPNN) and General Regression Neural Network (GRNN) appeared in 20%, and Convolutional Neural Network (CNN) in 10%. Hybrid/ensemble models such as Particle Swarm Optimization-Bagging-Artificial Neural Networks (PSO-Bagging-ANNs) and GSNE were found in 10% of the studies. Tree-based models like Classification and Regression Tree (CART) were applied in 20%, showing good interpretability. Figure 3 presents an overview of the predictive models used in the selected articles. These predictive models were categorized into Regression Models, Decision Trees, Tree-Based Models, Neural Networks, Optimization and Hybrid methods and Ensemble Models emphasizing the diversity of approaches applied to real estate price prediction and classification tasks.

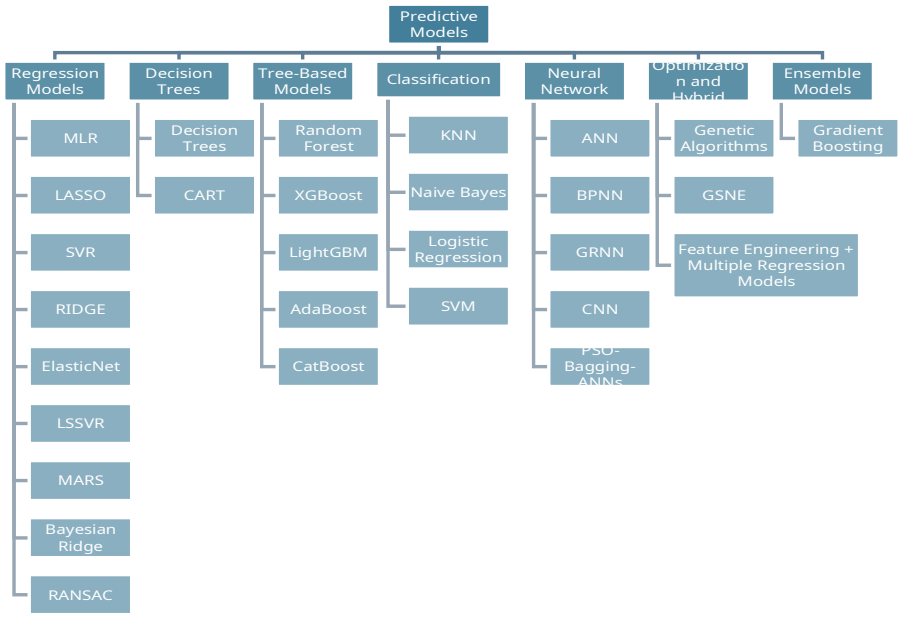


Figure 3: Taxonomy of predictive models

Table 4 to Table 8, Figure 4 and Table 9 show the studies that used the mentioned algorithms, ordered by decreasing number of papers that use them, each table for each of the categories.

Clustering techniques used by the studies were diverse. K-Means was the most used (20%), followed by Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Gaussian Mixture, Hierarchical Clustering, Spatially Constrained Multivariate Clustering Analysis (SCMCA), Cluster and Outlier Analysis, Prototype-Based Models, and Data Mining (10% each).

Table 4: Regression Models applied by selected articles

Algorithm	Papers
MLR	[1, 2, 3, 4, 6, 7]
LASSO	[1, 2, 4, 5, 8]
SVR	[2, 4, 5, 10]
RIDGE Regression	[1, 2, 4, 5]
ElasticNet	[1, 4]
LSSVR	[6]
MARS	[1]
Bayesian Ridge Regression	[5]
RANSAC Regression	[5]

Table 5: Decision Tree Models applied by selected articles

Algorithm	Papers
Decision Trees	[2, 3, 4]
Classification and Regression Trees	[6]

Table 6: Tree-Based Models applied by selected articles

Algorithm	Papers
Random Forest	[1, 2, 4, 10]
XGBoost	[1, 2, 4]
LightGBM	[1, 4, 5]
AdaBoost	[1, 2]
CatBoost	[1]

Table 7: Classification Models applied by selected articles

Algorithm	Papers
K-Nearest Neighbors	[1, 2, 4]
Naive Bayes	[4, 10]
Logistic Regression	[4]
SVM	[5]

Table 8: Neural Network Models applied by selected articles

Algorithm	Papers
ANN	[1, 4, 7]
BPNN	[4, 6]
GRNN	[4, 6]

CNN	[4]
PSO-Bagging-ANNs	[4]

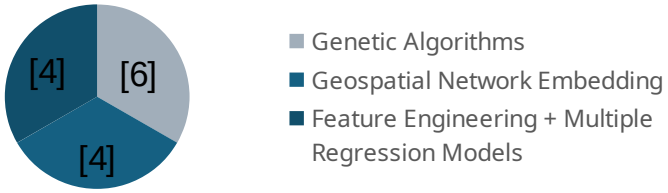


Figure 4: Optimization and Hybrid Models applied by selected articles

Table 9: Ensemble Models applied by selected articles

Algorithm	Papers
Gradient Boosting	[4]

Figure 5 shows the taxonomy of the clustering algorithms used in the selected articles. They were categorized as Partition-Based Models, Hierarchical Models, Prototype-Based Models and Spatial Models.
Table 10 to Table 13 show the studies that used the mentioned algorithms, ordered by decreasing number of papers that use them, each table for each of the categories.
Evaluating the performance of predictive models and clustering algorithms in the real estate market is essential to ensure the accuracy and effectiveness of the predictions and groupings generated. Next, are presented the main metrics used to evaluate prediction and clustering models applied to trend analysis in the real estate market, according to the studies analyzed.

- Regression Models: Mean Absolute Percentage Error (MAPE) was the most used metric (50%), followed by Mean Absolute Error (MAE) (40%), Root Mean Square Error (RMSE) and Normalized Mean Absolute Error (NMAE) (30%), Coefficient of Determination (COD), Percentage Root-mean-square Difference (PRD), Mean Squared Error (MSE) and R-squared (R^2) (20%), and Adjusted R^2, Root Mean Squared Logarithmic Error (RMSLE), Cross-validation (10%).
- Classification Models: Accuracy was the most frequent (30%), while True Positive Rate (TPR), True Negative Rate (TNR), Precision, Recall, F1-measure, and Support were used in 10% of the studies.

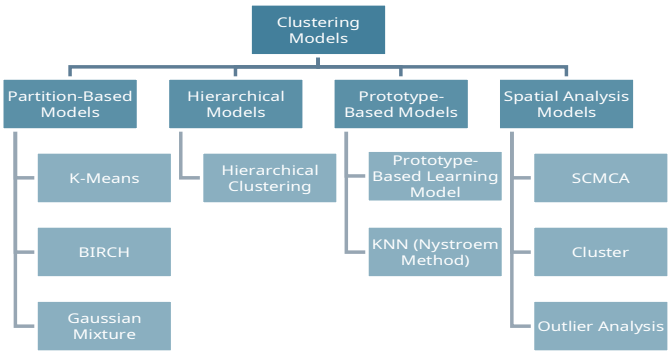


Figure 5: Taxonomy of clustering models

Table 10: Partition-Based Models applied by selected articles

Algorithm	Papers
K-Means	[1, 2]
BIRCH	[1]
Gaussian	[1]

Table 11: Hierarchical Models applied by selected articles

Algorithm	Papers
Hierarchical Clustering	[1]

Table 12: Prototype-Based Models applied by selected articles

Algorithm	Papers
Prototype-Based Learning	[8]
KNN with Nystroem method	[5]

Table 13: Spatial Analysis Models applied by selected articles

Algorithm	Papers
SCMCA	[9]
Cluster and Outlier Analysis	[9]

Table 14 presents the main conclusions drawn by each study.

Table 14: Conclusions from selected articles

Article	Conclusions
[1]	Best weak model: Ridge (MAPE 4.9%, MAE 30,756 TL); best meta-model: ANN (MAPE 3.2%, MAE 19,931 TL), outperforming all individual models.
[2]	The study concluded that the hybrid model (Linear Regression (LR), clustering analysis, KNN and Support Vector Regression (SVR)) is effective in predicting real estate prices, offering higher accuracy compared to traditional methods.
[3]	The application of ML algorithms (LR and Decision Tree) can improve accuracy in predicting property prices, facilitating more informed decisions in the property market.
[4]	ML was used in 61.5% of studies; Deep Learning, though promising, was underused (7.7%); Hybrid models (30.8%) offered better accuracy; ensemble methods like Random Forest, XGBoost, and Light Gradient Boosting Machine (LightGBM) were most effective; key factors included geographical, structural, and macroeconomic influences.
[5]	Combining clustering techniques with predictive algorithms improved the analysis of real estate market trends, although accuracy depends on the quality and detail of the data available.
[6]	LSSVR with genetic tuning and feature selection effectively predicts real estate prices; ML models are reliable for prediction using transaction data; feature selection boosts accuracy; Deep learning and unstructured data hold future potential.
[7]	The study demonstrated that personalization based on interactions with digital maps improved property recommendations and can help users make more informed decisions about purchasing and investing in real estate.
[8]	The prototype-based learning model can capture non-linear relationships and offers a degree of explainability by summarizing real estate datasets into a few representative “prototypes”.
[9]	The proposed methodology, which combines dataset optimization, cluster analysis and multiple regression, significantly improved the performance of massive property evaluation. Using geographic value clusters resulted in more accurate predictions than applying the model across the entire study area.

Article	Conclusions
[10]	K-Means was the top clustering model. For prediction, Naive Bayes, Random Forest, and Support Vector Machine (SVM) were evaluated, with Naive Bayes achieving 99.6% accuracy and Random Forest and SVM both reaching 99.9%.

5 Discussion

The choice of appropriate metrics to evaluate predictive models and clustering algorithms depends on the purpose of the analysis and the characteristics of the data. In the context of the real estate market, metrics such as RMSE, R^2 and PRD are essential for measuring the quality of predictions, while the Silhouette Score and Davies-Bouldin Index are useful for assessing the efficiency of market segmentation. In addition, for classification models, metrics such as accuracy and precision make it possible to measure the effectiveness of property categorization. The correct application of these metrics allows for a more robust and well-founded analysis, contributing to better strategic decisions in the real estate sector.

After conducting a thorough analysis of the studies, the RQs defined at the beginning of this work are answered.

RQ1: How do different ML approaches impact the efficiency of clustering and predictive models? Different models impact the efficiency of clustering and predictive algorithms depending on data characteristics and analysis objectives. For example, Regression Models like MLR helped estimate relationships between variables in various contexts [1]. Tree-based models such as Random Forest and XGBoost were widely used due to their ability to handle large volumes of data and improve prediction accuracy: Random Forest, a powerful ensemble method, aggregates multiple decision trees to enhance predictive accuracy [3]. In clustering, K-Means stood out for its efficiency in grouping data, appearing in 20% of the articles due to its effectiveness in partitioning data into distinct groups.

RQ2: How do supervised and unsupervised ML techniques complement each other in clustering and predictive modeling? Supervised and unsupervised techniques complement each other because clustering can be used as a preprocessing step before applying supervised predictive models. K-Means, for instance, can be employed to segment properties with similar characteristics before applying a regression model or neural network. Moreover, hybrid methods like PSO-Bagging-ANNs (Particle Swarm Optimization + Ensemble Learning) combine particle swarm optimization with an ensemble of neural networks to improve prediction accuracy [4]. This demonstrates that combining techniques can improve prediction effectiveness by leveraging both segmentation and supervised learning for better data analysis.

RQ3: How can ML algorithms be tailored to enhance both clustering and predictive performance? The adaptation of ML algorithms can improve both clustering and predictive performance by adjusting hyperparameters and selecting appropriate techniques. Regularization methods like LASSO, RIDGE, and ElasticNet were applied in 50%, 40% and 20%, respectively, of studies, highlighting their role in variable selection and reducing model complexity, helping to mitigate overfitting and enhance interpretability.

In predictive modeling, optimization strategies such as PSO-Bagging-ANNs (Particle Swarm Optimization + Ensemble Learning) combine Particle Swarm Optimization with an ensemble of Neural Networks to improve prediction accuracy [4]. Similarly, Genetic Algorithms (GA) are optimization techniques inspired by natural selection, often used to fine-tune parameters in ML models [6]. These methods help enhance model performance by optimizing hyperparameters and improving learning efficiency.

For clustering, approaches like SCMCA focused on geographic clustering to reveal spatial patterns [9]. These strategies collectively enable models to better fit the data and improve generalization capabilities.

RQ4: What are the key challenges in selecting and optimizing ML algorithms for clustering and prediction?

Selecting and optimizing ML algorithms for real estate analysis involves several intertwined challenges related to the complexity of the data, interpretability of models, and data quality. Real estate datasets often include a mixture of structured variables (such as price, area, or location) alongside unstructured inputs like textual descriptions, images, and user reviews, which add layers of complexity and require advanced techniques to analyze effectively [1, 4]. Additionally, some advanced ML models, while accurate, are difficult to interpret, posing a barrier to their practical adoption [4]. The evaluation of these predictive models relies on various performance metrics, with around 60% of studies favoring MAPE as the primary measure of accuracy. To overcome these challenges, methods such as sentiment analysis, Natural Language Processing

(NLP), and computer vision have been applied to extract and integrate richer information from diverse data types [4]. This shows that successful model optimization depends not only on choosing appropriate algorithms but also on tailoring them to the unique characteristics of the real estate data.

Building on these general challenges, several specific difficulties emerge in real estate market trend analysis and prediction, ranging from managing diverse and unstructured data to dealing with data availability, quality, and the inherent limitations of advanced predictive models. Furthermore, combining different analytical methods can add complexity, requiring careful handling. The main challenges and possible strategies to address them are outlined below:

1. Complexity of Data and Unstructured Variables

Real estate market analysis faces significant complexity due to the coexistence of structured data, such as price, area, or location, and unstructured inputs like textual descriptions, images, and user reviews [1]. Many critical variables, such as neighborhood perception or aesthetic features, are qualitative and difficult to quantify, often escaping traditional models [3]. NLP can help extract relevant information from inconsistent listing texts, while computer vision enables the evaluation of property condition through images [1, 3]. By combining numerical, textual, and visual data within multimodal models, predictive performance can be significantly enhanced [4], supporting more accurate valuations and tailored property recommendations.

2. Data Availability and Quality

The lack of a centralized infrastructure leads to fragmented and inconsistent data across real estate platforms, public records, and agencies [9]. Many datasets contain missing, outdated, or poorly recorded information, making accurate prediction difficult. Techniques like data imputation and integration platforms help address missing values and standardize disparate sources [9]. Furthermore, collaboration between public and private stakeholders could improve data reliability and ensure up-to-date and comprehensive datasets, especially when shared standards are adopted for information collection and categorization. Privacy regulations restrict direct access to individual transaction data, requiring the use of aggregated or complementary data sources to maintain data comprehensiveness and reliability [5].

3. Low Interpretability of Advanced Models

Despite their accuracy, advanced models such as deep neural networks, Random Forest, and XGBoost are often difficult to interpret [1, 2]. This limits their usability in practice, as real estate professionals need to understand the reasoning behind model predictions. To improve transparency, these models can be combined with more interpretable methods, such as logistic regression, or supported by Explainable AI (XAI) techniques [4]. Tools like SHAP and LIME allow for the identification of which variables influence specific predictions, providing clarity to end users [1].

4. Sensitivity to External Economic and Political Changes

The real estate market is heavily influenced by regional and socioeconomic factors such as income, demographics, and urban-rural differences. These variables reflect economic and social changes that models relying solely on historical data struggle to predict. Incorporating regional indicators improves model adaptability to external variations and enhances prediction accuracy [5].

5. Challenges in Integrating Clustering and Predictive Models

Combining clustering algorithms (like K-Means or DBSCAN) with predictive models presents computational challenges and risks compromising performance if parameters are poorly set [9]. For example, choosing the wrong number of clusters may result in data overgeneralization and reduced accuracy. Automated feature selection could help simplify model inputs, focusing only on the most relevant variables [1]. A promising solution lies in semi-supervised models, which combine unsupervised clustering with supervised learning, leveraging the strengths of both techniques for improved predictive performance [1].

6 Conclusion

This Literature review analyzed the use of clustering and predictive models in real estate, demonstrating their effectiveness in trend prediction and property segmentation. Combining clustering algorithms like K-Means, DBSCAN, and Hierarchical Clustering with predictive models such as Linear Regression, Random Forest, and Neural Networks enhanced prediction accuracy and categorization [1, 2, 3, 4]. Challenges included data complexity, unstructured variables, data quality, and interpretability of advanced ML models [1, 3, 4, 5, 9]. Emerging techniques like sentiment analysis, NLP, and computer vision could address these issues for deeper insights [1, 3, 4]. Future directions include integrating

multimodal models, real-time data systems, and AI-based early warning systems, which could transform market analysis and provide strategic support to stakeholders [1, 2, 4]. Overall, the synergy of clustering and predictive modeling marks a significant advance in real estate analysis, enabling more informed decisions and effective investment strategies.

This study has some limitations that should be acknowledged. The search string used, while precise, may have excluded relevant studies, for not using the root of some words instead of the words. This could have resulted in a smaller sample of articles than potentially available. Future research could benefit from broader search terms and more string variations to ensure a more comprehensive coverage of the existing literature on machine learning applications in real estate.

References

[1] C. Çılgın e H. Gökçen, «A Hybrid Machine Learning Model Architecture with Clustering Analysis and Stacking Ensemble for Real Estate Price Prediction», *Comput Econ*, ago. 2024, doi: 10.1007/s10614-024-10703-4.

[2] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, e P. Karadayı Ataş, «A Novel Hybrid House Price Prediction Model», *Comput Econ*, vol. 62, n.º 3, pp. 1215–1232, out. 2023, doi: 10.1007/s10614-022-10298-8.

[3] P. Furia e A. Khandare, «Real Estate Price Prediction Using Machine Learning Algorithms», em *Advanced Analytics and Deep Learning Models*, John Wiley & Sons, Ltd, 2022, pp. 19–32. doi: 10.1002/9781119792437.ch2.

[4] R. Naz, B. Jamil, e H. Ijaz, «Machine Learning, Deep Learning, and Hybrid Approaches in Real Estate Price Prediction: A Comprehensive Systematic Literature Review», *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, vol. 61, n.º 2, Art. n.º 2, jun. 2024, doi: 10.53560/PPASA(61-2)863.

[5] C. Kim e N. Bedorf, «Predicting housing sale prices in Germany by application of machine learning models and methods of data exploration», *Kwartalnik Nauk o Przedsiębiorstwie*, vol. 71, pp. 107–122, set. 2024, doi: 10.33119/KNoP.2024.71.1.7.

[6] Ping-Feng Pai e Wen-Chang Wang, «Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices». Acedido: 28 de março de 2025. [Online]. Disponível em: https://www.mdpi.com/2076-3417/10/17/5832

[7] M. Mubarak, A. Tahir, F. Waqar, I. Haneef, G. McArdle, M. Bertolotto, e M. Saeed, «A Map-Based Recommendation System and House Price Prediction Model for Real Estate». Acedido: 28 de março de 2025. [Online]. Disponível em: https://www.mdpi.com/2220-9964/11/3/178

[8] J. A. Rodriguez-Serrano, «Prototype-based learning for real estate valuation: a machine learning model that explains prices», *Ann Oper Res*, vol. 344, n.º 1, pp. 287–311, jan. 2025, doi: 10.1007/s10479-024-06273-1.

[9] S. Sisman e A. C. Aydinoglu, «Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis», *Land Use Policy*, vol. 119, p. 106167, ago. 2022, doi: 10.1016/j.landusepol.2022.106167.

[10] E. Hellsing e J. Klingberg, *It's a Match: Predicting Potential Buyers of Commercial Real Estate Using Machine Learning*. 2021. Acedido: 28 de março de 2025. [Online]. Disponível em: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-445229

[11] B. Kitchenham e P. Brereton, «A systematic review of systematic review process research in software engineering», *Information and Software Technology*, vol. 55, n.º 12, pp. 2049–2075, dez. 2013, doi: 10.1016/j.infsof.2013.07.010.

Chapter 20

Case Study - Implementation of the 7S Philosophy in a Food Industry

Contribution #43
Cite as:
João Pereira, Hugo Raposo, Jorge Raposo, Torres Farinha. Case Study - Implementation of the 7S Philosophy in a Food Industry. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-43,
  title ={Case Study - Implementation of the 7S
     Philosophy in a Food Industry},
  author={João Pereira and  Hugo Raposo and  Jorge
     Raposo and  Torres Farinha},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Implementation of the 7S Philosophy in a Food Industry

João Pereira[1,2*], Hugo Raposo[1,2], Jorge Raposo[1,2,3], Torres Farinha[1,2], J. Edmundo de-Almeida-e-Pais[1,2,4]

[1] Polytechnic University of Coimbra, Instituto Superior de Engenharia de Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.
[2] RCM[2+] Research Centre in Asset Management and System Engineering, Polytechnic University of Coimbra, Coimbra, 3030-199 Coimbra, Portugal
[3] Ci2 –Smart Cities Research Center, Polytechnic University of Tomar, Quinta do Contador | Estrada da Serra, 2300-313, Tomar, Portugal.
[4] CISE - Electromechatronic Systems Research Centre, University of Beira Interior, 6201-001 Covilhã, Portugal

*Corresponding author

* joaoandre1992dias@gmail.com; hugo.raposo@isec.pt; jorge.raposo@isec.pt; tfarinha@isec.pt; edmundo.pais@ubi.pt

Abstract

Due to the growing need for optimisation and efficiency of the production processes in the food industry, companies in this sector are looking to adopt methodologies that guarantee organisation, continuous improvement and waste reduction.

The implementation of management systems and the standardization of processes are fundamental elements to guarantee the competitiveness and sustainability of operations. Efficient management of finished products, as well as correct maintenance and organization of spaces, are crucial for the proper functioning of production activities. Thus, the application of Toyota Production System (TPS) tools emerges, with emphasis for the 5S and 7S philosophies, a methodology's that promotes the organization and the standardization of the work environment, allowing greater control over processes and, consequently, the reduction of errors and losses.

This study was developed with the aim of organisation of the production and storage areas, improving the flow and thus optimizing the production process of the company under study. Therefore, to assess the state of the organization of an industrial food company, periodic audits were carried out, using a Questionnaire Model, where the 5 and the 7 Senses were evaluated. For each Senses, questions were posed and evaluated according to criteria with a maximum weight of 10 points. The sum of all the criteria of the 5 Senses corresponds to a total of 250 values and the sum of all the criteria of the 7 Senses corresponds to a total of 380 values, making up 100% of the maximum score for each philosophy.

This case study aims to demonstrate that the application of TPS tools, especially 7S, can contribute significantly to the efficiency and organization of production, reinforcing the importance of continuous improvement practices in the food industry.

Keywords: Toyota Production System, TPS, 5S, 7S; Continuous Improvement; Food Industry

1 Introduction

1.1 Framework

The bakery and pastry industry represents a significant segment of the global food sector, with high energy

consumption, substantial raw material use, and considerable potential for waste generation [1]. As consumer demand shifts toward products produced under sustainable and safe conditions, companies in this sector face increasing pressure to align operational practices with the United Nations Sustainable Development Goals (SDGs) [2]. The SDGs adopted in 2015 as part of the 2030 Agenda for Sustainable Development offer a comprehensive framework for addressing environmental, social, and economic challenges, including responsible production (SDG 12), clean water and sanitation (SDG 6), affordable and clean energy (SDG 7), sustainable industrialization (SDG 9), and climate action (SDG 13) [3].

Lean manufacturing tools such as the 5S methodology have been widely recognized for their ability to enhance workplace efficiency, reduce waste, and foster continuous improvement [4,5]. Traditionally, 5S, Sort, Set in order, Shine, Standardize, Self Discipline has focused on improving organization and productivity. However, recent adaptations in food manufacturing settings have incorporated additional "senses" to address broader operational goals, including safety (ensuring the well-being of workers and product quality) and environmental sustainability (reducing resource consumption and environmental impact) [6].

In the bakery and pastry context, such expanded 5S frameworks can deliver tangible benefits. For example, energy-efficient ovens and optimized production layouts reduce operational costs and carbon emissions, directly contributing to SDG 7 and SDG 13 [7]. Similarly, rigorous hygiene and equipment organization minimize contamination risks, aligning with occupational safety requirements and enhancing food safety compliance [1]. Moreover, integrating sustainability into everyday operational routines can embed environmental responsibility into company culture, facilitating long-term alignment with global sustainability benchmarks [2].

This paper proposes a framework that connects the SDGs with sustainability initiatives in a bakery and pastry factory through the adapted 5S methodology, incorporating the additional senses of safety and environmental sustainability. The proposed framework provides a structured pathway for improving operational efficiency, enhancing workplace safety, and meeting international sustainability targets.

This article arises from the completion of an academic project in a company in the food sector. This project took place at São Silvestre Pastry Factory, Lda., a company that manufactures and sells bakery and pastry products, located in Lousã, Coimbra District. Its objectives were to implement Toyota Production System (TPS) tools, namely the 5S philosophy and subsequently the application of the 7S philosophy, a reformulation and update of the original "5S" philosophy.

Regarding the implementation of TPS tools, namely the 5S and 7S Philosophy, audits were carried out before the implementation of the two philosophies in order to verify which points should be changed by them. After these audits, new audits were carried out over a period of one year in order to verify the implementation of the respective philosophies.

First of all, it is important to briefly characterize the food industry in Portugal. In general, the food industry is within the scope of the processing industries, having very specific characteristics and whose objective is the production of food products. Based on the Brazilian Classification of Economic Activities (CAE) of the food industry, it can be subdivided into several branches of activity, as can be seen in Table 1.

Table 1 - CAE for the food and beverage industries

CAE	FOOD AND BEVERAGE INDUSTRIES
10	Food industries
101	Slaughter of animals, preparation and preservation of meat and meat products
102	Preparation and preservation of fish, crustaceans and molluscs
103	Preparation and preservation of fruits and vegetables
104	Production of animal and vegetable oils and fats
105	Dairy industries
106	Processing of cereals and legumes; manufacture of starches, starches and related products.

107	**Manufacture of bakery products and other flour-based products**
108	Manufacture of other food products
109	Animal feed manufacturing
11	Beverage industry

Recent studies show that the turnover of the Bakery and Pastry and Fruit and Vegetable Sector has been increasing successively, while the Dairy sector has been decreasing. Gross Value Added (GVA) is used to evaluate the result of productive activity over a given period. This value results from the difference between the value of production and the value of intermediate consumption. More recent studies show that the GVA of the Bakery and Confectionery sector has been increasing gradually, as all others experience variations over the years.

Therefore, the São Silvestre Pastry Factory, Lda., has a very acceptable turnover. Through graphical analysis (Figure 1), it is possible to observe that the company's revenue has been increasing significantly, reaching values of around 2,000,000 € in the last two years.

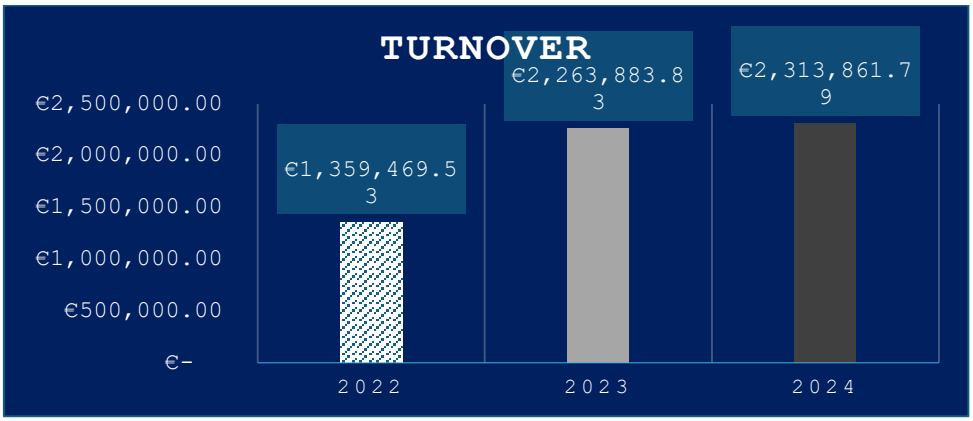


Figure 1 - Turnover of the São Silvestre Pastry Factory, Lda.

Gross Added Value (GVA) is the result of productive activity in a given period of time, which results from the difference between the value of production and the value of intermediate consumption, resulting in profit. By analysing Figure 2, it can be observed that the GVA has a small fluctuation in the three years analysed. In relation to the data presented by the author [8], it can be inferred that GVA values of Fábrica de Pastelaria São Silvestre, Lda. are stable.

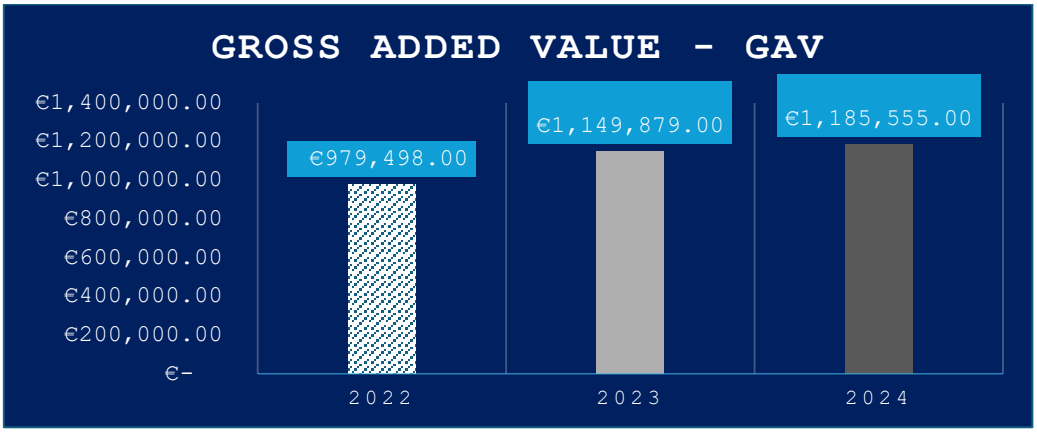


Figure 2 - GAB of the São Silvestre Pastry Factory, Lda.

During the course of an organization's activities, financial commitments may arise that must be met and achieved within a given period.

1.2 Methodology and Research Questions

The aim of this research is to create a methodology introducing two new senses, safety and sustainability. For this purpose, a five-step methodology was used:

a) Framing questions for a review

   i. Is it important for organizations to improve their safety and sustainability?

   ii. How can we measure organizational performance taking into account safety and sustainability?

   iii. How can we quantify safety and sustainability?

b) Identifying relevant work

   i. Section 2 provides a literature review.

c) Assessing the quality of studies

   i. Indexed papers from scientific libraries.

d) Summarizing the evidence

   i. Using available data to demonstrate the robustness of the proposed model through a case study.

e) Interpreting the findings

   i. The results obtained through the audits are discussed and the strongest and weakest aspects of the research are identified.

1.3 Paper Structure

This article is structured into the following sections:

- Section 1 presents the introduction, the objectives of the study and a brief characterization of the food industry in Portugal, methodology and research questions, finally the structure of the paper;
- Section 2 presents the theoretical framework and a description of all the concepts inherent to the TPS Methodology, Lean and its tools, as well as its foundations or the basis of the TPS house, the 5S philosophy and, based on this philosophy, a new and innovative concept was developed, the 7S;
- Section 3 – Presents the methodology adopted for the study, audits and implementation of the two philosophies 5S and 7S;
- Section 4 presents the Case Study where the models presented and developed in Section 2 are validated, in particular the implementation and audits of the 5S and 7S philosophies taking into account the TPS methodology;
- In Section 5 can be observed the results obtained from the audits and a brief discussion of them;
- In Section 6, the conclusions are presented, as well as the main contributions made in this study.

2 State of the art

In increasingly competitive markets, continuous improvement of industries prevails and must be increasingly valued. Nowadays, a large majority of companies implement tools from the TPS methodology to improve the organization of the factory in general and its production processes in particular. Applying TPS tools not only increases plant operational efficiency, but also improves customer experience, promoting company loyalty and repeat purchases [9].

TPS is the basis of Lean Manufacturing, a philosophy focused on eliminating waste and optimizing processes. In this way, Lean management also appeared and emerged as a strategic tool for increasing competitiveness and business performance [10].

The adoption of Lean management practices in the production of services, for example, in the context of after-sales services, has proven to be an effective strategy for eliminating waste, increasing operational efficiency and raising customer satisfaction levels. Also, in the maintenance and repair sector, such as the automotive sector, these principles are adapted to address the specific challenges of this context, promoting improvements in service quality and customer experience [10, 11].

Lean practices are focused on delivering value to the customer, which is reflected in more qualified maintenance services and better overall performance of production systems. Furthermore, the Lean philosophy helps to align maintenance activities with the organization's strategic objectives, ensuring that this area plays a relevant role in competitiveness and business success [10, 11].

A customer-centric approach is an essential element of Lean practices in after-sales services. Unlike traditional manufacturing, this sector involves the customer as an active participant in the process. Operational efficiency, in this case, depends considerably on compliance with appointments scheduled by customers, which requires clear, effective and targeted communication to keep the customer informed and committed to the services offered [10, 11].

Another significant advantage of the Lean approach is the ability to identify and resolve maintenance problems more quickly, improving response times and minimizing interruptions in production processes [10, 11].

Still within the scope of the TPS methodology, 5S tool stands out (separating what is useful from what is not used; organization; cleanliness; standardization; discipline) where the workplace is organized, facilitating the location of tools and materials. This organization reduces the time wasted searching for items or tools, as well as minimizes setup downtime and increasing productivity [12].

In summary, the TPS and Lean concept can be applied to these areas of an organization for its continuous improvement: process and equipment, production planning and control, maintenance management, human resources, product design, relationships with customers and suppliers.

2.1 The Seven Wastes

Waste can be defined as activities that are carried out and do not add value, making products and services more expensive and wasting unnecessary resources and time. This waste is divided into two categories:

- pure waste, which is waste that must be completely eliminated;
- necessary waste corresponds to the activities that must be carried out; however, they do not add any type of value, so they should be minimized.

Pinto [13] says that Ohno and Shingo, when developing the TPS, listed seven wastes. These seven wastes are: Transportation, Inventory, Waste of the process itself, Waiting, Unnecessary work, Excess production and Defects.

1º Inventory – "Accumulated stock for customer order"

Inventory (stock) is the raw materials or finished product that is retained, due to production being carried out in a larger quantity than the customer ordered. As a rule, inventories hide problems.

The most common causes for their appearance are: accepting them as normal, something that is part of the organization's assets; poor equipment layout: which leads to storage or transportation; long tool change times; existence of bottlenecks or bottlenecks in the processes; quality problems; and anticipation of production.

Eliminating this waste involves: reinforcing the planning and control of operations; regulation of the flow of operations; pull production; improvement of process quality; and quick tool change.

2º Waiting – "Postponing a task for some time"

Refers to the time people or equipment waste waiting for something.

The most common causes for this waste to appear are: obstructed flows (due to a breakdown; an accident); layout related issues; supplier problems and/or delays; and the balance between supply and demand.

The following methods can be applied to eliminate waste: making quick tool changes; improving planning and synchronization between work areas; balancing workstations; and opting for leveling operations (Heijunka).

3º Flaws – "Don't get it right the first time"

This type of waste includes defects and quality problems. Associated with this waste are the costs of inspection, responding to customer complaints and repairs.

The most common causes for its appearance are: transportation and movement of materials; human failures and errors; and the absence of self-control and inspection.

The methods used to eliminate waste are: standardization of operations; the presence of non-compliance detection devices; reduce the need to move parts and raw materials; and resort to process automation.

4º Overproduction – "Producing more than necessary"

It is the opposite of Just-in-Time (JIT) production, that is, continuing to produce even after the production order has been completed. Therefore, this is considered the most penalizing waste, contributing to the appearance of all the others.

Some of the causes of excess production or service are: the need to make the most of efforts made in activities that do not add value, such as transportation, inspections and setups; anticipation of production with the expectation of selling the product sooner; production of defective product; large batch production.

It is possible to apply methods to eliminate this type of waste, such as programming and standardizing the entire value chain; continuous flow (piece by piece); resort to pull production (JIT); levelling production (smaller batch production and flexible production - Heijunka); and resort to rapid tool change.

5º Unnecessary work – "Doing more than is necessary"

Refers to all operations and processes that are not actually necessary to perform operations, so they do not add value.

The most common causes for its appearance are: lack of employee training; incorrect layout; demotivation in people; instability in operations; undeveloped capabilities and skills and isolated operations.

To eliminate this waste, it is necessary to achieve a continuous flow of production/service; promote the standardization of work operations and invest in employee training and development.

6º Transportation – "Unnecessary movement of resources"

These are movements that raw materials and finished products undergo, and this not only waste time, but also damages products.

To reduce transport and movements, it is necessary to correct layouts, plan operations and choose smaller and faster transport systems.

The best way to reduce or eliminate this waste is: to have fluid and pulled production; there is operational flexibility; and use manufacturing cells.

7º Waste of the process itself – "Movement within the work area"

These are the operations and processes that are carried out and are not necessary. For example, an increase in defects may be a consequence of incorrect operations or processes. All processes generate losses that must be minimized or eliminated. This often happens due to a lack of work instructions, disorganization of the space, etc. In order to eliminate or reduce this waste, it is necessary to train employees and/or replace tools and processes with greater efficiency [13]

Pinto [13] also mentions that, in addition to the waste mentioned above, there are currently some more, such as:

- Waste of human potential – means not taking advantage of employees' capabilities, regardless of their role in the company;
- Waste from using an inappropriate system – means the incorrect use of existing technology in the company;
- Waste of resources – is the incorrect use of resources, such as water, electricity, gas, etc. This waste costs money and new standards must be created to ensure the sustainability of resources and the environment;
- Waste of materials – is the waste of material that could be used to produce other components at the end of its useful life;
- Wasting customer time – is when the customer is forced to wait longer than is actually necessary;
- Waste of human communication – this is when there is no effective communication within the company, and information may not be passed on correctly or even be lost.

2.2 TPS Tools and Techniques

There are a large number of TPS tools and techniques, such as:

- 5S and 7S Philosophy;
- Value Stream Mapping (VSM);
- Failure Mode and Effect Analysis (FMEA);
- Levelled programming;
- The Kanban control system;
- The cause and effect diagram;
- Standardized processes;

– The 5W2H formula (Who? What? Where? When? Why? How? How much?);

– SMED – Single Minute Exchange of Die;

– Among others.

In this project, the TPS Tool used is the 5S philosophy, which will be explained in the next sections, as well as the new 7S concept.

2.3 5S Philosophy

5S was developed as a philosophy of organization and standardization in the workplace. It emerged in Japan after World War II, when many Japanese companies were trying rebuild the economy and increase competitiveness in the global market. [14, 15, 16, 17].

The 5S Philosophy is an accessible and easy-to-understand management tool, allowing participation and involvement from all hierarchical levels of the company. This philosophy is an integral part of the TPS and was later disseminated by the Japan Industrial Standards Committee (JISC). Its aim is to organize work in a clean, efficient and safe manner in order to increase productivity and standardize work. The acronym 5S integrates five main Japanese words, each representing a fundamental principle of organization and efficiency (Figure 3):

- *SEIRI* – Organization (Separate the useful from the useless);
- *SEITON* – Tidying Up (A place for everything and everything in its place);
- *SEISON* – Cleaning (Clean the workplace and check for opportunities for improvement);
- *SEIKESTSU* – Normalization (Standardization);
- *SHITSUKE* – Self-discipline.

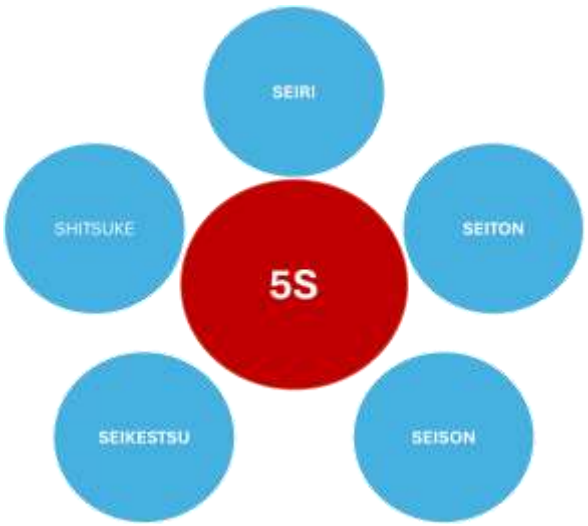


Figure 3 - 5S Philosophy

1º SENSE: *SEIRI* – Sense of Organization

The Seiri sense is the first step in the 5S methodology and consists of identifying the tools at the workplace (e.g. materials; equipment; utensils; etc.) distinguishing those that are really necessary from those that are not. The old habit of accumulating and keeping things because they might be useful does not contribute for cleanliness and the efficiency for searching for objects. This sense must identify not only waste, but also the associated causes; so, that preventive measures can be adopted to avoid its repeatability. In this step, it can be determined what should actually be kept in the workplace, what must be kept, and what should be thrown away.

Generally, the ABC classification system is used to identify the items that are really necessary:

- *A* – Daily use (items must be located at the workstation);
- *B* – Weekly or monthly use (items can be located next to the workstation, if there is space);
- *C* – Very rare use (items must be located outside the workstation, in a storage area).

It is often difficult to implement the method, as employees do not understand the need.

Some of the benefits of this sense are: rational use of the workplace, elimination of excess tools, reduction of time spent searching for tools and documents and reduction of stock and space costs.

2º SENSE: *SEITON* – Sense of Tidying Up

This step consists of organizing and identifying the workstation in order to make it functional, defining storage rules that allow any employee to find a tool as quickly as possible. So, this sense aims to organize and find the desired utensil in 30 seconds. This step allows the workplace to be more efficient, that is, it allows employees to always have what they are looking for at hand and makes them more productive. In Seiton is where, for example, can be painted the floor to check for dirt, to outline the tools found in a certain panel to make storage easier, etc. By applying this sense, they can be reached some benefits, like the fllowing: improved flows of people and materials, greater control of the workstation, greater efficiency when looking for a tool and purchasing unnecessary material.

3º SENSE: *SEISON* – Sense of Cleanliness

After the last steps, the third sense consists of cleaning the workstation and surrounding areas, that is, utensils, machines and other tools. This sense is not sequential to the two previous stages, but rather parallel. This means that each operator must clean each tool after using it, so that the next employee finds it clean. In this sense, the most important thing is not the act of cleaning, but the act of "not dirtying". In addition to cleaning, it is necessary to identify the source of dirt and its causes, in order to prevent them from occurring. To do this, it is necessary to define what needs to be cleaned, what means should be used, as well as the cleaning frequency. In a clean environment, if a non-conformity occurs, it is detected more easily and quickly. By implementing this sense, it is possible to obtain the following benefits: the elimination of sources of dirt, an improved image of the workplace and the preservation of equipment and tools. In short, a pleasant and clean work environment must be maintained to achieve greater productivity, health and safety in the workplace.

4º SENSO: *SEIKESTSU* – Sense of Normalization

The fourth sense consists of formalizing rules and defining standards with the help of employees. The participation of employees in this sense will facilitate the implementation and fulfilment of the assumptions established in the three previous senses. Standardization allows work to always be carried out in the same way and with the same tools and equipment, regardless of the employee; for this, it is

necessary that each equipment has instructions. Standardization also allows to know when cleaning and maintenance should be carried out and where it should be stored after use. This sense aims to prevent the return of old habits, and standardization is one of the most important principles of Lean and ISO 9001:2015 (Quality Management System). This sense brings with it benefits, such as: eliminating unsafe working conditions, using clean clothing, increasing the level of employee satisfaction with work and standardizing/disseminating the way of acting at the workplace.

5º SENSE: *SHITSUKE* – Sense of Self-Discipline

In this sense, control over the application of all the rules and decisions that were taken throughout all the previous senses is established, thus promoting continuous improvement (Kaizen). Whenever necessary, new changes must be modified and developed. To maintain the implemented method, it is necessary to carry out control audits regularly in order to evaluate the different positions. The objective of application control is for everyone to work autonomously, thus increasing productivity, quality and safety at the workplace, thus creating a pleasant daily work.

2.3.1 Implementation of the 5S methodology

It is very important to obtain evidence of the company's current situation, in the areas that need improvement. That said, you should meet with all employees in order to understand what improvement actions to take in their workplace, as you are better than anyone else to know where to intervene. As already mentioned, all employees must participate in the implementation of this method. Therefore, implementation goes through the following process [18]:

1. Motivate employees and senior management;
2. Train employees on the method;
3. Take stock of the facilities;
4. Define an experimental area;
5. Create a task management system;
6. Form the experimental working group;
7. Create a “5S dashboard”;
8. Start with the working group;
9. Implement the 5 senses;
10. Generalize to other areas.

The application of the 5S methodology is divided into two stages, which are raising to the appropriate level and maintaining the level reached. In the first stage, it involves the implementation of the first three senses and the second stage involves the implementation of the other two senses. The PDCA cycle must be part of the implementation of the 5S methodology, that is, it must be part of the organization's continuous improvement routine, that is, it must be incorporated into the organizational culture (sense of self-discipline).

2.3.2 Failure to Implement the 5S Methodology

The 5S methodology is a fundamental tool for any organization due to its simplicity, low implementation cost and achievement of short-term results. However, not all companies are able to achieve the desired results, as they limit themselves to carrying out small activities on the premises without making them habitual for employees. Some examples of failure are [19]:

– Lack of understanding of concepts (the method is understood as an organization and cleaning task rather than an educational process);

- Lack of a strategic plan (despite the simplicity of the method, it is important to have a person present who can check the activities carried out);
- Assume that the 5S methodology is standardized (i.e., it does not vary from company to company);
- Judge that the 5S methodology has a solution for everything (there must be a well-defined business structure, the company's technology must not be obsolete and salaries must be reasonable);
- Implement the 5S methodology to show to other entities (implementing the method just to show to customers or suppliers becomes a waste of time;
- Speed up the implementation of the methodology ("old habits" should not be changed drastically, this only makes it difficult for employees to implement it);
- Not drawing up a maintenance plan for the 5S method to maintain continuity (you should not just implement the method, it is important to continue the method and promote new improvement actions).

2.3.3 Other senses

Currently, due to the failure in implementing the 5S methodology, some authors have advocated the emergence of new senses, but there has been no consensus among all authors. The proposed senses are [19]:

- ***Shikari Yaro* (Sense of determination and unity)** – means committing senior management to the process of implementing the 5S method. It aims to improve interpersonal relationships and employee commitment to the desired results;
- ***Shido (*Sense of Education)** – This means training all employees in the concepts of the 5S methodology, that is, assessing needs, planning and executing the evaluation of the methodologies implemented by management;
- ***Setsuyaku* (Sense of economy)** – Means the reduction of waste, through the application of PDCA;
- ***Sekinin* (Sense of responsibility)** – means responsibility towards others, that is, each employee is responsible for their job;
- ***Seisan* (Sense of eliminating losses)** – means the involvement of all company employees to eliminate all types of losses;
- ***Sekinin Shakai* (sense of social responsibility)** – means empowering each employee who is not only responsible for their job, but also for society;
- ***Shitsukoku* (sense of persistence)** – This means not implementing 5S only on the eve of an audit or at its launch, that is, the 5S method must be part of the daily routine;
- ***Shukan* (Sense of habit)** – means making the 5S methodology a habit, independent of the workload;
- ***Shisei Rinri* (Sense of moral and ethical principles)** – means empowering employees, what they should or should not do, what is right or wrong.

2.4 7S Philosophy

It is important to highlight that companies are increasingly adopting these senses, and others that are essential for them. A sense that is widely used alongside all the others mentioned above is the sense of security. This sense is present in companies, as it is common to find emergency exits or blocked corridors,

equipment without safety protection, etc. By applying this concept, companies will not eliminate work accidents, only its application will allow them to be reduced. Through it, employees will be aware of the dangers and that the place is safe and will be able to focus solely on solving problems that exist in the workplace itself. Another sense that is also in fashion is the Sense of Sustainability.

Therefore, in this article we present a new concept: "7S Philosophy", which adds two senses to the 5S.

The 6S added a sixth principle:

6º SENSE: Safety

Accident prevention and worker safety. It includes measures such as correct use of PPE (Personal Protective Equipment), identification of risks and creation of safety protocols.

The 7S added a seventh principle:

7º SENSE: Sustainability

Efficient use of resources and reduction of environmental impact; encourages practices such as the rational use of energy, the reduction of waste and the proper disposal of waste.

Seven S (7S) philosophy adopted for the workplace organization by eliminating or reducing Muda, Mura and Muri. The 7S implementation consists of seven phases shown in Figure 4: Sort, Set in order, Shine, Standardize, Self Discipline, Safety and Sustainability.



Figure 4 - 7S Philosophy

Therefore, 7S is the extension of 5S with two additional S, namely Safety and Sustainability. Safety practices and safety training are helpful in reducing accidents in industry. The Sense of Sustainability encourages and motivates the correct use of resources, such as water, electricity, gas, etc. New standards must be created to ensure the sustainability of resources and the environment.

7S activities must be continuously monitored through regular audits and improved if there is deterioration. 7S improves the performance of organizations in terms of greater productivity, better quality, greater cleanliness and a healthier, safer and more sustainable work environment.

3 Methodology

To implement this study, a bibliographic search was carried out on platforms such as Google Scholar, B-on, SciELO, Science Direct, prioritizing studies/articles published in the last five years and considering the Scimago Journal Rank (SJR) classification. The search strategy was based on the use of keywords in Portuguese and English, including terms such as “TPS”, “Lean” and “5S”.

Before starting the study, data collection was carried out, identifying and analysing the jobs in the different production areas, using the following collection methods:

- Direct observation using auditing;
- Existing documentary analysis;
- Photographic record of before and after the audited areas.

In this way, data collection will be carried out on site, during audits at different points in the process. The data obtained will be processed and analysed with the help of Excel tools. The data will be presented visually, using graphs and tables, to facilitate its interpretation and analysis. Regarding the application of 5S and 7S, the methodology will consist of the following steps:

- Preparation of a checklist in digital format for audit assessment. The aim is that this tool can be used to evaluate the application of 5S and 7S, during the project and after the project;
- An initial audit will be carried out to assess the status and identify the degree of implementation of 5S and 7S on the factory floor where photographs of the current state will be collected. An awareness campaign will be carried out aimed at employees and those responsible, to clarify the importance of correctly implementing the systems and the benefits for the organization;
- Regular audits will be conducted to assess the implementation of each sense of 5S and 7S, in different areas of the factory. With each audit, progress on practices will be monitored and areas for improvement will be identified. In the final audit, photographs will be collected to demonstrate the effect of applying the measures.

4 Case Study – Implementation of the 5S and 7S Philosophy

In this section it can be observed the case study carried out in a sector of the food industry, namely the implementation on the factory floor of two philosophies: 5S and 7S. Some studies in the food industry sector are the following: [20, 21, 22].

4.1 Implementation of the 5S

In order to assess the state of the organization of the manufacturing part of Fábrica Pastelaria São Silvestre, Lda, at the time of the beginning of the implementation, and also to identify some aspects to be improved, an initial audit was carried out. For this purpose, a checklist was created (Figure 5), where the 5 senses were evaluated. For the census, questions were asked and evaluated according to criteria with a maximum weight of 10 points. The sum of all the criteria of the 5 senses corresponds to a total of 240 values, making up 100% of the maximum score. As we can see in the radar map of the first audit (Figure 6).

São Silvestre	5S Methodology Checklist											
	Date: 20.11.19	Location: Fáb. Pastelaria São Silvestre, Lda										
	Responsible for audit: João Pereira	Score										Observations
Category	Description	0	1	2	3	4	5	6	7	8	9	10	
Seiri	Are there only the items you need to perform your tasks on your desktop?					4							> Leaks have been found in the compressed air piping (Bakery and Dispatch) > Some hand labs are leaking water; > Check the maintenance plan, there is rust on some equipment; > Some furniture, equipment, utensils and everything that is really necessary needs to be identified; > There are record sheets, but they have not been filled in; > Replace the paper holders and soap dispensers with more economical ones.
	Are there no unused or non-compliant materials in the work area?					4							
	Is the desktop visually pleasing?								7				
	Are there necessary/relevant information (e.g. work instructions, log sheets, etc.) in the work area?					4							
	Are there any lack of information and/or outdated documents on the desktop?					4							
	Subtotal: "Separating the useful from the useless".	4,6											
Seiton	Are there no materials lying around in the work area (floor, corridor, etc.)?				3								> Pastry trolleys scattered throughout the factory area; > Lack of identification on raw materials and in storage spaces; > Shipping shelves are disorganized and taking up space for placing orders; > Disorganized order preparation support benches.
	Is there a properly identified place to put materials (e.g. spatulas, knives, sharps, etc.)?						5						
	Is the storage area identified?			2									
	Is there a place to put finished products and raw materials?								7				
	Subtotal: "A place for everything and everything in its place".	4,3											
Seisou	Is the work area clean?							6					> Some surfaces have accumulated dirt (e.g. the tops of cupboards); > The cleaning staff do not follow the methodologies described in the hygiene plan; > The cleaning staff do not keep hygiene/cleaning records; > The walls and ceilings have mold and some scaling.
	Is sanitization done according to a sanitization plan?						5						
	Are the equipment, utensils and tools clean and sanitized?						5						
	Are there cleaning check routines?				3								
	Are sanitizing materials available?										9		
	Subtotal: "Clean up the workplace and see if there are opportunities for improvement"	5,6											
Seiketsu	Are the equipment, utensils and tools in the right places?						5						> There are no working instructions for the equipment; > Some lockers show some deterioration.
	Are the hygiene and maintenance plans visible?									8			
	Are there working instructions for operating the equipment?	0											
	Are the working instructions for operating the equipment visible?	0											
	Do employees feel motivated?						5						
	Are the equipment in the audited area in good condition and hygiene?								7				
	Subtotal: "Make sure you don't fall back into old habits".	4,2											
Shitsuke	The last audit of method 5S was performed?		1										> Employees are unaware of the 5S methodology; > There is a lack of communication between departments.
	Do employees promote continuous improvement?		1										
	Do employees keep daily records?					4							
	Do employees follow the instructions/methodologies implemented?							6					
	Do employees follow the 5S methodology?	0											
	Subtotal: "Conscious use of resources".	2,4											
Total											21,0		
													Pontuação Total 250

Figure 5 - Checklist – 1st 5S Audit

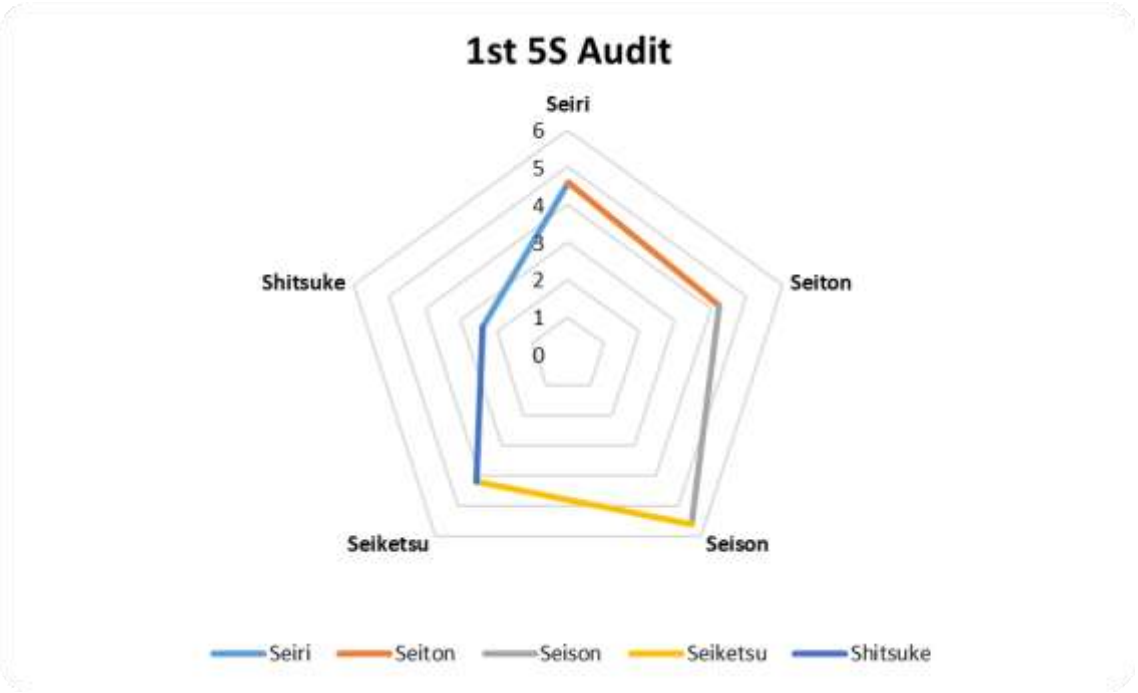


Figure 6 – Radar Map – 1st 5S Audit

In the first audit carried out, it was possible to verify some organizational problems, as can be seen in the Tables 2 to 7. With the help of the checklist, it was possible to verify the level of disorganization in which the company was at the manufacturing level, and this observation could be compared with the results of the first audit. That said, through the implemented methodologies, it can be observed a significant increase in the company's organization, since the result of the last audit was quite satisfactory. Tables 2 to 8 and Figure 7 show the results of the first audit, prior to the implementation of the 5S methodology.

Table 2 - Raw Materials Warehouse // Subsidiary Material Warehouse

Area:	**Raw Materials Warehouse // Subsidiary Material Warehouse**
	
Notes:	
– Raw materials and subsidiary materials are not all identified, in order to facilitate employees; – It is found that raw materials are ordered with low turnover, thus resulting in accumulated stock; – Especially the subsidiary material warehouse is disorganized; – It is necessary to create rules to avoid disorganization of the space; – In the raw material warehouse, it is possible to see that cleaning is not carried out daily;	

Table 3 - Pastry Area // Bakery

Area:	**Pastry Area // Bakery**



Notes:

- There is a water leak in the bakery sink;
- There is a compressed air leak in the bakery piping;
- In both areas it is possible to observe the lack of identification of equipment, shelves and it is not defined what to place in each of them;
- There are no work instructions for any equipment;
- There is an accumulation of dirt in some areas;
- During the work phase, tables are placed to obstruct the passage of some employees; they must be reorganized in such a way as to allow everyone to pass;
- It is necessary to carry out maintenance on some equipment, as rust is present.

Table 4 - Refrigeration Chambers // Freezing

Area:	**Refrigeration Chambers // Freezing**

Notes:

- It is possible to verify that there is some disorganization in the management of refrigerated raw material stocks;
- The products placed in the frozen chambers are all identified in order to ensure good stock management;
- There is some accumulated dirt due to humidity;

Table 5 - Heat Treatment Area // Shipping

Area:	**Heat Treatment Area // Shipping**






Notes:

- There is a leak in the compressed air pipe;
- The shipping shelves are disorganized and those that support manufacturing are not identified;
- The tables used to prepare orders are disorganized and therefore need to be identified;
- It can be seen that there are pastel carts spread throughout the area;
- There is some dirt present;
- The use of new rolls of film is verified, when there is still product to be used.

Table 6 - Social Areas

Area:	**Social Areas**

Notes:

- Some of the lockers are in a state of disrepair;
- Lack of organization in the changing rooms;
- The cafeteria is very disorganized (e.g.: dishes in various places)

Table 7 - General

Area:	**General**





Notes:

- Cleaning staff do not follow the methodologies described in the hygiene plan;

- Pastry/bakery trolleys (oven trolleys) spread throughout the factory;
- Cleaning staff do not keep hygiene records;
- There is a lack of communication between departments;
- Employees are not familiar with 5S methodologies;
- There are many damaged paper holders and soap dishes, it is recommended that they be re-placed with more efficient equipment (e.g. self-cutting paper and foam soap dishes);
- Disorganized administrative desks;
- Lack of a control record of customer, sales or visitor visits;
- Structure in need of intervention (e.g.: the roof of the bakery area needs general maintenance);
- The works must be completed as quickly as possible.

Therefore, the process of implementing the 5S method began in the production area, where objects and utensils that were already obsolete or that were not necessary for the task performed were removed (1st SENSE: SEIRI – Sense of Organization). It is important to note that there is poor communication between departments, as there are many communication failures regarding information given by the customer (e.g. complaints). It is worth noting that the employees have been with the company for a few years and are resistant to change. It was then necessary to make them understand the application of this sense, creating a pamphlet explaining the changes made with the application of this sense.

Table 8 - Comparison of before and after implementation

UTENSIL	LOCATION	
	Before Implementation	Post-Implementation
DOUGH CUTTERS	Scattered throughout the countertop drawers	Drawer of sanitized cutting tools
SHAPES (CUPCAKES; CORNUCOPIES, ETC..)	Several buckets (scattered throughout the oven area and confectionery area);	Form cabinet (paper, metal and sheets of paper);
LUBRICANT MATERIAL	Chef's Table	Eliminated
INVOICES/SHIPPING NOTES	Computer Table	Box placed on a shelf of the Computer Desk
PACKAGING (GOLDEN PLATES; LIDS; BASES)	Bakery area	Placed in the packaging cabinet

After completing the first sense, the implementation of the 5S method continued with the sense of tidiness (2nd SENSE: SEITON – Sense of Tidying). It was found that the warehouses were quite disorganized and needed some organization in order to help employees find what they were looking for in the shortest possible time. That said, the shelves were tidied up and the identification of raw materials and subsidiary material was done. In order to create better inventory management, raw materials were placed on the shelves, by supplier (Figure 7). Also, in the confectionery and bakery area, the work benches were identified. The pastry/bakery trolleys (oven trolleys) that were spread throughout the factory have been identified as the area where they should be stored so as not to obstruct the passage.



(A) (B) (C)

Figure 7 - (A) Organized Packaging Room; (B) Warehouse shelves organized by supplier; (C) Organized packaging support table

The sense of cleanliness (3rd SENSE: SEISON – Sense of Cleanliness) cannot be dissociated from the previous senses, as this one was applied as the shelves and cupboards were organized and tidied. With this sense, it was even possible to make the cleaners aware of the importance of keeping the production area clean and the need to fill out the spreadsheets. In this awareness, the importance of following hygiene plans and the safety measures that should be used when using cleaning products was also reinforced.

4th SENSE: SEIKESTSU – Sense of Standardization) according to this sense, rules must be described so that the work is always carried out in the same way and with the tools or equipment, regardless of the employee. That said, work instructions were created for the equipment to facilitate the operators' daily tasks. With this in mind, the equipment maintenance plan was also reviewed, as during the audits carried out, some non-conformities were found regarding the completion of the equipment registration forms.

5th SENSE: SHITSUKE – Sense of Self-Discipline, this sense consists of controlling all the other senses mentioned previously, in order to promote continuous improvement. To verify that there is continuous improvement, several audits were carried out, in order to improve all the less positive points until the final audit. It is important to emphasize that despite the 5S method having been implemented, and a final audit having taken place, it is necessary to continue carrying out audits in order to continue improving the negative aspects that arise over time. In this sense, the last audit carried out for this study obtained a final result of 85%, that is, a maximum score of 237.

4.1 Implementation of the 7S

At this point it is presented the application of the new 7S concept, a new checklist was created (Figure 8), where the 7 senses were evaluated. For each added sense (Safety and Sustainability), questions were asked and evaluated according to criteria with a maximum weight of 10 points. The sum of all the criteria of the 7 senses corresponds to a total of 380 values, making up 100% of the maximum score (Figure 9).

São Silvestre	5S Methodology Checklist											
	Date: 11.05.23	**Location:** Fáb. Panificação São Silvestre, Lda										
	Responsible for audit: João Pereira	**Score**										**Observations**
Category	**Description**	0	1	2	3	4	5	6	7	8	9	10	
Seiri	Are there only the items you need to perform your tasks on your desktop?						5						There is some obsolete material in the drawers; Employees sanitize and store some materials (e.g. buckets, bags) in the drawers that should be thrown away.
	Are there no unused or non-compliant materials in the work area?					4							
	Is the desktop visually pleasing?								7				
	Are there necessary/relevant information (e.g. work instructions, log sheets, etc.) in the work area?								7				
	Are there any lack of information and/or outdated documents on the desktop?									8			
	Subtotal: "Separating the useful from the useless".	**6,2**											
Seiton	Are there no materials lying around in the work area (floor, corridor, etc.)?						5						-
	Is there a properly identified place to put materials (e.g. spatulas, knives, sharps, etc.)?							6					
	Is the storage area identified?											10	
	Is there a place to put finished products and raw materials?											10	
	Subtotal: "A place for everything and everything in its place".	**8**											
Seiso	Is the work area clean?										9		-
	Is sanitization done according to a sanitization plan?											10	
	Are the equipment, utensils and tools clean and sanitized?									8			
	Are there cleaning check routines?									8			
	Are sanitizing materials available?										9		
	Subtotal: "Clean up the workplace and see if there are opportunities for improvement".	**9**											
Seiketsu	Are the equipment, utensils and tools in the right places?								7				-
	Are the hygiene and maintenance plans visible?								7				
	Are there working instructions for operating the equipment?								7				
	Are the working instructions for operating the equipment visible?								7				
	Do employees feel motivated?										9		
	Are the equipment in the audited area in good condition and hygiene?									8			
	Subtotal: "Make sure you don't fall back into old habits".	**7,5**											
Shitsuke	The last audit of method 5 S was performed?				3								The employees are familiar with the methodology, but it will be necessary to raise awareness again to remind them of some of the concepts.
	Do employees promote continuous improvement?						5						
	Do employees keep daily records?							6					
	Do employees follow the instructions/methodologies implemented?								7				
	Do employees follow the 5S methodology?						5						
	Subtotal: "Conscious use of resources".	**5,2**											
Security	Is the legislation in force complied with, respecting safety regulations and standards?									8			The company meets the requirements for ISO22000, but is not yet certified. It intends to start certification in 2025 and complete it in 2026.
	Do employees wear PPE?								7				
	Does the company use PPE?								7				
	Has the company implemented and complies with HACCP?											10	
	Does the company meet ISO 22001 requirements?									8			
	Does the company meet OHSAS 18001 requirements?									8			
	Does the company meet ISO 45001 requirements?							6					
	Do employees feel safe in the company?										9		
	Subtotal: "Ensuring everyone's safety".	**7,0**											
Sustainability	Are there no waste of raw materials or finished products?							6					There are returns of finished products because they are out of date. The company complies with some ISO14001 requirements, but it is not certified.
	There is no waste of water, energy or compressed air, etc.?								7				
	Recycling of recoverable components such as metals, plastics, electronic circuits and batteries?								7				
	Reuse of parts and accessories that can be reused in other equipment or sold for recovery?									8			
	Does the company meet ISO 14001 requirements?							6					
	Subtotal: "Ensuring sustainability".	**6,8**											
Total												50,6	

Figure 8 - Checklist – 1st 5S Audit



Figure 9 – Radar Map – 1st 7S Audit

It can be verified, through the checklist created for the 7S philosophy audit, that questions were added for the two new senses (Safety and Sustainability), through brainstorming meetings between the authors and collaborators of the new concept. The new questions for the Safety and Sustainability senses are presented below.

Safety:

- Is current legislation complied with, respecting safety regulations and standards?
- Do employees use PPE (Personal Protective Equipment)?
- Does the company use CPEs (Collective Protection Equipment)?
- Has the company implemented and complied with HACCP?
- Does the company meet ISO 22001 requirements?
- Does the company meet the OHSAS 18001 requirements?
- Does the company meet ISO 45001 requirements?
- Do employees feel safe in the company?

Sustainability:

- Is there no waste of raw materials or finished products?
- Is there no waste of water, energy or compressed air, etc.?
- Recycling of valuable components, such as metals, plastics, electronic circuits and batteries?
- Reuse of parts and accessories that can be reused in other equipment or sold for recovery?
- Does the company meet ISO 14001 requirements?

5 Results and Discussion

In this section, it will be presented the results and the discussion about the implementation of the 5S and 7S Philosophy, through the analysis of the audits carried out over time.

As previously stated, employees are highly resistant to change, but despite this, it was possible to increase the results obtained from implementing 5S from 42% to 95%. In Table 9 and according to Figure 10, it is possible to observe the evolution of the audits carried out from the first to the last audit carried out for the 5S philosophy implementation project.

Table 9 – Results of Audits in 5S Philosophy

Category	1st Audit	2nd Audit	3rd Audit	4 th Audit	5th Audit
Seiri	4,60	5,60	6,40	8,40	9,40
Seiton	4,25	5,70	6,75	8,25	9,25
Seison	5,60	7,20	8,00	9,00	10,00
Seiketsu	4,17	4,70	5,00	8,00	9,67
Shitsuke	2,40	3,80	4,60	7,80	9,00
Total	**21,02**	**27,00**	**30,75**	**41,45**	**47,32**

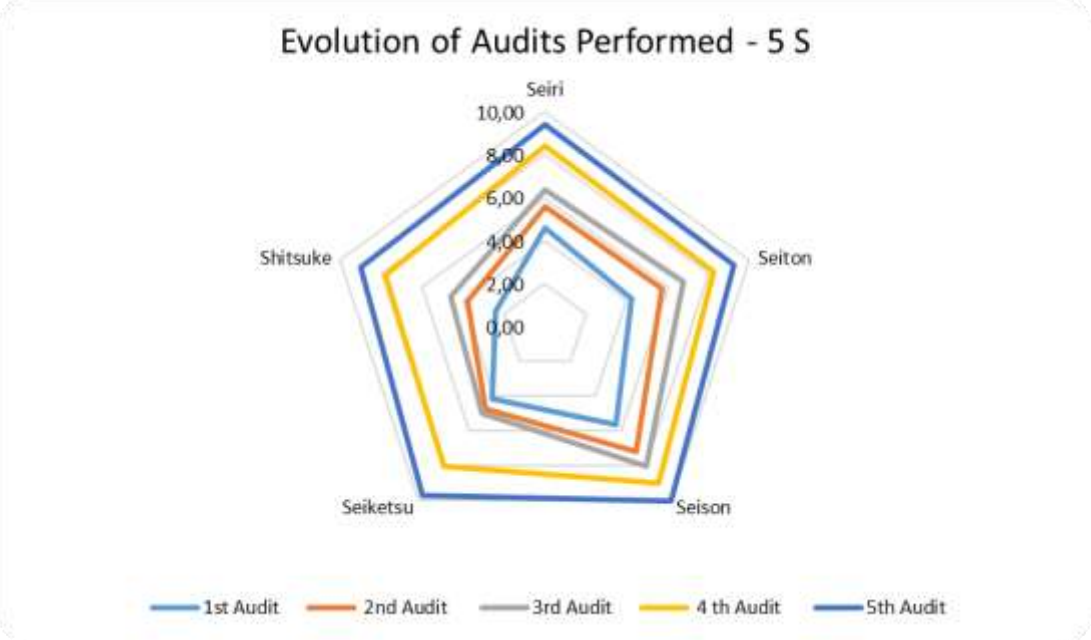


Figure 10 - Evolution of audits performed - 5 S

Throughout the audits, it was possible to observe a continuous improvement in the five senses. By analysing Figure 10, it is possible to verify that the 1st, 4th and 5th senses only showed a significant improvement after the third audit. This is due to resistance to change on the part of employees, the size of the team and the different schedules practiced, which makes it difficult to implement some new rules and self-discipline. As for the 2nd census, as would be expected in the first and second audit, there were no significant improvements, since, despite there being defined places to put the work tools, they often tend to place them in the old places. Regarding the 3rd sense, there was a gradual increase in the parameters evaluated, perhaps because the team was smaller and a little more open to change.

Below it is presented the first audit of the 7S implementation. In Table 10 and Figure 11, it is possible to see the results obtained in the audit carried out for the 7S philosophy implementation project.

Table 10 – Results of Audit in 7S Philosophy

Category	1st Audit	2nd Audit	3rd Audit	4th Audit	5th Audit
Seiri	**6,20**				
Seiton	**8,00**				
Seisou	**9,00**				
Seiketsu	**7,50**				
Shitsuke	**5,20**				
Security	**7,88**				
Sustainability	**6,80**				
Total	**50,58**				

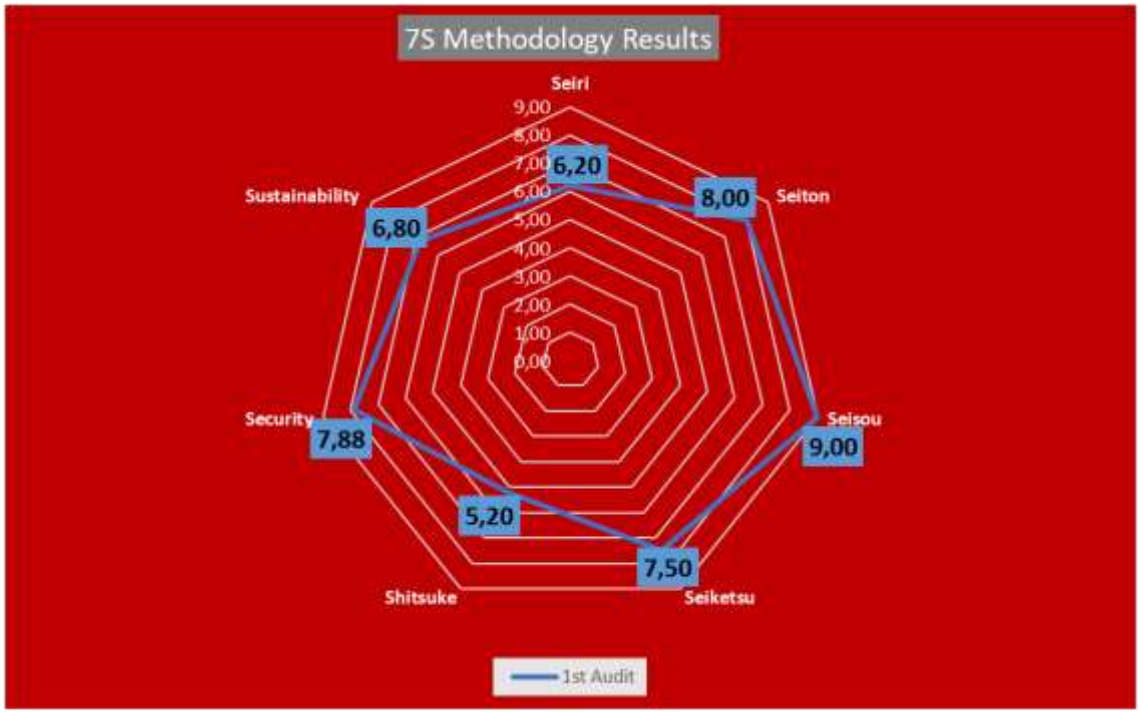


Figure 11 - Audit 7S philosophy

It can be seen from Table 10 and Figure 11 the results obtained in the first audit of the 7S philosophy, which were quite satisfactory, but it is important to do better. In this sense, now they can be addressed some points that can improve the performance of the 7S philosophy.

Below they are presented some points to improve the implementation of the 7S philosophy, namely in the last two senses: Safety and Sustainability.

6S - Safety: fewer errors, fewer injuries, fewer “run-overs”:

Safety is the condition of being protected against physical, social, spiritual, financial, political, emotional, occupational, psychological, educational, or other types or consequences of failure, damage, error, accident,

harm, or any other event that could be considered non-desirable.

Action Steps:

- Error Proofing (Poka-yoke);
- Safety related instructions & symbols;
- Alert, Warning, Hazard area identification and labelling with proper symbols;
- Safety trainings to employees.

Resources required:

- Safety related instruction and Symbols;
- Personal Protective Equipment (PPE);
- Collective Protection Equipment (EPC);
- Safety Trainer or Expert.

Target Outcomes:

- Avoid errors or mistakes;
- Reduces accidents;
- Safer working environment.

7S - Sustainability: efficient use of resources:

The "sense of sustainability" refers to the awareness and understanding of the importance of balancing human needs with the preservation of the environment and the guarantee of resources for future generations. It involves recognizing the finiteness of natural resources and adopting practices that minimize negative impacts on the planet, promoting development that is socially fair, economically viable and environmentally responsible.

Examples of environmentally sustainable measures are the following:

- Avoid wasting water;
- Use renewable and clean energy sources (geothermal, wind and hydraulic);
- Moderate the exploitation of mineral resources (coal, oil, minerals);
- Prioritize the production and consumption of organic foods;
- Prioritize the use of technologies that use renewable energy sources;
- Recycle;
- Prefer biodegradable products;
- Preserve green areas.

Examples of socially sustainable measures:

- Free educational and social projects, particularly for people with low incomes;
- Public and quality education;
- Social inclusion programs, aimed at people with special needs;
- Professional qualification of young people, through free courses;

- Basic sanitation, guaranteeing sewage and drinking water to all citizens;
- Wide access to electricity;
- Free or reduced-price Internet access for low-income people;
- Professional qualification of workers or unemployed people;
- Youth awareness programs on risk behaviours;
- Support programs for the neediest people.

Examples of economically sustainable measures:

- Use of clean and renewable energy, such as wind and solar;
- Treatment of organic waste, namely through recycling;
- Rational use of energy and water resources;
- More economical and less polluting transport of goods, such as rail and sea transport;
- Tax incentives for companies with sustainable practices;
- Water treatment and reuse systems;
- Use of recycled or biodegradable materials.

Finally, employees must be trained on the importance of maintaining the 7S philosophy in the company and its continuous improvement.

6 Conclusion

The present study was carried out at the São Silvestre Pastry Factory, Ltda. Through this work it was possible to implement TPS tools on the factory floor, notably the 5S philosophy and the 7S philosophy.

In this context and in relation to the implementation of the TPS tool, that is, the implementation of the 5S Philosophy, it was possible to verify that employees have some resistance to change; It has also been proven that it is necessary to provide them with training and information and give them attention, allowing them "make the change" themselves, which implies monitoring/guidance; so that, they feel the need and motivation to choose, for example, a place to put the cutting tools, without the person in charge being able to directly interfere in the choice of that same place. With this project it was also possible to see that, where teams are smaller, it is easier to encourage them to change.

Therefore, it is important to highlight that the results obtained were very satisfactory. It can even be said that the implementation of 5S was a success, although there is still much more to be done and changed at the São Silvestre Pastry Factory, that is, it is necessary to promote a spirit and philosophy of continuous improvement; It is necessary to maintain and continuously improve, always with the help and involvement of all employees.

The 7S philosophy is the extension of 5S with two additional senses, namely Safety and Sustainability. Safety practices and safety training are helpful in reducing accidents in industry. A sense of sustainability encourages and motivates team members to work and produce products sustainably. Thus, reducing the consumption of natural resources, in view of the current and global concern with the sustainability of the world.

The activities of the 7s must be continually monitored through audits over time and improved if there is degradation. 7S improves the performance of organizations in terms of greater productivity, better quality, fewer accidents, cleaner, safer, healthier and a more sustainable work environment.

Acknowledgments: When appropriate, include acknowledgments, such as funding information.

References

[1] Roldán, A., Pérez-Bosch, I., & Tello, M. (2025). Sustainable food production and the circular economy: Applying industry practices to reduce waste. Journal of Sustainable Food Systems, 7(1), 45–63. https://doi.org/10.1007/s44187-025-00461-z

[2] United Nations. (n.d.-a). Sustainable Development Goal 12: Ensure sustainable consumption and production patterns. Retrieved from https://sdgs.un.org/goals/goal12

[3] United Nations. (n.d.-b). Sustainable Development Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation. Retrieved from https://sdgs.un.org/goals/goal9

[4] Gapp, R., Fisher, R., & Kobayashi, K. (2008). Implementing 5S within a Japanese context: An integrated management system. Management Decision, 46(4), 565–579. https://doi.org/10.1108/00251740810865067

[5] Hirano, H. (1995). 5 Pillars of the visual workplace: The sourcebook for 5S implementation. Productivity Press.

[6] Sahoo, S., & Yadav, S. (2018). Lean manufacturing practices in Indian manufacturing SMEs and their effect on sustainability performance. Journal of Manufacturing Technology Management, 29(4), 733–762. https://doi.org/10.1108/JMTM-01-2017-0002

[7] AMF Bakery Systems. (2023). Nurturing sustainability in industrial baking: A comprehensive approach. Retrieved from https://amfbakery.com/news-insights/nurturing-sustainability-in-industrial-baking-a-comprehensive-approach/

[8] Gaspar, Pedro D., ELIAS, Miguel (2017) – Caracterização e análise dos processos produtivos em empresas agroalimentares. +Agro, mais estratégia, mais eficiência. ISBN:978-989-654-451-5.

[9] Raposo, H., Farinha, J.T., de-Almeida-e-Pais, J.E., Raposo, J. (2024). Life Cycle Management of Hospital Physical Assets Waste Elimination. In: Ball, A.D., Ouyang, H., Sinha, J.K., Wang, Z. (eds) Proceedings of the UNIfied Conference of DAMAS, IncoME and TEPEN Conferences (UNIfied 2023). TEPEN IncoME-V DAMAS 2023 2023 2023. Mechanisms and Machine Science, vol 151. Springer, Cham. https://doi.org/10.1007/978-3-031-49413-0_12.

[10] Arlinghaus, J. C., & Knizkov, S. (2020). Lean maintenance and repair implementation - A crosscase study of seven automotive service suppliers. Procedia CIRP, 93, 955–964. https://doi.org/10.1016/j.procir.2020.03.144.

[11] Mendes, D. S. F. T., Navas, H. V. G., Didelet, F., & Charrua-Santos, F. (2022). Improvement of maintenance management through Lean Philosophy and Industry 4.0. Maintenance, Reliability and Condition Monitoring, 2(1), 17–27. https://doi.org/10.21595/marc.2022.22472.

[12] Arslankaya, S., & Atay, H. (2015). Maintenance Management and Lean Manufacturing Practices in a Firm Which Produces Dairy Products. Procedia - Social and Behavioral Sciences, 207, 214–224. https://doi.org/10.1016/j.sbspro.2015.10.090.

[13] Pinto, João P. (2014) – Pensamento Lean. A filosofia das organizações vencedoras. Lisboa: 6a Edição, Lidel – Edições Técnicas Lda. ISBN: 978 989 975 203 27.

[14] Hayashi, N.: (2018). Toyota Production System and the roots of Lean (2018). https://vimeo.com/300443389. Accessed March 19, 2025.

[15] Kumar Saha, R., Minhaj Uddin, M., Mustafizur Rahman, M., Tutul Islam, M., & Mukshitul Mumin, M. (2023). Reduction of Changeover Time by Using the SMED Technique with the Assistance of Lean Manufacturing Tools in a Plastic Company. Proceedings of the 6th Industrial Engineering and Operations Management Bangladesh Conference Dhaka, Bangladesh, December 26-28, 2023. https://ieomsociety.org/proceedings/2023dhaka/30.pdf.

[16] Reke, E., Powell, D., & Mogos, M.F. (2022). Applying the fundamentals of TPS to realize a resilien and responsive manufacturing system. Procedia CIRP, 107, 1221–1225. https://doi.org/10.1016/j.procir.2022.05.135.

[17] Tahasin, T. A., Sen Gupta, H., & Tuli, N. T. (2021). Analyzing the Impact of 5S Implementation in the Manufacturing Department: A Case Study. Int. J. Res. Ind. Eng, 10(4), 286–294. https://doi.org/10.22105/riej.2021.229039.1230.

[18] Courtois, Alian, PILLET, Maurice, BONNEFOUS, Chantal Martin (2017) - Gestão da Produção. Lisboa: 5ª Edição, Lidel – Edições Técnicas, Lda. ISBN: 978-972-757-469-8.

[19] Sousa, Ana B., (2014) - Implementação do método 5S numa empresa de metalomecânica. Instituto Politécnico de Leiria, Escola Superior de Tecnologia e Gestão. Dissertação de Mestrado

[20] Santos, T., Silva, F. J. G., Ramos, S. F., Campilho, R. D. S. G., & Ferreira, L. P. (2019). Asset priority setting for maintenance management in the food industry. Procedia Manufacturing, 38, 1623–1633. https://doi.org/10.1016/j.promfg.2020.01.122.

[21] Veseli, A., Bajraktari, A., & Trajkovska Petkoska, A. (2024). The Implementation of Lean Manufacturing on Zero Waste Technologies in the Food Processing Industry: Insights from Food Processing Companies in Kosovo and North Macedonia. Sustainability (Switzerland), 16(14). https://doi.org/10.3390/su16146016.

[22] Pereira J, Raposo H, Farinha JT, de-Almeida-e-Pais JE (2023). Physical assets life cycle analysis in a food industry company - case study. Arch Food Nutr Sci. 2023; 7: 012-029. https://doi.org/10.29328/journal.afns.1001045.

Chapter 21

Trajectory Optimization for Collaborative Robots via the Deep Deterministic Policy Gradient Algorithm

Contribution #45
Cite as:
Zeinel Momynkulov, Azhar Tursynova, Olzhas Olzhayev, Akhanseri Ikramov, Sayat Ibrayev, Amandyk Tuleshov, Batyrkhan Omarov. Trajectory Optimization for Collaborative Robots via the Deep Deterministic Policy Gradient Algorithm. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-45,
  title ={Trajectory Optimization for Collaborative
     Robots via the Deep Deterministic Policy Gradient
     Algorithm},
  author={Zeinel Momynkulov and  Azhar Tursynova and
     Olzhas Olzhayev and  Akhanseri Ikramov and  Sayat
     Ibrayev and  Amandyk Tuleshov and  Batyrkhan
     Omarov},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Trajectory Optimization for Collaborative Robots via the Deep Deterministic Policy Gradient Algorithm

Zeinel Momynkulov[1,2], Azhar Tursynova[1,2], Olzhas Olzhayev[1,2], Akhanseri Ikramov[1], Sayat Ibrayev[1], Amandyk Tuleshov[1], Batyrkhan Omarov[1,2,*]

[1] Institute of Mechanics and Engineering Science, Almaty, Kazakhstan
[2] International Information Technology University, Almaty, Kazakhstan

*Corresponding author
batyahan@gmail.com

Abstract

This paper presents a deep reinforcement learning-based approach for trajectory optimization in collaborative robots using the Deep Deterministic Policy Gradient (DDPG) algorithm. The objective is to enable robots to generate smooth, efficient, and collision-free trajectories in continuous action spaces, particularly within dynamic and partially observable environments. The proposed method utilizes an actor-critic architecture, combining policy evaluation and action generation to optimize behavior through interaction with the environment. A carefully designed reward function incorporating distance to target, obstacle avoidance, and goal completion guides the learning process. Experiments were conducted in a simulated environment under two scenarios: single-goal navigation and multi-goal traversal. Over 200 training episodes, the agent demonstrated significant improvement in trajectory quality, progressing from random exploration to optimal path generation. Comparative analyses with classical methods such as A*, linear interpolation, and Bezier curves reveal that DDPG achieves competitive accuracy while offering superior adaptability and smoothness. Results confirm the effectiveness of DDPG in learning complex motion strategies without explicit path modeling. Although the algorithm showed promising outcomes, performance could be further enhanced with extended training and adaptive reward strategies. The findings suggest that DDPG is a viable and scalable solution for real-time robotic control, with potential for deployment in collaborative and industrial settings.

Keywords: DDPG; reinforcement learning; trajectory optimization; collaborative robots; continuous control; motion planning; obstacle avoidance; robot learning; real-time control

1 Introduction

The development of autonomous robotic systems capable of adaptive and efficient decision-making in dynamic environments has emerged as a critical objective in modern robotics research. Among the various challenges in this domain, trajectory optimization for collaborative robots remains a complex problem, especially when dealing with high-dimensional continuous action spaces and real-time environmental changes. Traditional trajectory planning methods, such as A* and spline interpolation, often fall short in dynamic scenarios due to their reliance on discretized state representations and static maps, which limits their capacity for adaptive control and smooth path generation [1].

Reinforcement learning (RL), particularly in its deep learning-augmented forms, has shown significant promise in addressing these limitations by enabling robots to learn control policies through environmental interaction [2]. One such technique, the Deep Deterministic Policy Gradient (DDPG) algorithm, extends deterministic policy gradient methods to continuous action domains, allowing for more precise motion control and policy convergence [3]. DDPG utilizes an actor-critic architecture, where the actor proposes actions and the critic evaluates them based on expected future rewards, ensuring efficient learning in high-dimensional control spaces [4].

Recent research has highlighted the advantages of DDPG in robotic motion planning and trajectory

generation. For instance, the integration of DDPG in robotic arms, mobile platforms, and industrial manipulators has demonstrated improved adaptability, smoother motion profiles, and better collision avoidance capabilities compared to classical algorithms [5]. Furthermore, the algorithm's stability is enhanced through the use of replay buffers and target networks, which mitigate the issues of correlated updates and high-variance learning typically encountered in RL [6].

This study investigates the application of the DDPG algorithm to optimize trajectories for collaborative robots. It focuses on constructing a reward function that balances trajectory smoothness, obstacle avoidance, and task completion, while leveraging deep reinforcement learning to achieve real-time, collision-free, and efficient motion planning in simulated environments.

2 Related Works

The field of robotic trajectory planning has experienced significant advances with the integration of deep reinforcement learning (DRL), particularly in continuous control tasks where traditional rule-based methods struggle. Conventional techniques such as linear interpolation, Bezier curves, and A* algorithms rely on predefined motion rules and often fail to generalize across dynamic or unstructured environments [7]. These limitations have motivated researchers to explore DRL algorithms capable of learning optimal behaviors directly from interaction with the environment.

One of the pivotal breakthroughs in this space is the application of the Deep Deterministic Policy Gradient (DDPG) algorithm, which operates effectively in continuous action spaces and enables fine-grained motion planning [8]. Unlike discrete action-based Q-learning methods, DDPG utilizes an actor-critic framework where the policy and value functions are jointly optimized, allowing robots to adapt to complex scenarios in real time [9-10]. The inclusion of target networks and experience replay further enhances the algorithm's stability, making it more viable for industrial applications where consistency and safety are critical [11].

In trajectory optimization, DDPG has demonstrated remarkable performance improvements in terms of motion smoothness and obstacle avoidance. For example, DRL-based methods have been employed to manage robotic arms performing high-precision assembly tasks, where DDPG enabled adaptive stiffness control for improved contact stability [12-13]. In another study, mobile robots equipped with DDPG learned to navigate cluttered environments, efficiently avoiding obstacles without the need for external supervision or global maps [14]. These capabilities are particularly important in collaborative settings where human-robot interaction demands high levels of autonomy and reliability.

Several works have further extended DDPG through architectural innovations. Hierarchical and memory-augmented variants of the algorithm have been proposed to enhance the temporal abstraction and contextual awareness of robotic agents [15]. In manufacturing contexts, DDPG has been integrated into fleet management systems for Automated Guided Vehicles (AGVs), demonstrating its utility in real-time task allocation and deadlock prevention [16]. Additionally, the algorithm has shown promise in robotic surgery, aerial drone navigation, and humanoid gait control, illustrating its broad applicability across diverse robotic platforms [17-19].

Despite these successes, challenges such as sample inefficiency, reward sparsity, and policy convergence remain open research problems. Recent efforts focus on addressing these limitations through reward shaping, imitation learning, and hybrid reinforcement learning frameworks that combine model-free and model-based techniques [20]. Collectively, these studies establish DDPG as a robust and adaptable tool for trajectory optimization in robotic systems, warranting further exploration and refinement.

3 Problem Statement

Trajectory optimization for collaborative robots operating in dynamic and unstructured environments remains a persistent challenge due to the need for real-time decision-making, continuous action control, and safety assurance. Classical methods such as graph-based planners or polynomial interpolation techniques are often limited by their reliance on static environmental models and lack of adaptability

when faced with unforeseen obstacles or moving targets. These methods generate suboptimal or infeasible trajectories in real-world scenarios, especially where high precision and responsiveness are required.

This research seeks to address the problem of generating optimal, collision-free trajectories for collaborative robotic agents using a learning-based framework. Specifically, we aim to leverage the Deep Deterministic Policy Gradient (DDPG) algorithm to learn a control policy $\pi(s_t \mid \theta^\pi)$ that maps observed states $s_t \in R^n$ to continuous control actions $a_t \in R^m$, optimizing the robot's motion through the environment.

The trajectory optimization task can be formulated as a Markov Decision Process (MDP) defined by the tuple (S, A, P, R, γ), where:

S is the continuous state space,

A is the continuous action space,

$P(s_{t+1} \mid s_t, a_t)$ denotes the transition probability function,

$R(s_t, a_t)$ is the reward function,

$\gamma \in [0,1]$ is the discount factor,

The objective is to maximize the expected cumulative reward:

$$J(\theta^\pi) = E_{s_t \sim p^\pi, a_t \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \quad (1)$$

subject to physical constraints of the robot and safety conditions of the environment.

The reward function $R(s_t, a_t)$ must encode key objectives, including proximity to the target, obstacle avoidance, and trajectory smoothness. A typical formulation integrates multiple components:

$$R(s_t, a_t) = -\alpha \cdot dist(s_t, s_g) - \beta \cdot coll(s_t) + \delta \cdot goal_reached(s_t) \quad (2)$$

Where

$dist(s_t, s_g)$ is the Euclidean distance to the goal s_g,

$coll(s_t)$ is a penalty for collision or proximity to obstacles,

$goal_reached(s_t)$ is a reward signal for reaching the goal,

α, β, γ are weighting coefficients,

This problem becomes more complex in multi-goal or time-varying environments, where the agent must generalize to new states without explicit reprogramming. Therefore, the central challenge is to train a policy that robustly generates smooth, safe, and efficient trajectories while minimizing learning time and computational overhead.

4 Materials and Methods

This section outlines the core components of the Deep Deterministic Policy Gradient (DDPG) algorithm as implemented in our trajectory optimization framework. The algorithm's structure, learning dynamics, and application to collaborative robot motion planning are detailed through dedicated subsections.

Q-function. The Q-function, or action-value function, plays a central role in evaluating the expected utility of an action a_t taken in state s_t under policy π. It is defined as:

$$Q^\pi(s_t, a_t) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k})\right] \quad (3)$$

Where $\gamma \in [0,1]$ is the discount factor, and R denotes the immediate reward at time t. The critic network approximates this function using parameterized weights θ^Q, allowing the agent to assess long-term consequences of its actions.

Actor-Critic Architecture. DDPG employs an actor-critic framework where two neural networks are trained concurrently:

The actor network $\pi(s_t \mid \theta^\pi)$ outputs the action a_t to be executed.

The critic network $Q(s_t, a_t \mid \theta^Q)$ estimates the Q-value of that action-state pair.

Policy parameters are optimized by maximizing the expected return $J(\theta^{\pi})$ using the deterministic policy gradient theorem:

$$\nabla_{\theta^{\pi}} \approx E_{s_t \sim D}\left[\nabla_a Q(s, a \mid \theta^Q)|_{a=\pi(s)} \nabla_{\theta^{\pi}} \pi(s \mid \theta^{\pi})\right] \quad (4)$$

This setup allows for continuous, smooth control updates across high-dimensional state-action spaces.

Target Networks. To stabilize training, DDPG maintains target networks π' and Q', which are time-delayed copies of the actor and critic. These are updated softly:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \quad \tau << 1 \quad (5)$$

This decouples the rapidly changing Q-value estimates from the target computation, thus preventing training oscillations and divergence.

DDPG Algorithm. The training algorithm proceeds as follows:

Algorithm: Deep Deterministic Policy Gradient (DDPG) for Trajectory Optimization in Collaborative Robots

---

Initialize actor $\pi(s \mid \theta^{\pi})$, critic $Q(s, a \mid \theta^Q)$, and target networks.

Initialize replay buffer D to store past experiences.

At each time step:

Select action $a_t = \pi(s_t \mid \theta^{\pi}) + N_t$, where N_t is noise.

Execute action, observe reward r_t and new state s_{t+1}

Store transition (s_t, a_t, r_t, s_{t+1}) in D:

Sample mini-batch from D, compute target:

$$y_t = r_t + \gamma Q'(s_{t+1}, \pi'(s_{t+1} \mid \theta^{\pi'}) \mid \theta^{Q'}) \quad (6)$$

Update critic by minimizing:

$$L = \frac{1}{N} \sum \left(Q(s_t, a_t \mid \theta^Q) - y_t\right)^2 \quad (7)$$

Update actor using gradient ascent on the expected return.

Soft-update the target networks.

End

Application of DDPG to the Problem of Trajectory Generation. In the robotic context, the agent's goal is to generate a path from an initial state s_0 to a goal state s_g, avoiding obstacles and minimizing trajectory cost. The state transition is defined as:

$$s_{t+1} = s_t + \Delta a_t = s_t + (\Delta x_t, \Delta y_t, \Delta z_t) \quad (8)$$

Where $\Delta a_t \in [-1,1]^3$ are continuous control outputs representing Cartesian displacement. The policy must generalize to environments with varying obstacle configurations, making DDPG well-suited due to its ability to learn smooth mappings under such complexity.

Reward Function. The reward function is carefully constructed to incorporate multiple objectives:

$$R(s_t, a_t) = -\alpha \cdot d(s_t, s_g) - \beta \cdot I_{collision}(s_t) + \delta \cdot I_{goal}(s_t) \quad (9)$$

$d(s_t, s_g)$: Euclidean distance to the goal,

$I_{collision}(s_t)$: collision indicator function,

$I_{goal}(s_t)$: indicator if goal is reached.

The α, β, δ parameters weight trajectory efficiency, safety, and completion respectively. The reward shaping encourages the agent to approach the goal while penalizing proximity to obstacles.

Action and Observation Spaces. The observation space comprises the current position and relative distance to the goal:

$$s_t = \left[x_t, y_t, z_t, x_g - x_t, y_g - y_t, z_g - z_t\right] \tag{10}$$

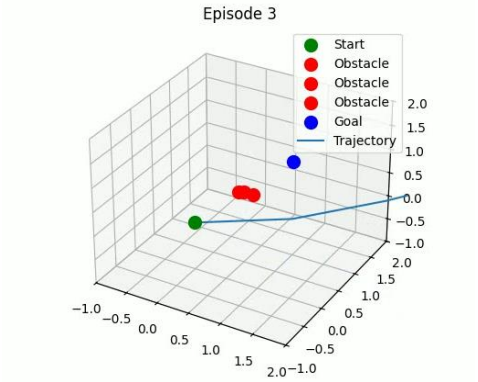
The action space consists of bounded displacement vectors in 3D space:

$$\alpha_t = \left[\Delta x_t, \Delta y_t, \Delta z_t\right] \in \left[-1,1\right]^3 \tag{11}$$

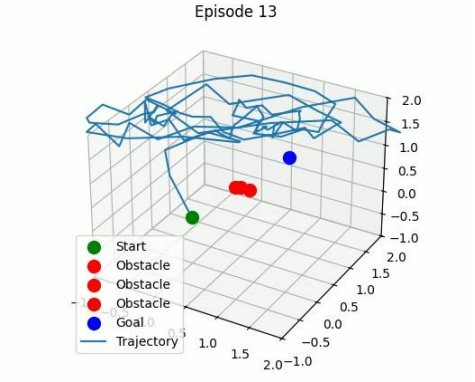
This design enables the agent to continuously navigate toward the goal while considering spatial relationships and dynamic constraints in a fully continuous control setting.

5 Results

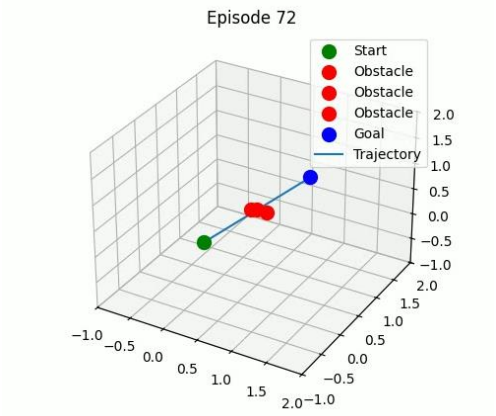
The training procedure was conducted in two distinct stages to evaluate the learning dynamics of the agent under controlled conditions. In the first experiment, the agent was assigned a single objective: to reach a fixed target while avoiding multiple static obstacles within the environment. To ensure computational representativeness without excessive training overhead, the agent was trained for a total of 200 episodes. The simulation and training were executed on an NVIDIA RTX 4070 GPU utilizing CUDA acceleration, with the complete training process. Throughout this stage, the agent relied solely on feedback from the reward function to explore the environment and iteratively improve its policy.



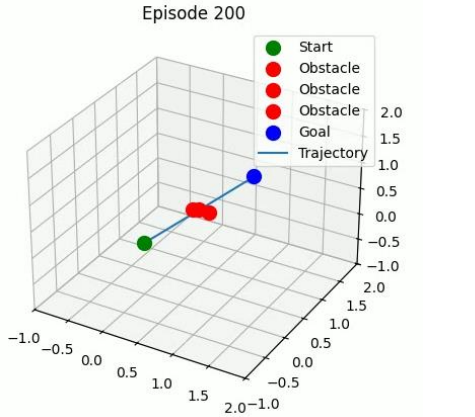
a) Agent trajectory in episode 3. No suffucient results achieved at this stage



b) Agent trajectory in episode 13. The agent starts to understand of approximate location of required points



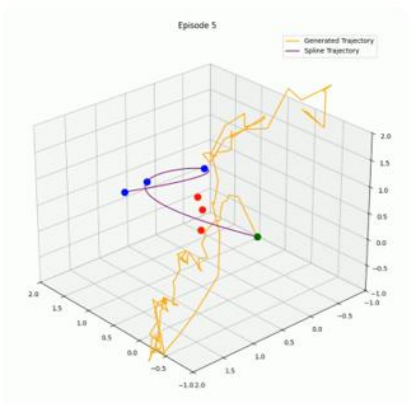
c) Agent trajectory in episode 72. First time agent reached the goal.



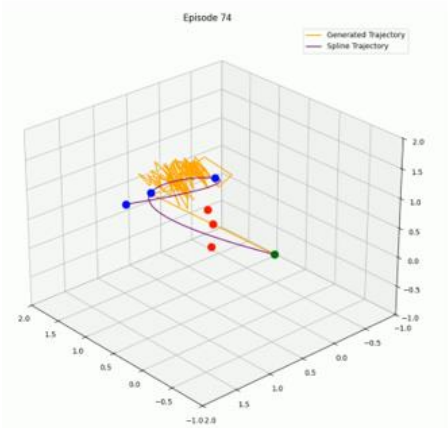
d) Agent trajectory in episode 200. At termination episode agent continues to straight and optimal line

Figure 1: Process of agent's learning (Experiment 1)

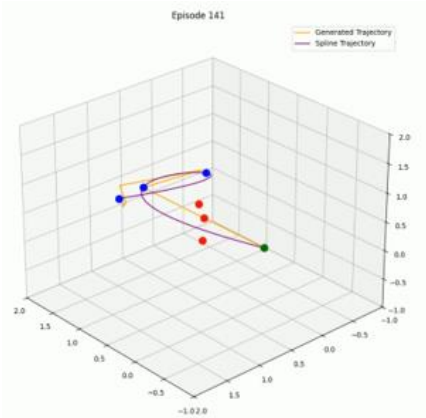
Figure 1 illustrates the evolution of the agent's behavior across selected episodes, highlighting its progressive adaptation. In the early stages (Episode 3), the agent demonstrates no meaningful navigation capability, failing to approach the goal and producing disorganized trajectories. By Episode 13, however, the agent begins to approximate the target's location, suggesting initial policy convergence. Notably, by Episode 72, the agent successfully reaches the goal for the first time, showing clear evidence of obstacle avoidance and path optimization. Finally, in Episode 200, the trajectory becomes nearly linear and efficient, indicating that the agent has learned to generate a smooth and direct path to the target. This gradual improvement confirms the efficacy of the Deep Deterministic Policy Gradient (DDPG) algorithm in learning optimal trajectories under sparse rewards and continuous control settings.



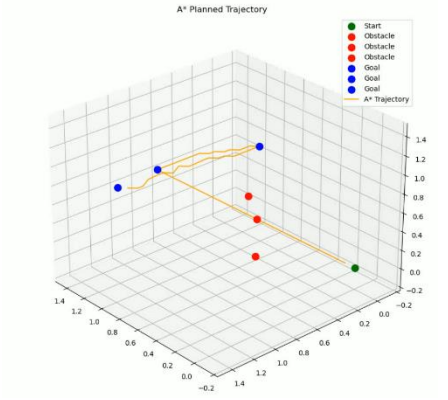
a) Agent trajectory in episode 5. Agent tries to find positive rewards



b) Agent trajectory in episode 74. Agent reaches the first target points and continues exploring



c) Agent trajectory in episode 141. Agent fully avoids the obstacles and reaches all target points.



d) Example of comparison with A* (A-star) path finding algorithm

Figure 2: Process of agent's learning (Experiment 2)

In the second experiment, the task complexity was heightened by requiring the agent to sequentially navigate through multiple target points rather than a single destination. To accommodate this spatial increase in trajectory depth, the visualization angle in Figure 2 was adjusted, offering a clearer perspective of the agent's evolving behavior. Early in training, as shown in Episode 5, the agent's trajectory is erratic, indicating exploration behavior in search of rewarding states. By Episode 74, it begins to identify the initial targets, showing partial mastery over the multi-goal path. Progress becomes more evident in Episode 141, where the agent successfully avoids obstacles and follows a trajectory that intersects all required goals. By Episode 200, the agent consistently executes an efficient and obstacle-free path, demonstrating policy stability and effective generalization. However, it is evident that increasing the number of training episodes and fine-tuning the reward function could

further refine the agent's performance and trajectory consistency.
When benchmarked against classical planning methods, such as A* or linear interpolation, the DDPG-based approach exhibits clear advantages, especially in dynamic and continuous action spaces. Unlike A*, which operates in discretized spaces and relies heavily on a complete environmental map, DDPG learns policies through interaction, enabling the agent to adapt to unknown or changing conditions. Similarly, linear interpolation generates rigid, static paths that fail to consider environmental constraints or obstacles. In contrast, DDPG optimizes trajectories using a learned value function that incorporates cumulative reward, allowing for dynamic trade-offs between safety, efficiency, and smooth motion. Furthermore, DDPG-trained agents can generalize across unseen scenarios after training, eliminating the need for manual recalculations and making the approach suitable for real-time applications in collaborative robotics. The comparison in Figure 2d underscores this difference, highlighting the flexibility and adaptability of learned policies over rule-based strategies.

Table 1: Comparison of Algorithms

Method	MAE	Max Error
Spline	0.0	0.0
Linear	0.386	0.808
Bezier	0.487	0.771
A*	0.370	0.777
DDPG	0.389	0.796

Table 1 presents a quantitative comparison of five trajectory planning methods—Spline, Linear Interpolation, Bezier Curve, A*, and Deep Deterministic Policy Gradient (DDPG)—based on two key performance metrics: Mean Absolute Error (MAE) and Maximum Error. As expected, the spline method serves as the reference trajectory with zero error, representing the ideal path. Linear interpolation demonstrates moderate accuracy (MAE = 0.386) but incurs the highest maximum error (0.808), indicating poor handling of curvature and obstacle constraints. The Bezier curve, while producing visually smooth trajectories, exhibits the highest MAE (0.487), reflecting its limited precision in adhering to multiple target constraints. A* performs slightly better with a lower MAE of 0.370, but still suffers from a relatively high maximum error (0.777) due to its discrete nature. The DDPG algorithm achieves a balance between accuracy and adaptability, with an MAE of 0.389 and a max error of 0.796, showcasing competitive performance against A* while offering the advantage of continuous, learned control. These results highlight DDPG's potential as a viable alternative to classical methods in dynamic and partially observable environments.

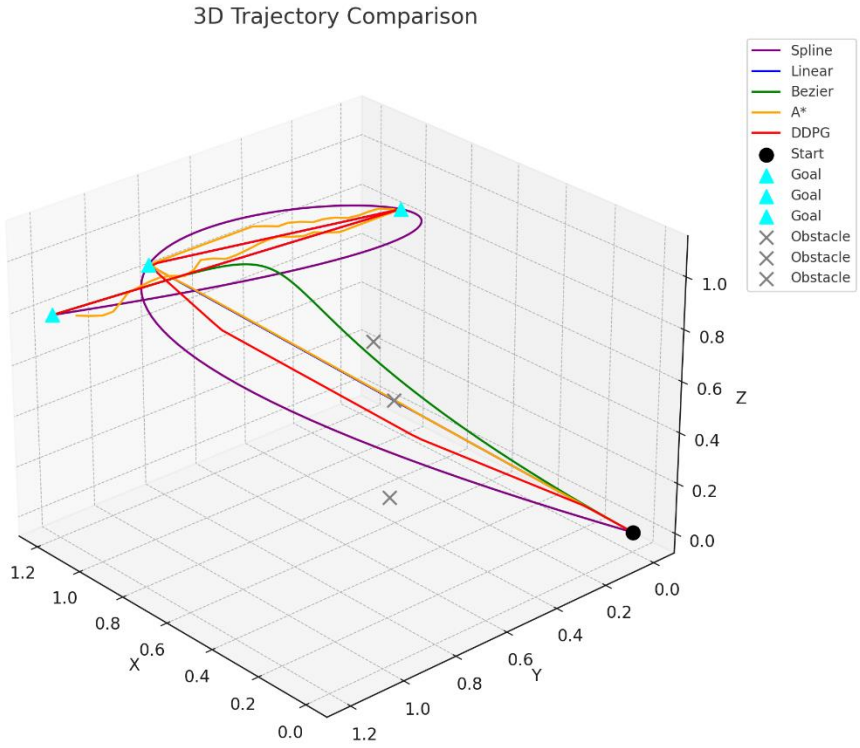


Figure 3: Visualization of different trajectories

Figure 3 provides a comparative visualization of the trajectories generated by different planning algorithms, including Spline, Linear Interpolation, Bezier Curve, A*, and Deep Deterministic Policy Gradient (DDPG). The spline trajectory, serving as the reference, exhibits an ideal, smooth path that accurately passes through all target points while avoiding obstacles. In contrast, the linear interpolation path connects waypoints directly with straight segments, resulting in abrupt directional changes and limited adaptability to environmental constraints. The Bezier curve produces a smooth but less precise trajectory, often deviating from critical waypoints due to its inherent lack of control over intermediate points. The A* algorithm, while capable of obstacle avoidance, displays discrete and somewhat rigid path characteristics, which may not be suitable for systems requiring fluid motion. The DDPG-based trajectory demonstrates a more adaptive and continuous path, effectively balancing smoothness and goal-directed behavior. Its learned policy enables the agent to navigate around obstacles and generalize across unseen configurations, emphasizing its potential in real-time robotic applications that demand both flexibility and precision.

6 Discussion

The experimental results demonstrate the efficacy of the Deep Deterministic Policy Gradient (DDPG) algorithm for optimizing trajectories in collaborative robotic systems operating within continuous and dynamic environments. Throughout both single-goal and multi-goal scenarios, the agent progressively learned to navigate towards its objectives while avoiding obstacles, reflecting the capacity of DDPG to converge toward optimal control policies under sparse rewards. The learning behavior, visualized over successive episodes, revealed a significant transformation—from initial exploration to smooth and goal-directed motion—indicating the effectiveness of the actor-critic architecture and reward shaping strategy.

Compared to traditional trajectory planning techniques, DDPG exhibited several compelling advantages. While methods like A* and linear interpolation offer deterministic paths, they lack the adaptability required for real-time robotic applications in environments with evolving constraints. In contrast, the DDPG-trained agent operated in a continuous action space, enabling it to respond fluidly to changes and generate trajectories that were not only efficient but also dynamically feasible. The inclusion of reward components for distance minimization, obstacle avoidance, and goal acquisition proved instrumental in guiding the agent's behavior without the need for explicit path engineering.

Nonetheless, several limitations remain. Although DDPG achieved competitive accuracy compared to A*, its slightly higher error margins may be attributed to the relatively short training horizon of 200 episodes and the fixed reward weights. This suggests that further tuning—either by extending the training duration or incorporating adaptive reward strategies—could yield improved performance. Additionally, integrating domain-specific knowledge or safety constraints into the learning process could enhance robustness, especially in high-risk industrial environments.

Overall, the findings affirm that deep reinforcement learning, and specifically DDPG, offers a powerful and flexible solution for trajectory planning in robotics. It opens pathways for further research into hybrid models, transfer learning, and deployment on physical platforms for real-time control in collaborative settings.

7 Conclusion

This study has presented an effective approach to trajectory optimization for collaborative robots using the Deep Deterministic Policy Gradient (DDPG) algorithm. By leveraging an actor-critic architecture and continuous control strategies, the agent demonstrated the ability to generate smooth, goal-directed, and obstacle-avoiding trajectories in both single-goal and multi-goal environments. Through iterative learning and interaction with the environment, the DDPG-based policy converged toward optimal motion paths, as evidenced by progressive improvements in trajectory quality and task completion rates across training episodes. Comparative evaluations against classical methods such as A*, linear interpolation, and Bezier curves revealed that DDPG not only achieves comparable accuracy but also provides enhanced adaptability in dynamic scenarios. While classical planners require explicit modeling and reconfiguration when the environment changes, the learned DDPG policy generalizes to novel conditions without manual intervention, making it well-suited for real-time robotic applications. Despite promising results, the algorithm's performance could be further enhanced by extending training epochs, refining reward functions, and incorporating domain-specific constraints to improve policy robustness. Future work will focus on real-world deployment with physical robotic platforms, enabling online adaptation, safety assurance, and integration into collaborative human-robot workspaces. Overall, the study affirms the potential of deep reinforcement learning in advancing intelligent and flexible robot motion planning.

Acknowledgments: This study was funded by the research project "BR24992947—Development of Robots, Scientific, Technical, and Software for Flexible Robotization and Industrial Automation (RPA) in Automotive Industrial Enterprises in Kazakhstan Using Artificial Intelligence".

References

[1] Gao, X., Yan, L., Li, Z., Wang, G., & Chen, I. M. (2023). Improved deep deterministic policy gradient for dynamic obstacle avoidance of mobile robot. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 53(6), 3675-3682.

[2] Wang, Z., Jin, X., Zhang, T., Li, J., Yu, D., Cheong, K. H., & Chen, C. P. (2022). Expert system-based multiagent deep deterministic policy gradient for swarm robot decision making. IEEE Transactions on Cybernetics, 54(3), 1614-1624.

[3] Li, Z., Guan, X., Liu, C., Li, D., He, L., Cao, Y., & Long, Y. (2025). Active Disturbance Rejection Control Based on Twin-Delayed Deep Deterministic Policy Gradient for an Exoskeleton. Journal of Bionic Engineering, 1-20.

[4] Kuo, P. H., Hu, J., Lin, S. T., & Hsu, P. W. (2022). Fuzzy deep deterministic policy gradient-based motion controller for humanoid robot. International Journal of Fuzzy Systems, 24(5), 2476-2492.

[5] Garg, A., & Jha, S. S. (2023). Deep deterministic policy gradient based multi-UAV control for moving convoy tracking. Engineering Applications of Artificial Intelligence, 126, 107099.

[6] Liu, J., Yap, H. J., & Khairuddin, A. S. M. (2025). Path Planning for the Robotic Manipulator in Dynamic Environments Based on a Deep Reinforcement Learning Method. Journal of Intelligent & Robotic Systems, 111(1), 1-15.

[7] Dong, R., Du, J., Liu, Y., Heidari, A. A., & Chen, H. (2023). An enhanced deep deterministic policy gradient algorithm for intelligent control of robotic arms. Frontiers in Neuroinformatics, 17, 1096053.

[8] Li, X., Xiao, J., Zhao, W., Liu, H., & Wang, G. (2022). Multiple peg-in-hole compliant assembly based on a learning-accelerated deep deterministic policy gradient strategy. Industrial Robot: the international journal of robotics research and application, 49(1), 54-64.

[9] Zhao, B., Wu, Y., Wu, C., & Sun, R. (2025). Deep reinforcement learning trajectory planning for robotic manipulator based on simulation-efficient training. Scientific Reports, 15(1), 8286.

[10] Zanchettin, A. M., Messeri, C., Cristantielli, D., & Rocco, P. (2022). Trajectory optimisation in collaborative robotics based on simulations and genetic algorithms. International Journal of Intelligent Robotics and Applications, 6(4), 707-723.

[11] Fan, J., Zhang, X., Zheng, K., Zou, Y., & Zhou, N. (2024). Hierarchical path planner combining probabilistic roadmap and deep deterministic policy gradient for unmanned ground vehicles with non-holonomic constraints. Journal of the Franklin Institute, 361(8), 106821.

[12] Farag, W. (2023). Robot arm navigation using deep deterministic policy gradient algorithms. Journal of Experimental & Theoretical Artificial Intelligence, 35(5), 617-627.

[13] Baltes, J., Akbar, I., & Saeedvand, S. (2025). Cooperative dual-actor proximal policy optimization algorithm for multi-robot complex control task. Advanced Engineering Informatics, 63, 102960.

[14] Mosali, N. A., Shamsudin, S. S., Alfandi, O., Omar, R., & Al-Fadhali, N. (2022). Twin delayed deep deterministic policy gradient-based target tracking for unmanned aerial vehicle with achievement rewarding and multistage training. IEEE Access, 10, 23545-23559.

[15] Tao, C., Li, M., Cao, F., Gao, Z., & Zhang, Z. (2024). A multiobjective collaborative deep reinforcement learning algorithm for jumping optimization of bipedal robot. Advanced Intelligent Systems, 6(1), 2300352.

[16] Jebrane, W., & El Akchioui, N. (2024). Advancing Distributed Distributional Deterministic Policy Gradients for Autonomous Robotic Control Across Diverse Environments. Journal of Control, Automation and Electrical Systems, 35(6), 1059-1077.

[17] Li, B., Qiu, S., Ye, H., Guo, Y., Wang, H., & Bai, J. (2025). Motion planning for 7-degree-of-freedom bionic arm: Deep deterministic policy gradient algorithm based on imitation of human action. Engineering Applications of Artificial Intelligence, 140, 109673.

[18] Chu, Z., Wang, Y., & Zhu, D. (2024). Local 2-d path planning of unmanned underwater vehicles in continuous action space based on the twin-delayed deep deterministic policy gradient. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54(5), 2775-2785.

[19] Chen, X., Gan, Y., & Xiong, S. (2024). Optimization of Mobile Robot Delivery System Based on Deep Learning. Journal of Computer Science Research, 6(4), 51-65.

[20] Guo, H., Liu, Z., Shi, R., Yau, W. Y., & Rus, D. (2023). Cross-entropy regularized policy gradient for multirobot nonadversarial moving target search. IEEE Transactions on Robotics, 39(4), 2569-2584.

Chapter 22

Implementation of an MVP for Healthcare Consumables Management: A Concept Paper for a Pilot Project in ULS Coimbra

Contribution #46
Cite as:
Edite Ravella, Silvino Capitão, Jorge Barreiros. Implementation of an MVP for Healthcare Consumables Management: A Concept Paper for a Pilot Project in ULS Coimbra. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-46,
  title ={Implementation of an MVP for Healthcare
     Consumables Management: A Concept Paper for a
     Pilot Project in ULS Coimbra},
  author={Edite Ravella and  Silvino Capitão and  Jorge
     Barreiros},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Implementation of an MVP for Healthcare Consumables Management: A Concept Paper for a Pilot Project in ULS Coimbra

Edite Ravella[1*], Silvino Capitão[1,2], Jorge Barreiros[1,3]
[1] Polytechnic University of Coimbra, Instituto Superior de Engenharia de Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.
[2] CERIS, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
[3] RCM[2+] Research Centre in Asset Management and System Engineering, Polytechnic University of Coimbra, Coimbra, 3030-199 Coimbra, Portugal

*Corresponding author
* a2024163363@isec.pt; capitao@isec.pt; jmsousa@isec.pt

Abstract

Hospitals face persistent challenges in managing medical consumables, including overstocking, shortages, and poor tracking, all impacting care quality and operational costs. Literature emphasizes the need for digital supply chains, integrating real-time data and automation to improve decision-making and operational efficiency. Medical consumables represent significant hospital supply expenses, requiring precise inventory tracking and efficient replenishment processes. This concept paper proposes the development of a Minimum Viable Product (MVP) of a COnsumables Management System (COMS) to be implemented in the hospital context of ULS Coimbra. It aims to integrate real-time monitoring of medical consumables, bedside direct debit of consumables, and improved communication between local and central warehouses, integrating digital solutions. The methodology combines a literature-based approach with action research, including stakeholder inquiries, requirement analysis, and phased system prototyping. Although the system has not yet been implemented, this paper outlines the theoretical framework and planned stages for development, implementation and evaluation. Expected outcomes include increased efficiency in medical consumables management, cost reduction in material waste, and improved local warehouse operations. This proposal advances digital hospital logistics and opens pathways for future empirical validation.

Keywords: MVP; Supply Chain Management; Warehouse Management Systems; COMS; CMS Medical Consumables; Digital Supply Chains; Hospital Logistics

1 Introduction

Hospitals are organizations with very specific management processes (regarding assets, data, resources, and others), and the information flows, processes and investments are often misaligned with the goals, vision and sustainability of that organization. Disorganized inventories, complex distribution processes and inadequate protocols for data keeping and processing are mostly inadequate, which can lead to significant investments in health assets that are not needed [1].

The shift to value-based care, reimbursement rate impacts, and higher costs have challenged hospitals to be more cost-effective and resourceful. GE Healthcare states that healthcare expenditure is expected to grow at an average annual rate of 5.7% until 2027 [1].

It is crucial to understand how specialized infrastructures such as healthcare institutions can be efficient in their processes regarding asset management, so that the expenditure is reduced, but also with the focus on the sustainability of the process and minimizing environmental impact.

This concept paper proposes the development of a Minimum Viable Product (MVP) of a COnsumables

Management System (COMS) that optimizes the stock of medical consumables in real time, guaranteeing an efficient flow of consumables between the central warehouse, local warehouse, healthcare professionals, and patients.

Being this a concept paper, the structure of it will be the following: the section after the introduction describes the background and purpose of doing this research, as well as the questions of research and objectives; the third section describes the research methodology; the fourth section presents the state-of-the-art of supply chain and warehouse management as well as some case studies regarding medical consumables management; section number five presents the COMS; section 6 shows the expectation of implementation of the COMS with a timeline. This concept paper concludes with the expected outcomes and challenges, and limitations.

2 Background

Projections made at the beginning of this year indicated that healthcare supply chain costs are expected to rise by 2.3% by the following year, 2026, specifically for medical and surgical equipment (such as gloves, needles, syringes, and others). This is one of the reasons why the demand for supply chain management systems and warehouse management systems is crucial at this point, to keep operations agile and resilient [2] [3].

At hospital level (public or private) there is often inefficient management of medical consumables in inpatient services, where sometimes there is too much waste on resources and materials that were not used (sometimes because there was an excess in inventory and then they expired) or sometimes there is not enough consumables in the local warehouse to answer the demands of a hospital ward.

The rate for purchasing versus using medical assets needs to change. There is an increase of 62% in the number of assets purchased per bed; however, the usage is only 40% [1]; either way, both situations lead to problems of waste, lack of essential materials, high costs and impact on the quality of care provided to patients. The effect of poor management of medical consumables is more significant in private institutions, as every medical consumable used on a patient that is not charged to them is a financial loss for the hospital itself.

It is essential to identify gaps in the current management of medical consumables and explore opportunities to maximize operational efficiency in managing these materials. The challenge here lies in building a consumable management system (COMS) that optimizes the stock of medical consumables in real-time for local warehouses (located in the hospital wards), guaranteeing an efficient flow of consumables between the central warehouse, local warehouses, healthcare professionals, and patients.

2.1. Research Questions

This concept paper aims to answer the following questions:

- What are the supply chain management processes for asset management?
- How do medical consumables integrate the supply and logistics chain of asset management?
- Are there technological systems that could be the leading technological approaches used for managing medical consumables, and what impact do they have?
- Could the integration of a COMS improve medical consumables management?
- Are there obstacles in the usage and integration of this technology?

2.2. Research objectives

The research will therefore have the following objectives:

- Understand the state-of-the-art of the current asset management
- Understand the challenges in the logistics management sector in a healthcare institution

• Identify gaps and challenges as opportunities for innovation in the supply chain of medical consumables;

• Develop a COMS MVP to be implemented in a specific ward in ULS Coimbra;

• Evaluate the impact and acceptance of this system by the different stakeholders (nurses, doctors and ward managers).

3 Research Methodology

This study adopts a mixed-methods approach, combining qualitative and quantitative methodologies within an action-research framework.

For the first part of the study, a literature review was conducted to support the development of a pilot project focused on implementing a COMS at ULS Coimbra Hospital. The review covered key concepts in supply chain and warehouse management, including warehouse types, layout, the role of Warehouse Management Systems (WMS), and research about some use cases in the healthcare sector. Articles were searched in B-on, Scielo, and Web of Science (WOS), and RCAAP, and complemented with Google Scholar and research in International Organization for Standardization (ISO) guidelines.

For the second part of the study, where it is proposed the development of an MVP for a COMS system, it is expected that, when reaching research level, the methodologies to be applied will be:

• Inquiries: to evaluate the impact and acceptance of the COMS of the different stakeholders;

• Action research: as a way to solve real-life practical problems, through a collaborative environment between the researcher and the stakeholders to find relevant solutions. In this practical case, the aim is to implement a COMS for a local warehouse and evaluate the impact of this technology on the hospital supply chain [4].

The study will also collect data from primary and secondary sources, and the quantitative data will be analised using descriptive statistics. It is also expected that the implementation of the COMS in a pilot project will involve a practical implementation and iterative testing.

4 State-of-the-art

The supply chain is a concept with multiple levels that encompasses all points, from the initial stage (raw materials) to the final stage (final product). Each step of this chain embraces unique characteristics and challenges that companies and businesses need to deal with and find strategies to improve the value of that chain sustainably [5].

Richey et. al [6] defined logistics and supply chain management as a combination of 3 perspectives: i) an approach as a whole integrated system to manage the inventory assets from the supplier to the final user; ii) an orientation for intra and inter operational capabilities synchronized as one; iii) an individualized source of customer value, leading to customer satisfaction.

APICS [7] presented the Supply Chain Operations Reference model (SCOR) that describes the business activities organized around 6 management processes: i) Plan; ii) Source; iii) Transform; iv) Return; v) Fulfill; and vi) Order.

A foundational perspective on supply chain management that brings together stakeholders, partners and customers: Responsiveness, as it was suggested by Richey et. al [6]. It can be defined as: "*A network of firms that is capable of creating wealth for its stakeholders in a competitive environment by reacting quickly and cost-effectively to changing market requirements*." [8].

The model proposed by Richey et. al [6] is based on 5 main dimensions: Adaptability (strategic adjustments according to market feedback); Flexibility (adjust the policies based on the interactions with the customers);

Agility (adjust the processes based on the reaction of the customers and external markets); Improvisation (adjustments planned on-the-go); and Resilience (sustain an amount of disturbance in the supply chain, without rupturing any of the communications flow) [6].

Some authors have explored the possibility of merging the LEAN, agile, resilient, and green paradigms in the supply chain management approach [9]. These frameworks promote inventory minimization, but the resilience approach demands the existence of strategic inventory buffers, but it is stated that all of these paradigms should be developed simultaneously for supply chain performance to be improved [9].

Figure 1 shows how practices influence the supply chain (the lean practices are inside the colored box) and to what extent those practices influence the supply chain, represented through the mathematical symbols of "+" and "-" meaning a positive or negative influence, respectively. For example, using a Just-in-Time practice leads to a higher replenishment frequency of assets and materials. In the same image, it is also possible to see how all the factors increase or decrease the cost in the supply chain [9].

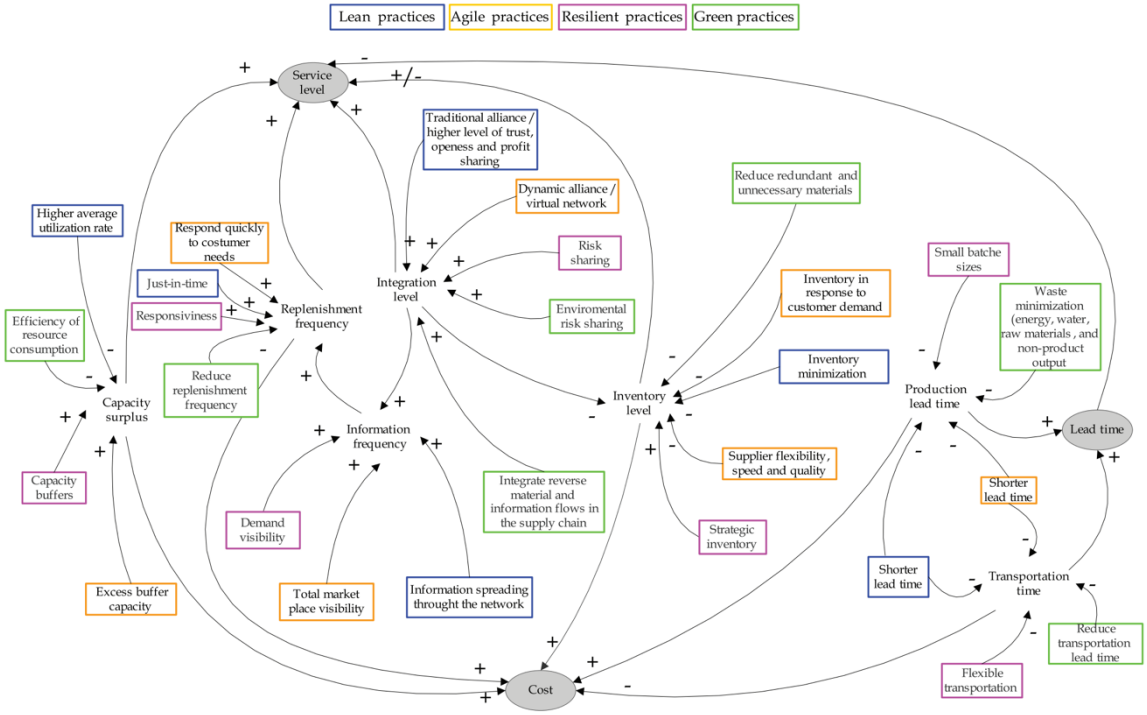


Figure 1: The conceptual model presented by Carvalho & Cruz-Machado (2011) [9]

A recent study done in 2023 by the American Productivity and Quality Center [10] surveyed more than 350 professionals from multiple industries and they found that 62% of the organizations did not reach the goals of performance planned for 2023.

4.1 Warehouses

Warehouses play a significant role in the journey of assets: storing, managing, and distribution, are essential to ensure a smooth flow of products and eliminate unnecessary costs [11].

Gurram et. al [12] highlight that warehouses play a crucial role in the entire process from the manufacture of

the asset to end-user delivery, reinforcing the idea that these structures have an impact on the efficiency of the supply chain.

Warehousing operations involve inventory management, order selection, packaging, shipping, and distribution [11]. The goal in managing all these operations is to enhance the efficiency of asset or equipment flow. Each of these transactions is associated with a logistics operation, along with efficient material handling of goods. On top of that, the data management and flows should always accompany the asset flow [13].

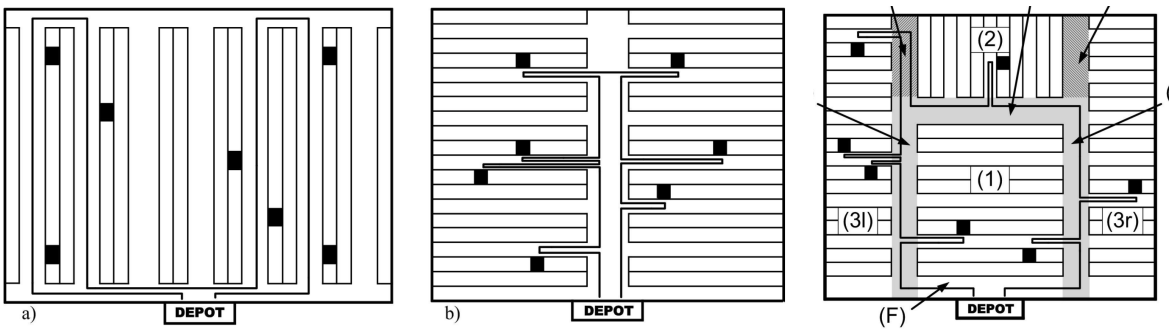
Traditional warehousing relies heavily on manual labor for asset handling, inventory tracking, and order picking, resulting in higher costs and limited scalability. In contrast, automated warehousing uses robotic systems, digital inventory tools, and WMS, enabling more efficient operations, reduced labor costs, and improved storage and workflow management [11].

The layout of warehouses is also an important aspect to consider, since they aim to give a proper answer to a market need and the orders' demands. The best warehouse layout is the one that maximizes the use of the available space and minimizes the number of travelling journeys and points of contact for a given asset [14]. The order-picking processes should be optimized, since they represent 55% of the total operating costs of a warehouse [15].

Albert, Ronnqvist & Lehoux [16] highlighted the forward area as an interesting methodology for warehousing efficiency. Their review identified different types of layout designs, where the storage zones and rack dimensions were the variables most studied by various researchers and were seen as having an impact on warehouse efficiency. These authors identified that non-conventional or heterogeneous warehouse types (cross-docking, JIT [Just-in-Time] environment) are common in the industry sector.

Non-traditional warehouses can have different layouts, for example, a V-shaped warehouse, where the picking aisles below the V are horizontal and above the V are vertical. These layouts show a reduction of workers' travels inside the infrastructure of 10%-15% when compared to a traditional warehouse [15].

Also, there can be warehouses that have a single-block warehouse (picking aisles are parallel to each other) in Figure 2a; a two-block warehouse (a vertical aisle with two blocks of horizontal picking aisles), in Figure 2b; or a U-shape layout (the central aisle is arranged in an upside-down U shape, with 2 vertical aisles connected by a horizontal cross aisle), in Figure 2c [17].



a) b) c)

Figure 2: Warehouse layouts: a) single-block; b) two-block; c) U-shape [17].

The conclusion was that the two-block layout has a greater performance when compared to the single-block layout when there is a high number of picking assets to be done. The U-shaped layout had the best performance for medium-sized numbers of picks, and the travelling distance of the workers could be reduced compared to the previous ones. The U-shaped is also best suited for saving space than the other ones [17].

4.2 Warehouse Management Systems

Automated warehousing represents a significant advancement in logistics management, integrating robotic systems, guided vehicles, and digital technologies to streamline operations. By leveraging Warehouse Management Systems (WMS), these warehouses enable real-time inventory control, optimized space utilization through tall racking structures, and improved operational efficiency [11].

It is crucial to have data from all steps and management processes within all parts of the supply chain, including data from warehouse management [18]. Major trends were identified: big data and analytics (65% of the professionals' answer); digitalization of the supply chain (64%); data management (62%), and process standardization (53%). These were the ones appointed to have a higher impact [10].

Data is collected into a WMS from various sources. The goal is to have direct access to stock information, based on real-time data from the sources. Barcode systems for picking and reducing human error are commonly used, as are mobile computers, wireless networks, and radio-frequency identification (RFID) sensors [11] [14]. Once the data is gathered from these sources, it is integrated into a central database in real-time, and it is that database that allows to track stock efficiently and provide reports about the status of the assets within the warehouse.

AI can also be a powerful solution for analyzing large amounts of data, and outcomes can be provided in seconds [18].

WMS are essential to guarantee the flow of assets for the customer, minimization of errors in picking the orders, reduce the risk of stock outs, and facilitate the flow of assets and their respective information. Having a system that can predictively forecast the shortage of an asset or issue alerts for problematic situations inside the warehouse is fundamental to maintaining the integrity of the processes and the chain [11] [18].

4.3 Case Studies of Medical Consumables

The setting for this pilot project is a healthcare institution, so it was necessary to understand the concepts and frameworks behind asset management and logistic chains in a healthcare setting.

Healthcare institutions are also responsible for operations related to asset management: planning, purchasing, managing, handling, tracking, and transporting. Supply chain management in a hospital requires strategies as well to manage costs, manage inventory and space available for the storing of assets (medication, medical equipment and others) [19].

The COVID-19 crisis brought to light a problem of supplies management, and in the study performed by Spieske et. al [20] all investigated hospitals reported a significant increase in demand regarding medical supplies in intensive care patients, leading to a sharp rise in that field. A significant proportion of hospital costs is devoted to logistics activities, which can account for up to 46% of the hospital budget [21].

Balkhi et. al [19] used a JIT approach in healthcare used for restocking inventory and placing an order for resources when a minimum value is reached. For each supply category, the order volume is determined to prevent stock shortage during the time between orders. Differences in demand between different types of supplies are expected and must be determined using robust software. JIT requires the establishment of a monitoring and evaluating system that can detect and determine minimal acceptable inventory levels.

Shashidharan & Anwar [13] integrated several technologies including RFID (Radio-Frequency Identification), robotics, QR (Quick Response) codes, barcodes, sensors and mobile devices. This software enables efficient inventory management with real-time replenishment and optimal asset quantities. The implementation of these systems allowed for effective warehouse management that reduced the constraints of data management and cost optimization. The results showed that workers gave more importance to systems that could help with data

management and accuracy, as well as the implementation of safety parameters.

Another proposal to improve warehouse performance was the implementation of KPI (Key Performance Indicators) by Marziali et. al [22]. The KPIs definition was based on the current level of stock, the goal intended to achieve (in this case was the maximum capacity of the product) and on a traffic light rating system (using the colors green, yellow and red). This work demonstrates that the development of indicators is an efficient and helpful approach, as well as incorporating more sophisticated technologies in data collection, to add flexibility to the process [22].

Another use case involved implementing a Warehouse Management System (WMS) for asset management within the warehouse, utilizing data from the assets themselves and barcodes, as well as Bluetooth and other communication protocols. Even though demonstrated empirically, with synthetic dataset, the results show that this system increased the warehouse efficiency, by utilizing the space consciously and reducing retrieval times for the assets, based on real-time data provided by the sensors and other devices [12].

In a different sector, but comparable, Machado [23] researched the use of barcodes in the assets when arriving at the warehouse and to manage the materials inside the warehouse. Results indicated that the flow of the asset significantly reduced the time spent on picking processes by 35%.

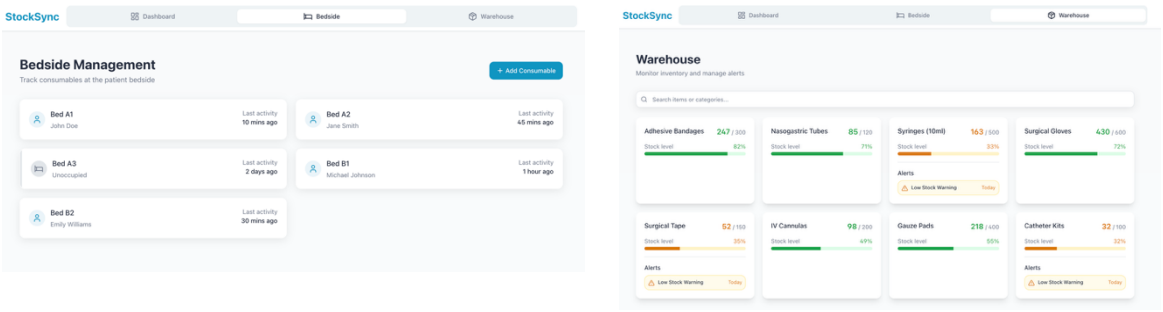
A specific use case identified in the ULS Coimbra, from Baeta [24], which developed a Power BI tool to provide information for the logistics department. The implementation of the data-driven decision tool could lead to cost savings of up to 7% and a decrease in the average stock of 37% for the items tested.

Another use case in a Portuguese hospital involved the application of methodologies such as the Kanban system, and the definition of KPIs was used to enhance the efficiency of the hospital's ambulatory pharmacy. With the implementation of these tools, there was a decrease in the stock-out levels (4.5% in the first phase and 1.5% in the second phase of implementation). Overall, there was a reduction of 30% in out-stocks situations [25].

5 COMS

The COnsumables Management System (COMS) system to be build will help in the optimization of the stock of medical consumables in real-time, for the local warehouses of the hospital, and that can guarantee an efficient flow between the central warehouse, local warehouse, healthcare professionals, and patients.

Figure 3 represents a prototype of the COMS to be developed and implemented on the hospital ward. The Figure 3a) represents the page of the system where medical consumables can be inserted into the platform, associated with an identified patient. Figure 3b) represents the current stock management of the local warehouse, with the different types of materials and with alerts for low (in orange) and critical stock (not represented).



a) b)

Figure 3: Prototype of the COMS

Next steps to be followed on the development of this system are to characterize the healthcare institution where the pilot project will take place, doing a brief description about the units, services, human resources, cost allocation and other data that can help describe the institution as its whole; and to divide the pilot project into 3 different phases:

- pre-implementation phase, where there is market research to understand the warehouses' layout, data flows for the assets and the information, and the characterization of WMS;
- implementation phase, divided into phase 1 (the development of a COMS prototype) and phase 2 (the COMS MVP). Following both phases, a report is compiled detailing the evaluation and feedback provided by stakeholders.
- post-implementation phase, where the acceptance and impact of the COMS are evaluated.

6 Timeline

The planning and implementation of the pilot project follow a series of steps. Those are described as follows:

- Step 1 – review the methods and data collection sources;
- Step 2 – validation & requirements analysis, which includes: surveys of different stakeholders, validation for the need of implementation a COMS; requirement definition, identification of a ward to implement the pilot project;
- Step 3 – authorization requests and access to a list of medical consumables in the ward selected;
- Step 4 – phase 1 of the pilot project: validation of the prototype of the COMS;
- Step 5 – phase 2 of the pilot project: implementation of the MVP of the COMS;
- Step 6 – evaluation of impact, acceptance, and feedback about inefficiencies in the system or its use;
- Step 7 – presentation to the stakeholders of ULS Coimbra Hospital.

Table 1 shows the timeline of implementation, with an expected duration for each step.

Table 1 – Timeline of implementation of the COMS

Step of Research	Duration
1. Practical Methodology Review	2 months
2. Validation & Requirements	2 months
3. Authorizations & Consumables	2 months
4. Pilot Project: Phase 1	2 months and half
5. Pilot Project: Phase 2	2 months and half
6. Evaluation	2 months
7. Report & Result Dissemination	10 months

7 Conclusion

The efficient management of medical consumables in healthcare settings is increasingly critical as operational demands, supply chain complexities, and cost pressures continue to intensify.

This concept paper proposes the development and implementation of a pilot project for a COMS at ULS Coimbra to address key inefficiencies in current stock management practices. The proposed COMS aims to enable real-time tracking of consumables, streamline communication between hospital warehouses, and automate material debit processes at the point of care.

Through a mixed-methods and action-research approach, the project aims to understand user needs, test technological solutions in a real-world hospital ward, and evaluate the system's impact and stakeholder acceptance. Grounded in current literature and supported by relevant case studies, the concept of a COMS MVP is designed not only as a technical solution but also as a strategic move toward the digital transformation of hospital logistics.

While implementation is still forthcoming, the theoretical and practical foundations laid in this paper provide a strong basis for initiating the pilot project. The expected benefits include improved resource planning, cost reduction through waste minimization, enhanced workflow for healthcare professionals, and ultimately, better patient care.

Having an effective process, strategies and approaches to supply chain management leads an organization to reduce the costs and cut the delivery times to customers. Simplifying the steps of the production and distribution chain can be a significant measure to achieve those goals [18].

Closing Remarks

Despite its promising potential, implementing a COMS in healthcare settings is not without challenges. The cost of digital systems, integration issues with existing infrastructure, and the slower adoption rate in healthcare due to regulatory and operational constraints present notable barriers. Additionally, physical space limitations and outdated inventory practices in hospital warehouses can complicate deployment.

Future research should focus on evaluating system scalability across different hospital departments, measuring long-term cost-effectiveness, and exploring integration with broader hospital information systems. Addressing these limitations through iterative development, stakeholder involvement, and adaptive technologies will be essential to ensure the successful implementation and sustainability of digital supply chain solutions in healthcare environments.

References

[1] S. Elston, "Moving Beyond Break-Fix Maintenance: Holistic Healthcare Technology Management," GE Healthcare, 2020. [Online]. Available: https://www.gehealthcare.co.kr/-/jssmedia/1fda3fcdc4544004afa0ecfdb9e6aa65.pdf.

[2] Becker's Hospital Review, "Healthcare supply chain costs expected to rise 2.3%: Report," Becker's Hospital Review, January 2025. [Online]. Available: https://www.beckershospitalreview.com/supply-chain/healthcare-supply-chain-costs-expected-to-rise-2-3-report/.

[3] Fortune Business Insightss, "Healthcare Supply Chain Management Market Source: https://www.fortunebusinessinsights.com/industry-reports/healthcare-supply-chain-management-market-101051," Fortune Business Insightss, May 2025. [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/healthcare-supply-chain-management-

market-101051.

[4] A. Melnikovas, "Towards an Explicit Research Methodology: Adapting Research Onion Model for Futures Studies," *Journal of Futures Studies,* vol. 23, no. 2, pp. 29-44, http://dx.doi.org/10.6531/JFS.201812_23(2).0003.

[5] ISO, "Automatic identification and data capture techniques — Supply chain applications of RFID — Product tagging, product packaging, transport units, returnable transport units and returnable packaging items," ISO, 2023. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:17360:ed-1:v1:en.

[6] R. G. Richey, A. S. Roath, F. G. Adams and A. Wieland, "A Responsiveness View of logistics and supply chain management," *Journal of Business Logistics,* vol. 43, no. 1, pp. 62-91, 2022. http://dx.doi.org/10.1111/jbl.12290.

[7] APICS, "Supply Chain Operations Reference Model: SCOR," 2017. [Online]. Available: https://www.apics.org/docs/default-source/scor-training/scor-v12-0-framework-introduction.pdf.

[8] A. Gunasekaran, K.-h. Lai and T. E. Cheng, "Responsive supply chain: A competitive strategy in a networked economy," *Omega,* vol. 36, no. 4, pp. 549-564, 2008. http://dx.doi.org/10.1016/j.omega.2006.12.002.

[9] H. Carvalho and V. Cruz-Machado, "Integrating Lean, Agile, Resilience and Green Paradigms in Supply Chain Management (LARG_SCM)" in *Supply Chain Management*, Croácia, InTech, 2011, pp. 27-48. https://doi.org/10.1007/s12351-022-00732-y.

[10] Supply Chain Management Review, "Preparing for 2024 supply chain challenges and priorities," 2024. [Online]. Available: https://www.scmr.com/article/preparing_for_2024_supply_chain_challenges_and_priorities.

[11] O. F. Odeyinka and O. G. Omoegun, "Warehouse Operations: An Examination of Traditional and Automated Approaches in Supply Chain Management," in *Operations Management - Recent Advances and New Perspectives*, IntechOpen, 2024. http://dx.doi.org/10.5772/intechopen.113147.

[12] A. Gurram, K. M. Chowdary, N. P. Neelesh, R. S. T. Reddy and P. R. A. Bangalore, "Optimal Warehouse Slotting in Supply Chain Management System," in *2024 5th International Conference for Emerging Technology (INCET)*, Belgium, 2024. http://dx.doi.org/10.1109/INCET61516.2024.10593363.

[13] M. Shashidharan and S. Anwar, "Importance of an Efficient Warehouse Management System," *Turkish Journal of Computer and Mathematics Education,* vol. 12, no. 5, pp. 1185-1188, 2021. Retrieved from https://turcomat.org/index.php/turkbilmat/article/view/1784.

[14] I. H. Mohamud, M. A. Kafi and S. A. Shahron, "The Role of Warehouse Layout and Operations in Warehouse Efficiency: A Literature Review," *Journal Européen des Systèmes Automatisés,* vol. 56, no. 1, pp. 61-68, 2023. http://dx.doi.org/10.18280/jesa.560109.

[15] G. Dukic and T. Opetuk, "Warehouse Layouts," in *Warehousing in the Global Supply Chain: Advanced Models, Tools and Applications for Storage Systems.*, London, Springer, 2014. doi 10.1007/978-1-4471-2274-6_3

[16] P.-W. Albert, M. Ronnqvist and N. Lehoux, "Trends and new practical applications for warehouse allocation and layout design: a literature review," *SN Applied Sciences,* vol. 5, no. 378, 2023. https://doi.org/10.1007/s42452-023-05608-0.

[17] S. Henn, S. Koch, H. Gerking and G. Wascher, "A U-Shaped Layout for a Manual Order Picking System," *Logistics Research,* vol. 6, no. 4, 2011. http://dx.doi.org/10.1007/s12159-013-0104-6.

[18] B. M. Mohsen, "Developments of Digital Technologies Related to Supply Chain Management," *Procedia Computer Science,* vol. 220, 2023. http://dx.doi.org/10.1016/j.procs.2023.03.105.

[19] B. Balkhi, A. Alshahrani and A. Khan, "Just-in-time approach in healthcare inventory management: Does it really work?," *Saudi pharmaceutical journal,* vol. 30, no. 12, pp. 1830-1835, 2022. http://dx.doi.org/10.1016/j.jsps.2022.10.013.

[20] A. Spieske, M. Gebhardt, M. Kopyto and H. Birkel, "Improving resilience of the healthcare supply chain in a pandemic: Evidence from Europe during the COVID-19 crisis," *Journal of Purchasing and Supply Management,* vol. 28, no. 5, 2022. http://dx.doi.org/10.5281/zenodo.12665400.

[21] F. Jawab, Y. Frichi and S. Boutahari, "Hospital Logistics Activities," in *International Conference on Industrial Engineering and Operations Management Bandung*, Indonesia, 2018. https://www.ieomsociety.org/ieom2018/papers/31.pdf.

[22] M. Marziali, D. Rossit and A. Toncovich, "Warehouse Management Problem and a KPI Approach: a Case Study," *Management and Production Engineering Review,* vol. 12, no. 3, pp. 51-62, 2023. http://dx.doi.org/10.24425/mper.2021.138530.

[23] L. P. Machado, "Operations improvement in an internal warehouse of a construction sector company," 2022. https://hdl.handle.net/1822/83220.

[24] J. Baeta, "Inventory Management in Healthcare: Case Study in a Public Hospital," 2024. https://estudogeral.uc.pt/retrieve/275241/JoseDiogoBaeta_2018278543.pdf.

[25] C. Castro, T. Pereira, J. C. Sa and G. Santos, "Logistics reorganization and management of the ambulatory pharmacy of a local health unit in Portugal," *Evaluation and Program Planning,* vol. 80, 2020. http://dx.doi.org/10.1016/j.evalprogplan.2020.101801.

Chapter 23

Multi-Head Loss-Driven Brain Stroke Segmentation Using Auxiliary Mask Supervised Swin Transformer

Contribution #47
Cite as:
Batyrkhan Omarov, Zhanseri Ikram. Multi-Head Loss-Driven Brain Stroke Segmentation Using Auxiliary Mask Supervised Swin Transformer. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-47,
  title ={Multi-Head Loss-Driven Brain Stroke
     Segmentation Using Auxiliary Mask Supervised Swin
     Transformer},
  author={Batyrkhan Omarov and  Zhanseri Ikram},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Multi-Head Loss-Driven Brain Stroke Segmentation Using Auxiliary Mask Supervised Swin Transformer

Batyrkhan Omarov[1,2,*], Zhanseri Ikram[1]

[1] Narxoz University, Almaty, Kazakhstan
[2] International Information Technology University, Almaty, Kazakhstan

*Corresponding author
batyahan@gmail.com

Abstract

Accurate and timely segmentation of stroke lesions in brain MRI is critical for effective clinical diagnosis and treatment planning. This paper proposes a novel architecture that combines the power of Swin Transformers with auxiliary mask supervision and a multi-head loss strategy to address the challenges inherent in ischemic stroke lesion segmentation. The model processes axial MRI slices using a patch embedding layer followed by a four-stage Swin Transformer, enabling hierarchical feature extraction with global contextual awareness. An auxiliary branch, guided by bounding box-derived pseudo-segmentation masks, enhances structural learning and refines segmentation outputs through a dedicated mask prediction head. Simultaneously, classification and localization heads are employed to detect stroke presence and regress lesion coordinates, respectively. These outputs are jointly optimized using a composite loss function integrating binary cross-entropy, mean squared error, and Dice loss components. Evaluation on the ISLES 2024 dataset demonstrates that the proposed architecture achieves high segmentation accuracy, robust lesion localization, and reliable stroke detection across diverse cases. Visual results confirm its ability to identify and delineate complex lesion structures with high confidence. This approach establishes a strong foundation for transformer-based medical image analysis and offers significant promise for clinical deployment in automated stroke assessment systems.

Keywords: Brain stroke segmentation; Swin Transformer; auxiliary mask supervision; multi-head loss; medical image analysis; MRI; ischemic stroke; deep learning; lesion localization

1 Introduction

Brain stroke remains one of the leading causes of death and long-term disability worldwide, necessitating timely and precise diagnosis to improve patient outcomes [1]. Magnetic Resonance Imaging (MRI) plays a pivotal role in stroke evaluation due to its high soft-tissue contrast and capacity to detect early ischemic changes. However, manual delineation of stroke lesions is labor-intensive, subject to inter-observer variability, and unfeasible in time-critical scenarios. Consequently, the development of automated and accurate segmentation systems is of paramount importance for clinical decision-making [2], [3].

Recent advances in deep learning have significantly enhanced medical image segmentation. Convolutional neural networks (CNNs), particularly U-Net variants, have demonstrated strong performance in segmenting lesions from brain scans [4]. However, these models often suffer from limited receptive fields and poor global context understanding, which is especially detrimental in cases of heterogeneous stroke presentations. Transformers, originally designed for natural language processing, have emerged as powerful alternatives due to their self-attention mechanisms that capture long-range dependencies [5]. The Swin Transformer, in particular, introduces hierarchical representations with shifted windows, offering both computational efficiency and strong feature expressiveness in vision tasks [6].

Despite these advances, challenges remain in achieving fine-grained and context-aware stroke lesion segmentation. Auxiliary supervision and multi-task learning have proven effective in guiding the model toward better generalization and structural awareness [7]. Moreover, integrating pseudo-labels or bounding

box-guided supervision has shown promise in improving segmentation performance, especially in weakly supervised scenarios [8]. In this study, we propose a novel architecture that combines the strengths of Swin Transformers with an auxiliary mask supervision mechanism and a multi-head loss strategy. Our model is designed to enhance the segmentation of brain stroke lesions by fusing global contextual understanding, fine spatial details, and multi-objective optimization. This approach is validated on benchmark datasets to demonstrate its superiority over existing methods.

2 Related Works

The task of brain stroke segmentation has been extensively studied, with early approaches relying on hand-crafted features and traditional machine learning classifiers. However, these methods often failed to generalize due to the high variability in lesion appearance and size across patients [9]. The advent of deep learning brought a paradigm shift, particularly with the introduction of encoder-decoder convolutional neural networks (CNNs), which have become the de facto standard in medical image segmentation [10]. Among these, the U-Net architecture and its derivatives have demonstrated considerable success in segmenting ischemic and hemorrhagic stroke lesions in MRI scans [11], [12]. Despite their effectiveness, CNN-based models struggle to capture long-range dependencies, which limits their performance in segmenting diffuse or small lesions.

To overcome these limitations, researchers have increasingly adopted transformer-based architectures, which leverage self-attention mechanisms to model global spatial relationships [13]. The Vision Transformer (ViT) introduced a compelling approach for visual tasks by treating images as sequences of patches; however, its high computational cost and lack of inductive biases posed challenges in medical imaging [14]. The Swin Transformer improved upon this by introducing hierarchical feature learning and shifted windows, enabling scalable and efficient vision processing [15]. Several studies have successfully applied Swin Transformers to medical imaging tasks, including tumor segmentation and organ delineation [16], [17].

Beyond architectural advancements, auxiliary supervision has been employed to enhance model training through multi-task learning, guiding the network with additional objectives such as edge detection or anatomical priors [18]. Incorporating auxiliary mask heads has shown potential in improving segmentation accuracy by providing intermediate gradient signals and promoting feature reuse [19]. Additionally, the use of bounding box-guided pseudo-segmentation masks has emerged as a promising strategy for weakly supervised and semi-supervised learning scenarios [20], [21]. These methods allow for leveraging coarse annotations to generate dense label maps, which can be refined through network feedback.

Multi-head loss functions have also gained attention as a means to balance diverse objectives such as classification, localization, and segmentation [22]. This approach enables more robust training dynamics and better convergence, especially in complex medical imaging tasks [23]. Recent works have demonstrated the benefits of combining global feature extraction with localized mask refinement, particularly when dealing with ambiguous lesion boundaries [24], [25]. Building on these findings, our work integrates a Swin Transformer backbone with an auxiliary mask head and a multi-loss strategy to achieve precise and context-aware brain stroke segmentation.

3 Materials and Methods

The proposed architecture for brain stroke segmentation, illustrated in Figure 1, is designed to exploit the hierarchical representational power of Swin Transformers alongside an auxiliary supervision strategy guided by bounding box pseudo-segmentation masks. The network incorporates three main components: the Swin Transformer backbone, the auxiliary mask branch, and multi-head output modules optimized via a composite loss function.

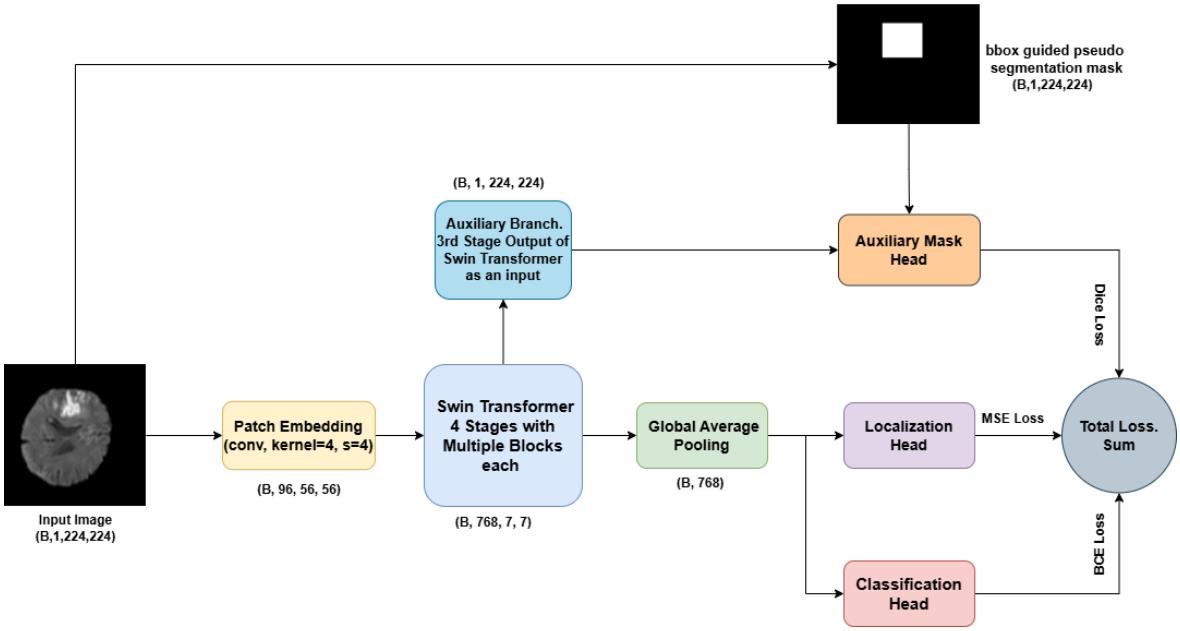


Figure 1: Proposed architecture for brain stroke segmentation using a Swin Transformer with auxiliary mask supervision and multi-head loss optimization.

The model accepts a pre-processed axial slice from a brain MRI scan, denoted as an input tensor $X \in R^{B\times1\times224\times224}$, where B is the batch size. The input is passed through a Patch Embedding layer, which applies a convolution with kernel size 4×4 and stride $s=4$, transforming the spatial resolution and generating non-overlapping patches:

$$P = Conv_{4\times4}(X) \in R^{B\times96\times56\times56} \quad (1)$$

These patch tokens are then fed into a 4-stage Swin Transformer module, each comprising multiple shifted window-based transformer blocks. The final output of the Swin Transformer is a hierarchical representation:

$$F_{main} \in R^{B\times768\times7\times7} \quad (2)$$

which is subjected to Global Average Pooling:

$$G = GAP(F_{main}) \in R^{B\times768} \quad (3)$$

This pooled feature is branched into two output heads:

1. Classification Head: Performs binary classification of the presence of stroke.
2. Localization Head: Regresses coordinates for coarse lesion localization.

Simultaneously, the third stage output of the Swin Transformer, $F_3 \in R^{B\times1\times224\times224}$, is routed to the Auxiliary Branch, which operates as a high-resolution refinement path. This feature map, concatenated with a bounding box-guided pseudo-segmentation mask $\hat{M} \in R^{B\times1\times224\times224}$, is input into the Auxiliary Mask Head to recover fine-grained pixel-wise lesion predictions.

The total loss function L_{total} is formulated as a weighted sum of three components:

$$L_{total} = \lambda_{cls}L_{BCE} + \lambda_{loc}L_{MSE} + \lambda_{mask}L_{Dice} \quad (4)$$

Where L_{BCE} is the binary cross-entropy loss for classification, L_{MSE} is the mean squared error for localization, and L_{Dice} is the Dice loss for segmentation supervision. The hyperparameters λ_{cls}, λ_{loc},

λ_{mask} balance the contributions of each loss. By integrating auxiliary supervision with transformer-driven global feature modeling, this architecture improves the structural coherence and segmentation accuracy for stroke lesions in MRI data.

Dataset. In this study, we utilize the ISLES 2024 (Ischemic Stroke Lesion Segmentation) dataset [26], a benchmark collection widely adopted for evaluating automated methods in brain stroke lesion detection and segmentation. The dataset consists of multi-modal MRI scans acquired from acute ischemic stroke patients, encompassing diffusion-weighted imaging (DWI), fluid-attenuated inversion recovery (FLAIR), and apparent diffusion coefficient (ADC) modalities. Each case is annotated by expert neuroradiologists, providing voxel-wise ground truth segmentations of stroke lesions. The dataset is partitioned into training, validation, and test sets, with corresponding metadata detailing lesion size, stroke onset time, and imaging protocols. These attributes offer a diverse and realistic representation of clinical stroke variability, enhancing the robustness of algorithmic validation.

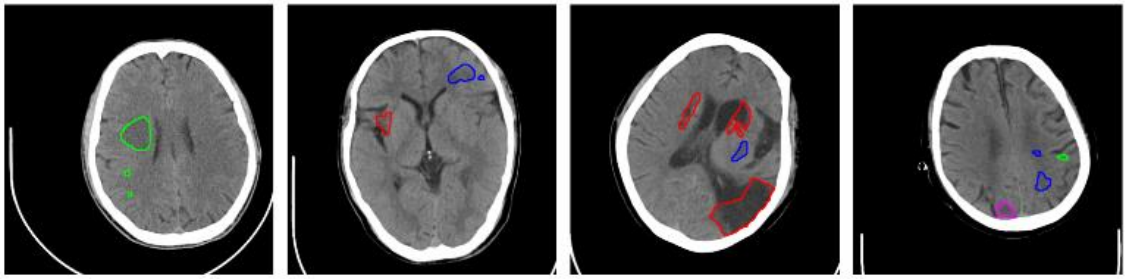


Figure 2: Sample axial slices from the ISLES 2024 dataset showing stroke lesions with varying sizes, shapes, and intensities across different MRI modalities.

Figure 2 illustrates representative slices from the ISLES 2024 dataset, showcasing the heterogeneity in lesion appearance, size, and location. As seen in the figure, some lesions appear hyperintense in FLAIR and DWI modalities, while others present subtle textural differences, making them challenging to detect. Such variations underscore the need for models that can effectively capture both global contextual relationships and fine-grained local features. The proposed architecture is trained on the training subset using preprocessed and skull-stripped axial slices, with input images normalized to a fixed resolution of 224×224 pixels. During training, data augmentation techniques, including rotation, flipping, and intensity scaling, are applied to enhance model generalization.

4 Results

To comprehensively assess the performance of the proposed architecture for brain stroke segmentation, we employ a set of widely accepted evaluation metrics that quantify both detection accuracy and segmentation quality. These metrics are computed on the validation and test subsets of the ISLES 2024 dataset and serve to evaluate the model's ability to localize stroke regions, predict segmentation masks, and classify stroke presence.

For segmentation accuracy, we use the Dice Similarity Coefficient (DSC) [27], which measures the spatial overlap between the predicted segmentation mask $\hat{M}$ and the ground truth mask M, defined as

$$DSC = \frac{2 \cdot \left| \hat{M} \cap M \right|}{\left| \hat{M} \right| + \left| M \right|} \tag{5}$$

This metric ranges from 0 to 1, with higher values indicating better agreement between prediction and ground truth.

To evaluate the model's localization performance, we use the Mean Squared Error (MSE) [28] between predicted and actual bounding box coordinates. Given the predicted coordinates $\hat{b}$ and ground truth b,

MSE is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{b}_i - b_i\right)^2 \tag{6}$$

where n is the number of test instances.

For stroke presence classification, we report the Binary Cross-Entropy (BCE) loss, as well as classification accuracy, precision, recall, and F1-score. These metrics are crucial in evaluating the discriminative power of the classification head, particularly for imbalanced datasets where stroke cases may be underrepresented.

Together, these evaluation metrics provide a comprehensive understanding of the model's performance across multiple dimensions—segmentation precision, localization accuracy, and binary classification robustness—validating the effectiveness of the proposed auxiliary mask-supervised, multi-head loss-driven Swin Transformer architecture.

Figure 3 presents qualitative results of the proposed architecture applied to samples from the ISLES 2024 dataset. The displayed axial MRI slices depict both the model's predicted stroke regions (red bounding boxes) and the ground truth annotations (green bounding boxes), alongside the predicted stroke probabilities for each case. These visualizations demonstrate the model's strong localization and segmentation capabilities across diverse lesion presentations.

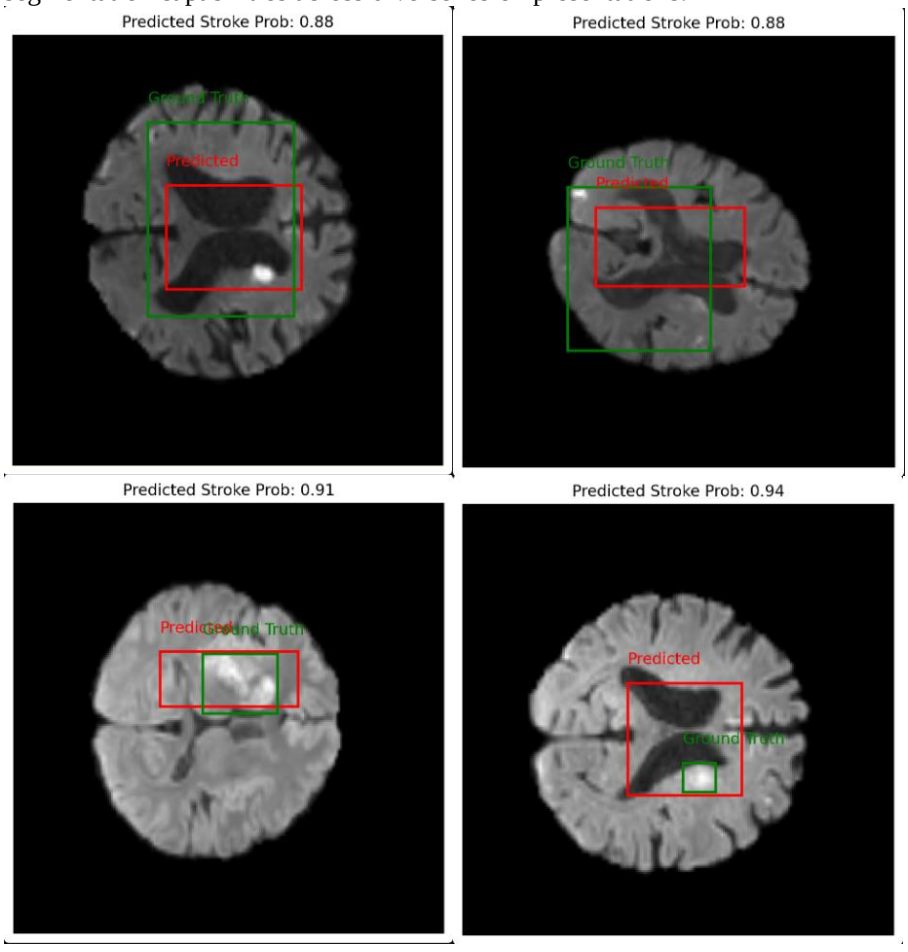


Figure 3: Sample axial slices from the ISLES 2024 dataset showing stroke lesions with varying sizes, shapes, and intensities across different MRI modalities.

Despite the inherent variability in lesion size, contrast, and anatomical location, the proposed Swin Transformer-based architecture consistently identifies stroke regions with high confidence (probabilities ranging from 0.88 to 0.94). In several cases, the predicted bounding boxes closely overlap the annotated ground truth, highlighting the model's capacity to capture complex spatial patterns and contextual cues. Notably, in challenging examples with diffuse or low-contrast lesions, the auxiliary mask supervision enhances the model's structural awareness, enabling it to maintain accurate predictions. These results visually support the quantitative improvements observed in segmentation metrics, affirming the effectiveness of the multi-head loss optimization strategy and the added value of pseudo-mask guidance during training.

6 Discussion

The proposed architecture introduces a novel integration of Swin Transformer blocks, auxiliary pseudo-mask supervision, and a multi-head loss formulation to tackle the complex problem of brain stroke segmentation. Through extensive experiments on the ISLES 2024 dataset, our model demonstrated both high segmentation accuracy and consistent stroke localization across diverse MRI cases. The architecture was explicitly designed to address critical limitations in conventional convolutional networks, such as their restricted receptive fields and limited capacity to capture global contextual information.

One of the key design innovations is the adoption of the Swin Transformer as the core backbone, which leverages a shifted window self-attention mechanism to efficiently model long-range dependencies within MRI slices. Unlike traditional CNNs, which tend to focus on local structures, transformer-based architectures are inherently better suited for capturing spatially dispersed features—an essential characteristic in ischemic stroke lesions that vary significantly in shape, size, and location [29]. Our findings are consistent with recent literature, where transformers have outperformed CNNs in several medical image segmentation tasks [30].

Another pivotal component is the auxiliary mask head supervised by bounding box-guided pseudo-segmentation masks. This weakly-supervised learning technique proved beneficial for refining spatial representations without requiring densely annotated masks for each training sample. By injecting intermediate segmentation guidance at the third stage of the Swin Transformer, the auxiliary branch promotes structural awareness during feature learning. This echoes prior work demonstrating the efficacy of auxiliary supervision in guiding deep models toward semantically meaningful representations.

The multi-head loss design—comprising Dice loss for segmentation, mean squared error for localization, and binary cross-entropy for classification—enables the model to simultaneously learn from multiple objectives. This multi-task formulation helps regularize the network, reducing overfitting and enhancing generalization. It aligns with trends in contemporary research where composite loss functions are leveraged to achieve balanced optimization in complex learning tasks [31].

Qualitative results (Figure 3) further validate our architectural choices. The proposed model accurately identifies stroke regions even in difficult cases with low contrast or subtle texture changes. The predicted bounding boxes largely overlap with the annotated regions, and the predicted stroke probabilities show high confidence. These observations are critical, particularly in clinical settings where timely and precise lesion identification can significantly influence treatment decisions.

However, the approach is not without limitations. First, although pseudo-masks provide a cost-effective supervision signal, they may introduce noise if the bounding box annotations are imprecise. This can affect the convergence stability and accuracy of the auxiliary segmentation head. Second, the current framework operates on individual 2D slices, which may limit the model's understanding of 3D anatomical context. Future work could address this by extending the architecture to 3D Swin Transformers or incorporating inter-slice attention mechanisms. Additionally, while the ISLES 2024 dataset offers high-quality annotations, domain adaptation to

other datasets or institutions may require fine-tuning to account for scanner variability and patient diversity.

In summary, the proposed architecture demonstrates a strong capability for brain stroke segmentation through the synergistic integration of transformer-based modeling, auxiliary supervision, and multi-objective optimization. This framework provides a promising foundation for future research aimed at enhancing clinical stroke diagnosis with reliable and interpretable deep learning systems.

7 Conclusion

This study presented a novel deep learning architecture for brain stroke segmentation that integrates a Swin Transformer backbone, auxiliary mask supervision, and a multi-head loss optimization strategy. The model effectively leverages the hierarchical self-attention mechanisms of the Swin Transformer to capture long-range dependencies and complex spatial patterns in brain MRI data. By incorporating an auxiliary branch guided by bounding box-based pseudo-segmentation masks, the framework enhances fine-grained feature learning, enabling the network to produce more accurate and spatially coherent lesion predictions. Furthermore, the multi-head loss design, encompassing Dice, MSE, and BCE components, supports robust training dynamics across segmentation, localization, and classification tasks. Experimental results on the ISLES 2024 dataset demonstrate the model's high performance in terms of segmentation accuracy, stroke detection confidence, and lesion localization. Visual examples confirm the model's ability to delineate stroke regions across a variety of complex cases. While the approach offers substantial improvements, future work may explore 3D volumetric integration and adaptation to heterogeneous clinical datasets to improve cross-domain generalization. Overall, the proposed framework contributes an effective and scalable solution for automated brain stroke analysis, holding potential for real-world clinical deployment in decision support systems and radiological workflows.

Acknowledgments: This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 "Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis".

References

[1] García-García, S., Cepeda, S., Arrese, I., & Sarabia, R. (2024). A fully automated pipeline using swin transformers for deep learning-based blood segmentation on head computed tomography scans after aneurysmal subarachnoid hemorrhage. World Neurosurgery, 190, e762-e773.

[2] Xie, H., Gu, H., Li, M., Zhu, L., Wang, T., Li, Z., & Wu, H. (2025). Carotid artery segmentation in computed tomography angiography (CTA) using multi-scale deep supervision with Swin-UNet and advanced data augmentation. Quantitative Imaging in Medicine and Surgery, 15(4), 3161.

[3] Li, T., An, X., Di, Y., Gui, C., Yan, Y., Liu, S., & Ming, D. (2024). SrSNet: accurate segmentation of stroke lesions by a two-stage segmentation framework with asymmetry information. Expert Systems with Applications, 254, 124329.

[4] Kirillov, K., Mikhailapov, D., Tulupov, A., & Berikov, V. (2024, September). Segmentation of 3D Non-Contrast CT Brain Images Using Transformer Neural Networks. In 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 173-177). IEEE.

[5] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In International MICCAI brainlesion workshop (pp. 272-284). Cham: Springer International Publishing.

[6] Wang, P., Yang, Q., He, Z., & Yuan, Y. (2023). Vision transformers in multi-modal brain tumor MRI segmentation: A review. Meta-Radiology, 1(1), 100004.

[7] Liu, Y., Ma, Y., Zhu, Z., Cheng, J., & Chen, X. (2024). TransSea: hybrid CNN-transformer with semantic awareness for 3D brain tumor segmentation. IEEE Transactions on Instrumentation and Measurement.

[8] Aboussaleh, I., Riffi, J., Fazazy, K. E., Mahraz, A. M., & Tairi, H. (2024). STCPU-Net: advanced U-shaped deep learning architecture based on Swin transformers and capsule neural network for brain tumor segmentation. Neural Computing and Applications, 36(30), 18549-18565.

[9] Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., & Li, J. (2023). W-Net: A boundary-enhanced segmentation network for stroke lesions. Expert Systems with Applications, 230, 120637.

[10] Chen, Y., Lu, X., & Xie, Q. (2023). Collaborative networks of transformers and convolutional neural networks are powerful and versatile learners for accurate 3D medical image segmentation. Computers in Biology and Medicine, 164, 107228.

[11] ElZemity, A., ElFdaly, M., Abdelfattah, S., Abdelwahab, A., Ramadan, M., Zakzouk, S., ... & Darweesh, M. S. (2023, November). A transformer-based deep learning architecture for accurate intracranial hemorrhage detection and classification. In 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 215-220). IEEE.

[12] Chen, Y. T., Huang, Y. C., Chen, H. L., Lo, H. C., Chen, P. C., Yu, C. C., ... & Lin, W. C. (2025). Automatic segmentation of white matter lesions on multi-parametric MRI: convolutional neural network versus vision transformer. BMC neurology, 25(1), 5.

[13] Lyu, J., Bartlett, P. F., Nasrallah, F. A., & Tang, X. (2024). Masked Deformation Modeling for Volumetric Brain MRI Self-supervised Pre-training. IEEE Transactions on Medical Imaging.

[14] Lyu, Y., & Tian, X. (2025). MWG-UNet++: Hybrid Transformer U-Net Model for Brain Tumor Segmentation in MRI Scans. Bioengineering, 12(2), 140.

[15] Zhang, Y., Xi, R., Wang, W., Li, H., Hu, L., Lin, H., ... & Liu, J. (2024). Low-Contrast Medical Image Segmentation via Transformer and Boundary Perception. IEEE Transactions on Emerging Topics in Computational Intelligence.

[16] Wu, B., Zhang, F., Xu, L., Shen, S., Shao, P., Sun, M., ... & Xu, R. X. (2023). Modality preserving U-Net for segmentation of multimodal medical images. Quantitative Imaging in Medicine and Surgery, 13(8), 5242.

[17] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., ... & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. Artificial intelligence in medicine, 102900.

[18] Wang, D., Shang, K., Liang, D., & Zhu, Y. (2024). Hybrid-ctunet: a double complementation approach for 3D medical image segmentation. International Journal of Machine Learning and Cybernetics, 1-16.

[19] Wang, Y., Yu, X., Guo, X., Wang, X., Wei, Y., & Zeng, S. (2023). A dual-decoding branch U-shaped semantic segmentation network combining transformer attention with decoder: DBUNet. Journal of Visual Communication and Image Representation, 95, 103856.

[20] Yan, Y., Liu, R., Chen, H., Zhang, L., & Zhang, Q. (2023). CCT-Unet: A U-shaped network based on convolution coupled transformer for segmentation of peripheral and transition zones in prostate MRI. IEEE Journal of Biomedical and Health Informatics, 27(9), 4341-4351.

[21] Li, J., Fan, J., Wang, Y., Yang, Y., & Zhang, Z. (2023). Coarse mask guided interactive object segmentation. IEEE Transactions on Image Processing, 32, 5808-5822.

[22] Hou, Q., Peng, Y., Wang, Z., Wang, J., & Jiang, J. (2023). Mfd-net: modality fusion diffractive network for segmentation of multimodal brain tumor image. IEEE Journal of Biomedical and Health Informatics, 27(12), 5958-5969.

[23] Liu, H., Hu, D., Li, H., & Oguz, I. (2023). Medical image segmentation using deep learning. Machine Learning for Brain Disorders, 391-434.

[24] Nizamani, A. H., Chen, Z., Nizamani, A. A., & Shaheed, K. (2023). Feature-enhanced fusion of U-NET-based improved brain tumor images segmentation. Journal of Cloud Computing, 12(1), 170.

[25] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., ... & Merhof, D. (2024). Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[26] Riedel, E. O., de la Rosa, E., Baran, T. A., Petzsche, M. H., Baazaoui, H., Yang, K., ... & Kirschke, J. S. (2024). ISLES 2024: The first longitudinal multimodal multi-center real-world dataset in (sub-) acute stroke. arXiv preprint arXiv:2408.11142.

[27] Liu, X., Du, P., Dai, Z., Yi, R., Liu, W., Wu, H., ... & Liu, J. (2024). SRTRP-Net: A multi-task learning network for segmentation and prediction of stereotactic radiosurgery treatment response in brain metastases. Computers in Biology and Medicine, 175, 108503.

[28] Hu, X., Li, Z., Wu, Y., Liu, J., Luo, X., & Ren, J. (2024). Neighbouring-slice guided multi-view framework for brain image segmentation. Neurocomputing, 575, 127315.

[29] Ge, P., Yan, T., Wong, P. K., Li, Z., Chan, I. N., Yu, H. H., ... & Gao, S. (2024). Simultaneous Segmentation and Classification of Esophageal Lesions Using Attention Gating Pyramid Vision Transformer. IEEE Transactions on Emerging Topics in Computational Intelligence.

[30] Neethi, A. S., Kannath, S. K., Kumar, A. A., Mathew, J., & Rajan, J. (2024). A comprehensive review and experimental comparison of deep learning methods for automated hemorrhage detection. Engineering Applications of Artificial Intelligence, 133, 108192.

[31] Huang, L., Jiang, B., Lv, S., Liu, Y., & Fu, Y. (2023). Deep-learning-based semantic segmentation of remote sensing images: A survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 8370-8396.

Chapter 24

Exploration of Textual Information in Points of Interest Recommendation

Contribution #48
Cite as:
Karla Giovanna Jiménez Enriquez, Ana Alves. Exploration of Textual Information in Points of Interest Recommendation. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-48,
  title ={Exploration of Textual Information in Points
     of Interest Recommendation},
  author={Karla Giovanna Jiménez Enriquez and  Ana
     Alves},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Exploration of Textual Information in Points of Interest Recommendation

Karla Giovanna Jiménez Enríquez[1] Ana Alves[2]

Coimbra Institute of Engineering (ISEC), Polytechnic University of Coimbra, Portugal
{a201910068a, aalves}@isec.pt

Abstract

The growing volume of data and widespread use of location-based services have highlighted the need for effective Point of Interest (POI) recommendation systems, particularly in fields such as tourism and urban navigation. This study addresses the challenge of enhancing POI recommendations by leveraging user interactions and rich textual information associated with locations. A hybrid recommendation model was developed, integrating collaborative and content-based filtering, and applied to a case study focused on Portugal. Textual data including POI descriptions, names, and categories were extracted from a location-based social network and processed using Natural Language Processing (NLP) techniques. The LightFM library served as the foundation for the system, with performance evaluated using metrics such as precision, recall, and AUC. The proposed method demonstrated strong performance, delivering personalized and relevant recommendations. These results underscore the value of incorporating textual information into recommendation models and contribute to the advancement of hybrid systems in the domain. While the study is geographically limited to Portugal, its methods are adaptable to other contexts, highlighting the potential for broader applicability and future research.

Keywords: Recommendation of Points of Interest (POI); Collaborative Filtering; Content-Based Filtering; Hybrid Recommendation Systems; Natural Language Processing (NLP); Machine Learning; Textual Information

1 Introduction

With the expansion of mobile technologies and the increasing availability of location-based data, Point of Interest (POI) recommendation systems have become central to enhancing user experiences in domains such like tourism and urban mobility. Traditional systems, however, often rely solely on check-in histories and geospatial data, underutilizing rich contextual information such as POI descriptions, names, and categories.

This reliance on structured data presents a significant drawback, especially when aiming to deliver personalized and context-aware recommendations. Recent studies emphasize the growing importance of incorporating unstructured textual data into recommendation models, yet few practical implementations have leveraged this approach in the context of POIs.

This research seeks to address that gap by designing a hybrid recommendation system that integrates collaborative and content-based filtering, enhanced with textual features obtained from a location-based social network. Portugal's linguistic diversity and geographic richness led to its selection as the application domain.

The main goals are to: (1) investigate how textual data can enrich POI recommendation models and (2) assess the performance of a hybrid system in comparison with existing approaches. The remainder of the paper is structured as follows: Section 2 reviews the literature, Section 3 details the methodology, Section 4 presents results, Section 5 discusses the findings, and Section 6 concludes.

2 State of the art

Point of Interest (POI) recommendation systems play a crucial role in helping users discover relevant locations, particularly in tourism, urban navigation, and social discovery platforms. These systems are commonly implemented within Location-Based Social Networks (LBSNs), which collect a wide range of user-generated data such as check-ins, photos, reviews, and geolocation. Traditionally, POI recommendation models have been built using collaborative or content-based filtering techniques. While collaborative filtering leverages user-item interactions to identify similar behavior patterns, content-based filtering relies on item metadata such as category, location, and tags.

However, both methods present significant limitations. Collaborative filtering suffers from the cold-start problem and data sparsity, particularly in sparse interaction matrices. Content-based models tend to produce limited novelty and diversity, often recommending items like those previously consumed. To overcome these issues, hybrid recommendation systems have emerged, combining the strengths of both approaches to improve accuracy and relevance.

Figure 1 illustrates the overall process of a recommendation system, emphasizing the importance of user input data in shaping the model. This highlights how personalization based on user actions and preferences is essential for the effectiveness of such systems.

(Gao et al., 2015) introduced graph-based representations to integrate social and geographical data, while (Guo et al., 2017) enhanced collaborative filtering using user trust networks. (Aliannejadi & Crestani, 2018) highlighted the importance of contextual and local preferences, which are often underrepresented in traditional systems. (Chang et al., 2018) proposed multi-source fusion methods that combine LBSNs and communication-based social networks (CBSNs) to enrich user profiles.



Figure 1 - Recommendation system scheme.

Despite advances in hybrid models, textual information such as POI names, descriptions, and user comments remains underutilized. (Werneck et al., 2020) noted that among 73 reviewed POI recommendation studies, only seven employed of textual features, signaling a critical gap in the literature. These unstructured data contain rich semantic context that can enhance recommendation relevance, especially in multilingual and dynamic environments.

Recommendation systems systems are increasingly incorporating Natural Language Processing (NLP. Techniques such as TF-IDF, word embeddings (e.g., all-MiniLM-L6-v2[1]), and topic modeling allow for the transformation of text into meaningful numerical vectors. (Xiong et al., 2020) explored the synergy between LBSNs and CBSNs, illustrating how communication metadata can complement spatial data. Additionally, approaches involving clustering algorithms like (Wang et al., 2017) have been applied to segment POIs based on geographic density, enabling more context-aware recommendations.

This study builds upon these prior contributions by proposing a hybrid recommendation model that integrates collaborative filtering with textual data via NLP techniques. In contrast to previous research, the suggested approach is assessed in a multilingual setting (Portugal) to showcase the advantages of textual enrichment for POI recommendation performance.

3 Methodology

Using structured and unstructured data from Points of Interest (POIs), this study presents a hybrid recommendation system that integrates collaborative and content-based filtering techniques. The methodology followed a step-by-step process: data acquisition, preprocessing, feature engineering, model development, and evaluation.

3.1 Data Acquisition

The dataset was collected from Foursquare[2], a Location-Based Social Network (LBSN). The goal was to extract POIs from Portugal to create a geographically focused use case. Data was obtained via Foursquare's API and included:

- Names, categories, and descriptions of POIs
- Geographic coordinates (latitude and longitude)
- User ratings and interactions (where available)

Custom Python scripts were developed to retrieve and structure the data, resulting in an initial dataset of 1,379 POIs.

3.2 Preprocessing and Textual Feature Engineering

After the initial data acquisition, a thorough cleaning process was conducted to remove duplicate, incomplete, or inconsistent records. The textual fields - POI names, categories, and descriptions - were specially considered due to their importance to the model's content-based component.

A multi-step pipeline for Natural Language Processing (NLP) was implemented to prepare the text for analysis. The language of each POI entry was first detected using Langdetected[3], allowing the dataset to be divided into Portuguese, English, and multilingual subsets and ensuring that language-specific patterns were preserved during training.

Next, standard preprocessing techniques were applied using SpaCy[3], including tokenization, lowercasing, stop-word removal, and lemmatization. These steps were essential for reducing noise and

[1] www.huggingface.co/sentence-transformers/all-MiniLM-L6-v2 *(last access on 2025-05-01)*

[2] www.foursquare.com *(last access on 2025-05-01)*

[3] www.spacy.io *(last access on 2025-05-01)*

normalizing the vocabulary across POIs. the SentenceTransformers[4] library library's all-MiniLM-L6-v2 model was then used to convert the cleaned text into numerical vector representations. This embedding process captured the semantic meaning of each POI's description, enabling a more nuanced similarity comparison between items.

In the implemented code, the language models en_core_web_md and pt_core_web_md from SpaCy were used to process textual information such as name, description, root category, and specific category. Each textual value was converted into a 300-dimensional vector, allowing for deeper semantic similarity analysis. For each entry in the dataset, a SpaCy Doc object was created to store the vector and additional attributes such as the norm (vector magnitude). In parallel, DBSCAN clustering was applied to the POI coordinates (latitude and longitude) to group nearby locations. This allowed the model to incorporate geographical context and focus recommendations within meaningful regional clusters.

3.3 Model Implementation

The recommendation system was implemented using LightFM[5], a hybrid model framework designed to support both collaborative and content-based filtering. The collaborative component was driven by user-item interaction data, while the content-based side utilized the previously generated text embeddings and metadata attributes such as category and location.

The model was trained using the Bayesian Personalized Ranking (BPR) objective function, which is optimized for ranking rather than rating prediction. This choice aligns with the real-world goal of prioritizing the most relevant POIs for each user, based on both their behaviour and item characteristics.

3.4 Evaluation Procedure

Initially, exploration data analysis was performed to understand the structure and characteristics of the dataset. The distribution of user–POI interactions, category frequency, and POI geographic concentration were all investigated using descriptive statistics, helping to identify patterns, such as the dominance of certain categories or regional preferences, which are critical for contextualizing recommendation outcomes.

Model Experiments

The LightFM model was subjected to multiple experimental setups to assess its behaviour under different linguistic and structural configurations. The dataset was segmented by language, producing separate subsets for Portuguese and English POIs and this enabling a direct comparison of model performance across monolingual and multilingual scenarios. Furthermore, cross-language experiments were conducted training the model in one language and testing in another to evaluate its generalization capacity across linguistic contexts.

To further refine model input, various combinations of textual features were tested: full feature sets (name, root category, specific category, and description), selected pairs (e.g., name and category, name and description), and isolated features. These tests allowed for a detailed assessment of which textual elements contributed most significantly to recommendation quality.

Additionally, continuous variables were discretized using binning techniques, which were adapted for each language segment. These bin intervals aimed to preserve linguistic nuances and enhance model training, particularly in dealing with sparsity and semantic variation.

[4] www.huggingface.co/sentence-transformers/all-MiniLM-L6-v2 *(last access on 2025-05-01)*

[5] www.making.lyst.com/lightfm *(last access on 2025-05-01)*

The system's effectiveness was measured using multiple ranking and relevance metrics:

- **Precision@k**: the fraction of recommended POIs in the top-k that are relevant
- **Recall@k**: the proportion of relevant POIs that are successfully retrieved within the top-k
- **MRR (Mean Reciprocal Rank)**: how soon the first relevant recommendation appears in the ranking
- **AUC (Area Under the ROC Curve)**: the model's ability to correctly distinguish relevant from irrelevant POIs

These metrics align with best practices in the evaluation of recommender (Aggarwal, 2016) and provided a multidimensional view of model performance. The technical implementation of the system relied on a variety of open-source tools and libraries, which are summarized in Table 1. These tools supported each project phase, from data acquisition and preprocessing to model training and evaluation.

Multilingual Considerations

Special attention was given to the multilingual nature of the dataset, predominantly composed of Portuguese and English entries. Experiments assessed whether linguistic segmentation improved the relevance and diversity of recommendations or if multilingual blending introduced semantic noise. Results showed that language segmentation could enhance contextual accuracy for specific user profiles, while multilingual data occasionally enabled broader generalization across semantically similar POIs in different languages.

Table 1 Tools and libraries used in the system implementation.

Tool	Purpose
Foursquare API	Data acquisition (POI names, descriptions, etc.)
SpaCy	NLP preprocessing (tokenization, lemmatization)
Langdetected	Language detection of POI text fields
SentenceTransformers	Embedding generation with all-MiniLM-L6-v2
DBSCAN	Spatial clustering of POIs
LightFM	Hybrid recommendation engine
Sweetviz	Data visualization and profiling
Google Colab / Spyder	Development and model testing environments

4 Results

The evaluation of the hybrid recommendation system was conducted through three different language-based models: multilingual, English-only, and Portuguese-only. These configurations allowed for the comparison of how linguistic characteristics affect recommendation performance.

3.5 Multilingual Model Performance

The multilingual model, trained with 11,003 interactions across 1,393 POIs, achieved balanced results. Precision@5 was 0.15 (train) and 0.13 (test), while Recall@5 reached 0.67 (train) and 0.61 (test). The AUC values remained consistently high—0.99 (train) and 0.98 (test)—demonstrating excellent discrimination

between relevant and irrelevant items. However, the MRR values were modest (0.36 and 0.32), suggesting that while relevant items were often recommended, they did not consistently appear at the top of the list.

3.6 English Model Performance

This model, trained with 923 interactions and 298 POIs, delivered the best overall performance. It achieved Precision@5 of 0.21 (train) and 0.20 (test), and Recall@5 of 1.00 (train) and 0.99 (test). The MRR of 0.92 (train) and 0.90 (test) confirmed that relevant recommendations were frequently ranked at the top. AUC was stable at 0.99 for both sets, reinforcing its robustness.

3.7 Portuguese Model Performance

Although this model included more data—5,582 interactions across 1,045 POIs—it showed the weakest performance. Both Precision@5 and Recall@5 were nearly zero. AUC was 0.60 (train) and 0.59 (test), and MRR stayed at 0.02, reflecting difficulties in capturing semantic patterns in Portuguese, possibly due to language complexity, orthographic variance, and less informative descriptions.

As shown in Figure 2, the multilingual model delivered balanced results, while the Portuguese model underperformed significantly across all metrics.

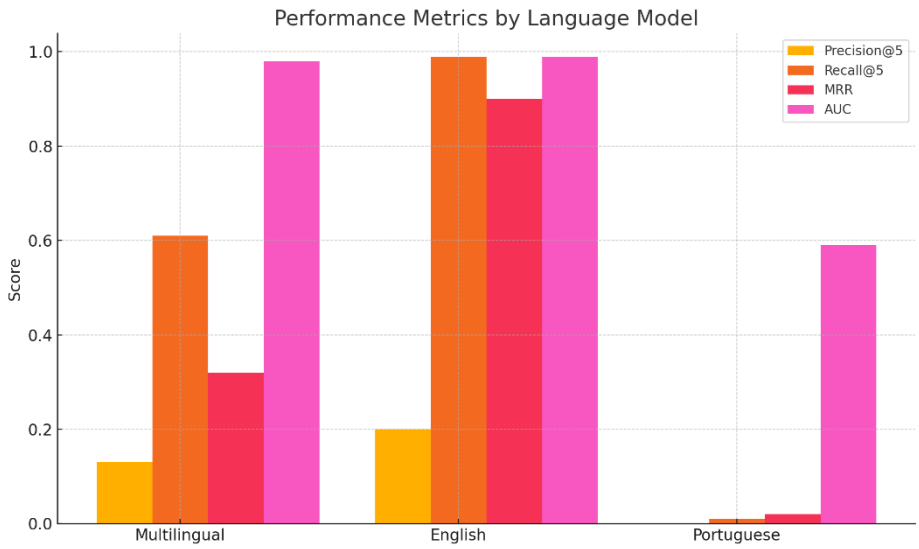


Figure 2 Performance metrics across multilingual, English, and Portuguese models.

5 Discussion

The experimental results offer valuable insights into the influence of language, text structure, and feature selection on the performance of POI recommendation systems. The findings confirm that textual information, when properly pre-processed and embedded, can significantly enhance recommendation accuracy, particularly in language-specific contexts.

The results also confirmed that the English-only model fared better than the others on all metrics. This superior

performance may be attributed to several factors: the consistency and richness of the English descriptions, the larger volume of pre-trained language models optimized for English, and possibly a more balanced semantic structure in the English POIs, highlighting the advantage of aligning training data with language-optimized embeddings and resources.

Even with a comparatively big dataset, the Portuguese-only model performed poorly on every evaluation metrics. This discrepancy suggests that Portuguese text may exhibit greater linguistic variability or that the available embeddings were less effective in capturing semantic nuance. Additionally, shorter or less descriptive POI texts in Portuguese may have hindered the model's ability to learn meaningful patterns. This reinforces the importance of dataset quality and text richness in non-English NLP contexts.

Interestingly, the multilingual model delivered stable and relatively high performance in both AUC and Recall@5, though it lagged in MRR and Precision. These results suggest that while the model could identify relevant POIs, it struggled to consistently rank the most relevant items at the very top. This effect might stem from semantic noise introduced by language mixing, particularly when word embeddings do not fully align across languages.

The cross-language experiments, which involved training the model in one language and testing in another, further demonstrated that language segmentation has a direct impact on generalization. Models trained in English showed moderate robustness when tested on multilingual or Portuguese data, while the inverse was not true. These results underscore the need for more robust multilingual embedding techniques or translation-aligned representations when deploying recommender systems in diverse linguistic environments.

Another critical observation was the impact of textual feature selection. Experiments that used combinations of name, category, and description performed better than those relying on isolated features. This confirms that different types of text contribute complementary semantic information, and their joint use increases the depth of representation available to the model.

Overall, the discussion supports the hypothesis that language-aware hybrid systems provide better performance but also highlights the challenges of working with under-resourced languages. Improving embeddings, enriching POI descriptions, and incorporating cross-lingual NLP techniques could be key directions for future improvements.

6 Conclusion

This study proposed a hybrid POI recommendation system that integrates collaborative filtering with content-based filtering, enriched using textual data and Natural Language Processing (NLP) techniques. By applying the model to a real-world, multilingual dataset from Portugal, the research demonstrated the potential and limitations of incorporating unstructured text into recommendation pipelines.

The results confirmed that language plays a critical role in recommendation performance. While the English-only model achieved the best accuracy, the Portuguese model faced challenges related to linguistic variability and data quality. The multilingual model offered a balanced alternative, though its performance was influenced by cross-language semantic inconsistencies. These findings reinforce the importance of considering linguistic context when designing and training recommendation systems.

Moreover, the experiments showed that combining different textual features such as names, categories, and descriptions enhanced recommendation quality more effectively than isolated inputs. This underscores the value of rich, semantically diverse data in improving system relevance and personalization.

Despite promising results, the study has some limitations. It relies on data from a single platform (Foursquare) and is geographically restricted to Portugal. These constraints open several directions for future research.

Future Work

A key avenue for expansion is the geographic broadening of the dataset to include other Portuguese-speaking countries, such as Brazil, Angola, and Mozambique. These regions have distinct cultural, touristic, and behavioural patterns that could enrich the recommendation engine's diversity and adaptability.

Integrating data from other platforms such as Google Maps, TripAdvisor, or Yelp could also provide more up-to-date and comprehensive information, including user reviews, images, operating hours, and real-time ratings. This would further increase the relevance and dynamism of recommendations.

Additionally, leveraging generative language models (e.g., ChatGPT) presents an opportunity to move beyond simple item lists toward explanatory and contextual recommendations. These models could generate personalized justifications for each suggestion, increasing user trust and engagement.

Another promising direction involves the use of Multimodal Machine Learning (MML), combining text, images, and user behaviour data to generate richer, culturally contextualized descriptions of POIs. Incorporating contextual variables such as time of day, season, or weather conditions could also help tailor recommendations to the user's current situation.

Finally, building interactive systems that allow users to refine preferences in real-time would enhance user agency and satisfaction. As data technologies, generative AI, and machine learning algorithms continue to evolve, recommender systems are becoming more adaptive, dynamic, and intelligent capable of delivering not only relevant suggestions but also a more personalized and immersive experience.

References

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook* (1st ed.). Springer Publishing Company, Incorporated.

Aliannejadi, M., & Crestani, F. (2018). Personalized Context-Aware Point of Interest Recommendation. *ACM Transactions on Information Systems (TOIS)*, *36*(4). https://doi.org/10.1145/3231933

Chang, B., Park, Y., Park, D., Kim, S., & Kang, J. (2018). Content-aware hierarchical point-of-interest embedding model for successive POI recommendation. *IJCAI International Joint Conference on Artificial Intelligence*, *2018-July*, 3301–3307. https://doi.org/10.24963/IJCAI.2018/458

Gao, H., Tang, J., Hu, X., & Liu, H. (2015). Content-Aware Point of Interest Recommendation on Location-Based Social Networks. *Proceedings of the National Conference on Artificial Intelligence*, 1721–1727. https://doi.org/https://dl.acm.org/doi/10.5555/2886521.2886559

Guo, Q., Sun, Z., Zhang, J., Chen, Q., & Theng, Y. L. (2017). Aspect-aware point-of-interest recommendation with geo-social influence. *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, 17–22. https://doi.org/10.1145/3099023.3099066

Wang, H., Fu, Y., Wang, Q., Yin, H., Du, C., & Xiong, H. (2017). A location-sentiment-aware recommender system for both home-town and out-of-town users. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F129685*, 1135–1143. https://doi.org/10.1145/3097983.3098122

Werneck, H., Silva, N., Viana, M. C., Mourão, F., Pereira, A. C. M., & Rocha, L. (2020). A Survey on Point-of-Interest Recommendation in Location-based Social Networks. *ACM International Conference Proceeding Series*, 185–192. https://doi.org/10.1145/3428658.3430970

Xiong, X., Qiao, S., Han, N., Xiong, F., Bu, Z., Li, R. H., Yue, K., & Yuan, G. (2020). Where to go: An effective point-of-interest recommendation framework for heterogeneous social networks. *Neurocomputing*, *373*, 56–69. https://doi.org/10.1016/J.NEUCOM.2019.09.060

Chapter 25

Storm damage and planting success assessment in Pinus pinaster stands using Mask-RCNN

Contribution #50
Cite as:
Raúl Salas-González, Ivon Brandão, Beatriz Fidalgo, Alexandra Amaral. Storm damage and planting success assessment in Pinus pinaster stands using Mask-RCNN. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-50,
  title ={Storm damage and planting success assessment
     in Pinus pinaster stands using Mask-RCNN},
  author={Raúl Salas-González and  Ivon Brandão and
     Beatriz Fidalgo and  Alexandra Amaral},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Storm damage and planting success assessment in *Pinus pinaster* stands using Mask-RCNN

Raúl Salas-González[1,2], Ivon Brandão[2] *, Beatriz Fidalgo[2] Alexandra Amaral[2]
[1] RCM2+ Research Centre for Asset Management and Systems Engineering, ISEC/IPC, Rua Pedro Nunes, 3030-199 Coimbra, Portugal
[2] Polytechnic Institute of Coimbra, Coimbra Agriculture School, Bencanta, 3045-601 Coimbra, Portugal

*Corresponding author
ivon.brandao@esac.pt

Abstract

This study investigates the application of deep learning neuronal techniques, specifically the Mask R-CNN architecture, for the automatic detection of trees in *Pinus pinaster* stands using RGB and multispectral imagery captured by a drone. The research addresses two distinct forest scenarios, resulting from disturbances intensified by climate change. The first concerns the detection of fallen trees following an extreme weather event, to support damage assessment and inform post-disturbance forest management. The second focuses on detecting trees in a recently established plantation after wildfire, to evaluate the effectiveness of ecological restoration efforts. The collected images were processed to generate high-resolution orthophotos and orthomosaics, which were used as input for tree detection using Mask R-CNN. Preliminary results suggest that integrating drone-based imagery with deep learning models can significantly enhance the efficiency of forest assessments, reducing the need for fieldwork effort and increasing the reliability of the collected data. Results demonstrated high detection rates, identifying more than 90% of fallen trees and 75% of recently planted trees while also allowing for the extraction of spatial metrics relevant to forest monitoring. Overall, the proposed methodology shows strong potential for rapid response in post-disturbance environments and for monitoring the early development of forest plantations.

Keywords: Mask R-CNN tree detection; Drone forestry monitoring; Fallen tree detection; Plantation success assessment; Deep learning in forestry

1 Introduction

Monitoring forest stands is essential for sustainable forest management providing critical information on forest structure. Information on forest structure, species composition, and regeneration, support operational planning, evaluating ecosystem health and ensuring forest resilience to climate change [1], [2]. Under current climate change scenarios, Europe is projected to experience a temperature increase ranging from 2.4 °C in optimistic projections to up to 6 °C in the worst-case scenarios. These climatic shifts are expected to intensify forest disturbances, including more severe and frequent droughts, pest outbreaks, wildfires, and extreme weather events.

Portugal, with its Mediterranean climate, is particularly vulnerable to these extremes and has witnessed increasingly intense wildfires in recent decades, notably in 2003 and 2017 [3]. Consequently, post-fire restoration has become a national priority, and assessing the success of reforestation and natural regeneration is critical for improving recovery strategies. Alongside wildfires, the rising incidence of extreme weather events such as storms has further disrupted forest stands, causing structural damage and tree mortality. These disturbances reduce timber value and impair key ecosystem services such as carbon sequestration, reinforcing the need for advanced tools to monitor forest condition and resilience under climate change [4].

Traditionally, forest inventories are conducted through field-based surveys, which offer high levels of accuracy but are often limited in spatial coverage, time efficiency, and accessibility,

especially in the aftermath of storms or in recently planted stands located in rugged terrain. Recent advances in remote sensing, particularly the use of unmanned aerial vehicles (UAVs) equipped with high-resolution RGB and multispectral sensors, have provided a more flexible and cost-effective alternative for forest monitoring [5], [6]. UAVs enable fast and repeated data acquisition over targeted areas, facilitating the timely detection of structural changes and vegetation conditions. [7].

When integrated with artificial intelligence, particularly deep convolutional neural networks (CNNs), such as Mask R-CNN [8] and YOLOv8 [9], UAV imagery can support automated detection of individual trees, fallen trunks, or plantation gaps, with minimal human intervention and high spatial precision [10], [11], [12], [13].

Recent studies have demonstrated the potential of deep learning to automate the assessment of forest damage following storms or tornadoes. For instance, Nasimi and Wood [11] employed a deep learning approach using YOLOv8x-seg to detect tornado-induced tree falls in forested regions of Kentucky and Tennessee. Their method achieved a mean average precision of more than 80% in instance segmentation of fallen trees from 2 cm resolution UAV imagery, illustrating the effectiveness of 2D image analysis for post-disturbance mapping.

Treefall detection is not the only application where this approach is valuable. According to Guo et al. [12], UAV imagery with ultrahigh spatial resolution is effective for recognising young trees in complex plantation environments, even where overlapping vegetation and irregular spacing are present. Candiago et al. [13] also demonstrated the utility of vegetation indices (NDVI, GNDVI) for detecting gaps in plantations. However, the combination with deep learning algorithms is what enables individual tree identification at scale, making this process both accurate and automated [14].

Despite these advances, applying deep learning in forestry is not without challenges. One of the most significant is the need for large, annotated datasets to train the models, especially given the variability in forest structures, species, lighting conditions, and terrain. Diez et al. [15] emphasised that although deep learning models generally outperform traditional approaches in image-based forestry applications, generalization across forest types, canopy structures, and illumination conditions remains a major challenge. Miao et al. [16] also found that although multispectral UAV classifiers achieved over 93% accuracy, active learning was required to generalise across seasonal conditions.

Building on previous research that has demonstrated the value of integrating UAV imagery and deep learning for forest monitoring, this work focuses on the specific application of these methods to Pinus pinaster stands. It addresses two key research questions: (i) Can Mask R-CNN accurately detect fallen trees in post-storm conditions using UAV imagery? (ii) Is it possible to identify individual young trees in recently planted maritime pine stands using multispectral UAV data? By addressing these questions, the study aims to contribute to the development of automated, scalable tools to support forest monitoring, post-disturbance assessment, and restoration planning in Mediterranean pine-dominated landscapes.

The remainder of this paper is structured as follows. Section 2 presents a review of the state of the art in remote sensing and deep learning applications in forestry. Section 3 describes the study areas, UAV platforms, data processing, and model training procedures. Section 4 reports and discusses the results of the two case studies. Finally, Section 5 summarises the main findings and outlines directions for future work.

2 State of the art

Mask R-CNN is one of the most widely used deep learning architectures for object detection in forestry applications [8], [17], [18]. This model detects objects in an image and simultaneously generates a segmentation mask for each instance. It is particularly suited for individual tree detection due to its ability to handle overlapping crowns and produce spatially precise outputs. Kislov and Korznikov [19] applied Mask R-CNN to satellite RGB orthophoto imagery and a Mask R-CNN model trained on a manually annotated dataset for detecting windthrown trees in boreal forests. The model achieved over 85% precision and recall in detecting fallen trees, outperforming manual photo interpretation and significantly reducing processing time.

Han et al. [14] used a similar approach to detect individual trees in plantations and orchards. using RGB drone images processed with a CNN-based object detection framework. Results showed a detection accuracy above 90% under varying canopy density and lighting conditions.

In more recent research, Yao et al. [20], used RGB and DSM layers to distinguish live and dead trees in mixed forests, highlighting the potential of combining spatial structure with instance segmentation.
These works also demonstrate that, while traditional spectral analysis has value, the inclusion of deep learning enables fine-scale, object-based monitoring that aligns better with forest management goals. For example, Morakot et al. [21] employed deep networks (Faster R-CNN and Mask R-CNN) on UAV RGB images to detect and classify coconut trees in plantations with high accuracy, highlighting the potential of these approaches for individual tree-level monitoring

YOLO (You Only Look Once) is a real-time object detection framework known for its computational efficiency. Nasimi and Wood [11] investigated the application of YOLOv8x-seg for mapping tornado-induced treefall in forests across Kentucky and Tennessee. Results show over 80% mean average precision (mAP) in detecting and vectorising fallen trees. The method reduced the need for time-consuming field assessments and enabled near-real-time storm damage estimation. Similarly, recent advances in lightweight architectures, such as YOLO-UFS [22] and RSD-YOLOv8 [23], have extended their applicability to real-time fire and pest detection in forestry, enabling deployment on low-power edge devices.

While the application of deep learning in forestry has demonstrated considerable potential, several methodological challenges remain unresolved. One of the primary limitations concerns data requirements, as most deep learning models demand large volumes of annotated training data—typically ranging from hundreds to thousands of labelled instances. However, the forestry sector continues to lack open-access, standardised datasets suitable for robust model development and validation, as noted by Diez et al. [15]

A second major challenge is model generalization. Predictive performance often deteriorates when models are applied across forest types that differ in species composition, canopy density, or environmental conditions. Diez et al. [15] highlight this issue, and Soto-Vega et al. [24] similarly observed reduced generalisation in tropical plantations when convolutional neural networks (CNNs) trained on temperate forest data were transferred without adaptation.

A further complication relates to the influence of overlapping tree crowns and shadow effects, which remain problematic, particularly in dense stands, and negatively affect the performance of object detection frameworks such as Mask R-CNN and YOLO-based models [25]. Yao et al. [20] addressed this issue by incorporating digital surface model (DSM) layers, which helped reduce the misclassification of overlapping crowns.

Another underexplored yet promising avenue lies in the integration of multispectral data. While models trained solely on RGB imagery can achieve satisfactory results, especially when combined with segmentation algorithms and vegetation indices (VIs), the fusion of spectral, structural, and contextual features could further enhance model robustness and transferability. Allen et al. [26] demonstrated that RGB-based monitoring can yield high predictive accuracy when supported by VIs, confirming its potential for low-cost applications. Nevertheless, other studies [11], [12] underscored the added value of incorporating multispectral or contextual information for improving classification performance under variable conditions.

Addressing these challenges will require coordinated efforts to develop annotated datasets, adopt domain adaptation strategies, and integrate spectral and spatial data sources within deep learning frameworks for forest monitoring.

3 Methodology

This study was conducted in two pilot sites located in central Portugal (Figure 1), both integrated in the *Transform Agenda* under the *Better Forests Program*, which aims to promote climate-resilient forest landscapes. The selected study areas included mixed stands dominated by *Pinus pinaster*, either mature or recently planted, allowing the evaluation of two distinct forest monitoring challenges: detection of fallen trees following an extreme weather event, and identification of plantation gaps in early-stage forest regeneration.

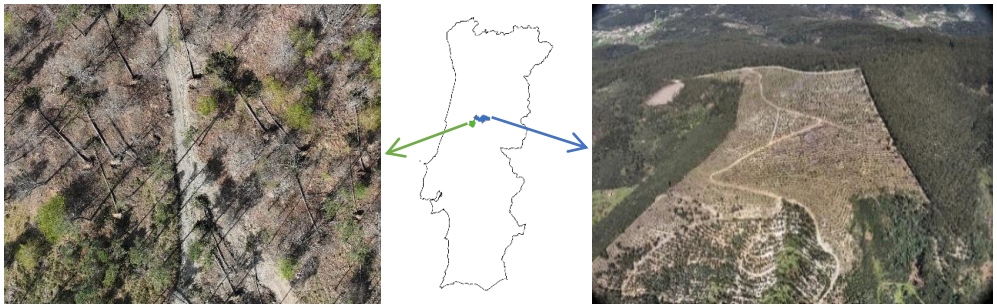


Figure 1 – Pilot site in Serra da Lousã / Pilot site in Serra do Açor

In Serra da Lousã, the focus was on detecting windthrown trees caused by storm Martinho (March 2025) in a mature stand composed primarily of *Pinus pinaster*, with presence of *Castanea sativa*, *Pseudotsuga menziesii* and *Quercus robur*. This stand is part of forest area established by the Portuguese Forest Service and now managed by ICNF, characterized by a transition from monocultures to more structurally and ecologically diverse forests. The second site, located in Serra do Açor (municipality of Arganil), includes over 2,500 hectares of community-managed land undergoing ecological restoration after the 2017 wildfires. The selected stand served as a case study for detecting planting failures and evaluating the spatial distribution and density of newly planted trees, particularly *Pinus pinaster* and *Quercus robur*.

A DJI Mavic 3M UAV was employed for aerial data acquisition. This drone integrates RGB and multispectral cameras, with real-time kinematic (RTK) support and multi-constellation GNSS (GPS, Galileo, BeiDou, GLONASS), ensuring high precision georeferencing. Flights were conducted at a fixed altitude of 120 meters, with 80% forward and side overlap, and nadir camera orientation (−90° tilt). In both sites, flight parameters were optimized to capture consistent and high-resolution imagery suitable for orthoimage generation and object detection.

Field data were collected to serve both georeferencing and model validation. A Spectra Precision SP60 GNSS antenna and a MobileMapper 50 device were used to acquire ground control points (GCPs) and reference tree positions. Tree biometric data, including diameter at breast height (DBH) and height, were obtained using a diameter tape and a TruPulse 200B laser rangefinder. These field measurements supported model evaluation through spatial accuracy and detection metrics.

All imagery was processed in ESRI Drone2Map software, generating orthophotos (Figure 2). RTK corrections were applied during processing to improve image location accuracy. The “Initial Image Scale” and “Matching Neighborhood” parameters were adjusted to enhance tie-point generation. GCPs

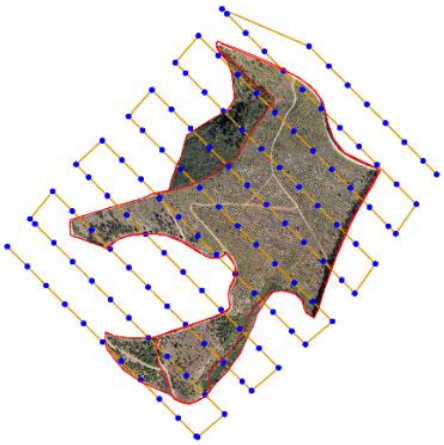


Figure 2 - Orthophoto created with imagery from a recently established plantation

were manually marked to further refine orthoimage alignment. Final products were projected to ETRS89 / PT-TM06 and exported as GeoTIFFs for use in ArcGIS Pro.

The deep learning workflow involved two stages: manual annotation and model training. Using ArcGIS Pro's "Label Objects for Deep Learning" tool, objects of interest (fallen trees and planted seedlings) were delineated as polygons on orthophotos. These labeled datasets were exported with the "Export Training Data for Deep Learning" tool, which produced tiled imagery (typically 76×76 to 512×512px) and metadata in RCNN Masks format. These tiles formed the input for training instance segmentation models based on the Mask R-CNN architecture.

For fallen tree detection, 10 different training datasets were prepared, each including between 29 and 40 annotated trees. Over 28,000 image tiles were produced, encompassing more than 26,000 tree instances. The datasets varied in object density (1.13 to 2.46 trees per tile) and mask size (2.30 to 24.31 units), capturing the variability in tree crown shapes, illumination, and background complexity. The most robust dataset included 13,308 tiles and 26,950 instances. For plantation gap detection in Serra do Açor, training datasets included between 161 and 534 labeled trees per set, generating over 66,000 images and 126,000 annotated instances. The largest dataset contributed 20,006 images and 72,674 trees. Smaller sets with higher tree densities (up to 12.73 trees/tile) were also included to improve model performance in densely planted areas. Object sizes were generally small (1.31 to 2.31 units), matching the typical scale of young trees.

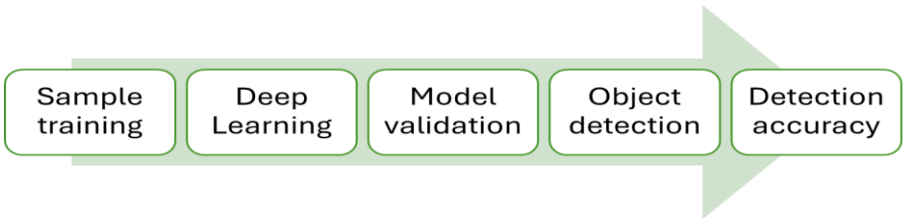


Figure 3 – Data processing pipeline

All models were trained using the "Train Deep Learning Model" tool in ArcGIS Pro, using a ResNet-50 backbone. Model performance was evaluated through training/validation loss curves and average precision (AP) scores. The trained models were then applied to the full orthophotos using the "Detect Objects Using Deep Learning" tool, with adjusted parameters such as confidence threshold (0.6–0.9), test-time augmentation (TTA), and non-maximum suppression (NMS) to optimize performance (Figure 3). Ground-truth data were used to compute precision, recall, F1-score, and intersection-over-union (IoU) as equation 1, which served as the principal evaluation metric for segmentation accuracy.

$$IoU = \frac{A_intersection}{A_union} \qquad (1)$$

Where A_intersection is the overlapping area between predicted and reference masks, and A_union is the total combined area.

4 Results

Fallen tree detection

The best-performing configuration was identified as model_metrics9 (Figure 4), which applied a learning rate slice of 7.5858×10^{-6} to 7.5858×10^{-5}. This model achieved the highest precision score among all experimental runs, with a value of 0.91, indicating a strong ability to correctly segment tree crowns relative to the total number of predicted instances. Notably, the model performed robustly under visually challenging conditions, including heterogeneous canopy structures and complex background textures. These results confirm both the effectiveness of the selected architecture and the adequacy of the training strategy and underscore the importance of a well-annotated dataset in supporting accurate instance segmentation in post-disturbance forest environments (Figure 5).

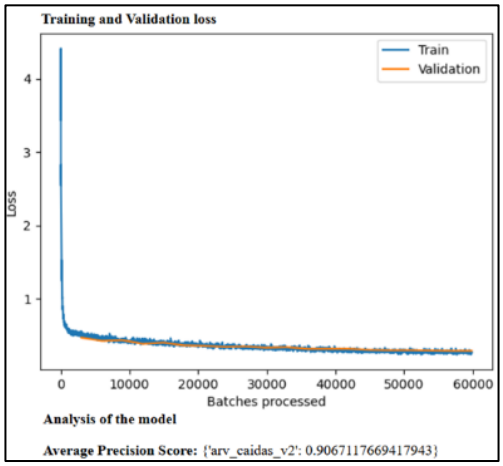


Figure 4 - Results of the "model_metrics9" MaskRCNN / Backbone: resnet50

In terms of operational performance, the best-performing model configuration—trained using the previously established labelled components—was evaluated not only for its detection accuracy but also for computational efficiency. The total training time amounted to 12 hours and 5 minutes, reflecting the size of the dataset and the fine spatial resolution adopted (tile size of 400, stride of 96, and 20% rotation augmentation). Despite the relatively long training phase, the model demonstrated efficient inference capabilities: the object detection stage required approximately 3 minutes and 12 seconds per sample, while the subsequent "label components" process was completed in just 5 minutes and 40 seconds. These results indicate that the model can be feasibly applied in operational contexts without prohibitive computational costs.

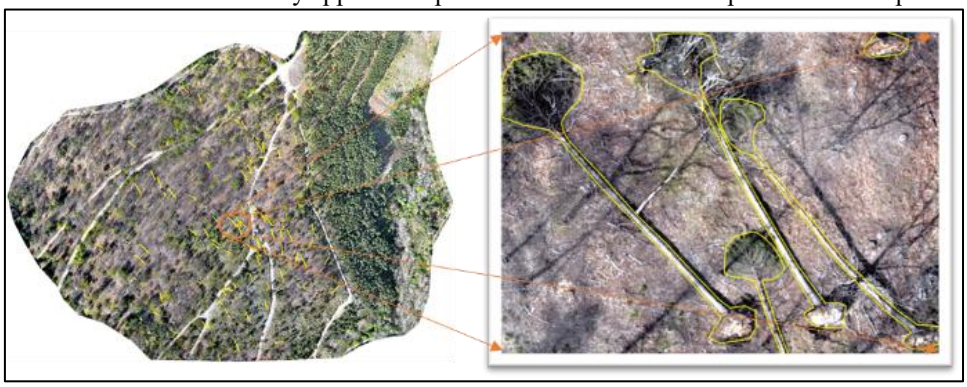


Figure 5 - Fallen trees detection with deep learning

Additional post-processing procedures, including non-maximum suppression (NMS) and test-time augmentation, were also implemented and evaluated. These techniques yielded acceptable trade-offs between computational cost and spatial accuracy, further refining the model's performance. Overall, the results indicate that, beyond achieving a high precision score (0.91), the model demonstrates computational viability and is well-suited for near-operational deployment in post-disturbance forest monitoring applications.

Planting failures

The best-performing model, designated RCNN_V14_76_30_20rot (Figure 6), achieved an average precision score of 0.7479. This configuration employed a learning rate slice of 9.1201×10^{-6} to 9.1201×10^{-5} and integrated rotation-based data augmentation techniques, which contributed to enhanced generalisation across heterogeneous spatial configurations and complex background textures.

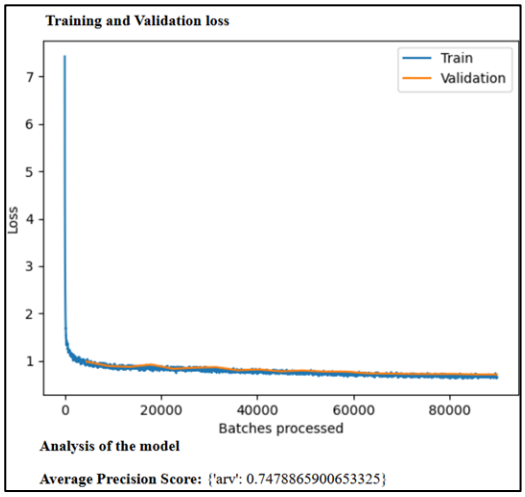


Figure 6 - Results of the "model_metrics14 - RCNN_V14_76_30_20rot MaskRCNN / Backbone: resnet50

Other competitive models included RCNN_V15_192_64_20rot (precision = 0.7163) and RCNN_v10_v4 (precision = 0.6813) (Figure 7), further illustrating the sensitivity of model accuracy to both hyperparameter tuning and input preprocessing strategies such as image resolution and augmentation. By contrast, RCNN_v10 achieved the lowest precision score (0.3504), highlighting the critical role of appropriate architectural tuning and parameter configuration. Collectively, these findings reinforce the suitability of the Mask R-CNN architecture, for instance, segmentation tasks in post-disturbance plantation monitoring. Moreover, they emphasize the importance of a well-annotated dataset that can capture spatially heterogeneous canopy patterns, thereby enabling effective model training and evaluation in applied forestry contexts.

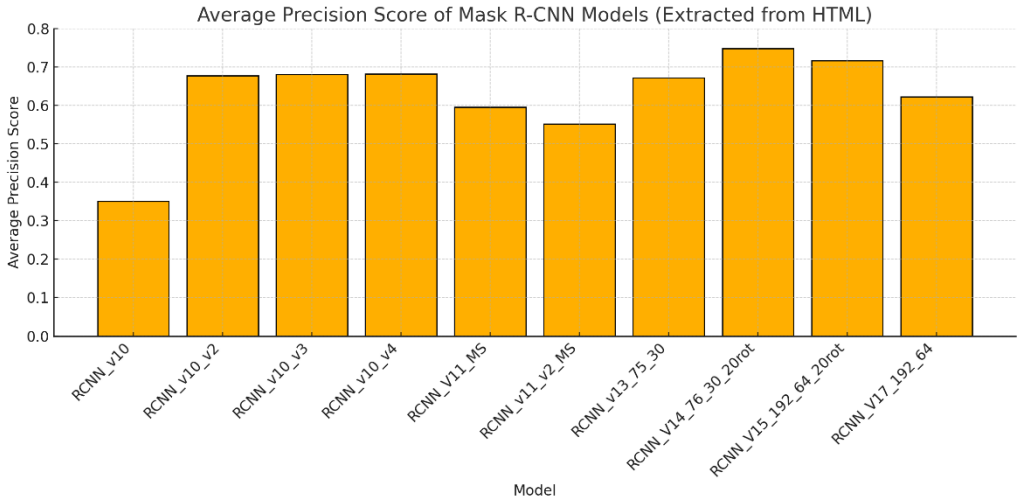


Figure 7 - Mask RCNN models precision comparison

In terms of operational feasibility, the top-performing model for detecting planting failures, RCNN_V14_76_30_20rot, was also evaluated for computational performance. The final training process required approximately 9 hours and 46 minutes, largely due to the high-resolution input configuration (tile size 76, stride 30, with 20% rotation) and the generation of over 49,000 image tiles. Despite the extended training

duration, inference remained computationally manageable: the object detection stage, incorporating both non-maximum suppression (NMS) and test-time augmentation (TTA), required approximately 1 hour and 14 minutes, while the label assignment phase was completed in just 4 minutes and 16 seconds per sample.

Other model configurations presented similar performance trade-offs. For instance, **RCNN_V15_192_64_20rot** completed training in approximately 2 hours and required 7 minutes and 13 seconds for object detection. Although more efficient computationally, this model achieved a slightly lower precision score (0.716) compared to **RCNN_V14_76_30_20rot** (0.748). These results underscore the operational feasibility of deep learning-based gap detection in forestry applications, while also highlighting the critical role of tile configuration and augmentation strategies in shaping both model accuracy and processing time. Striking an appropriate balance between computational efficiency and segmentation performance is essential for the large-scale and practical deployment of such models in forest monitoring workflows.

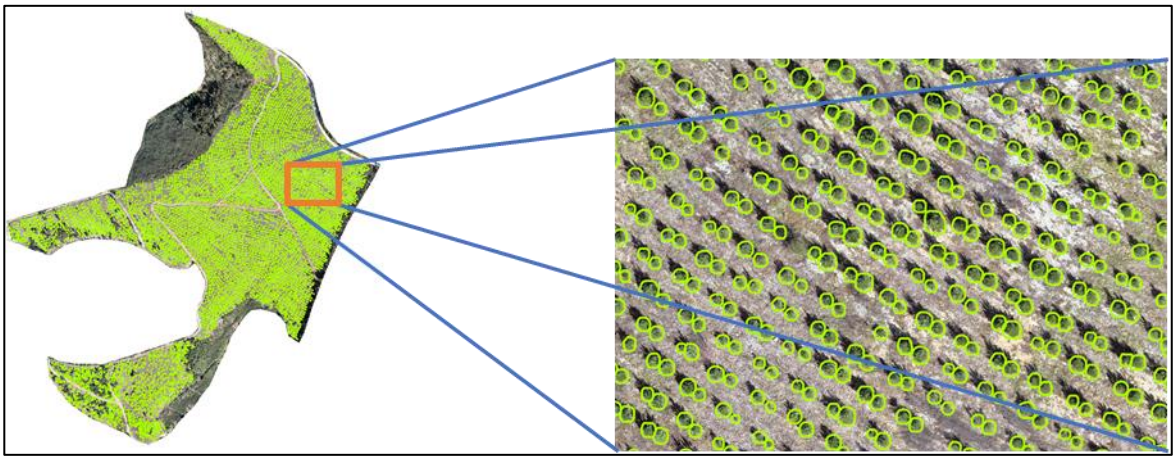


Figure 8 - Tree detection on a recent plantation using Mask-RCNN

This analysis finalises the evaluation of instance segmentation models for detecting planting failures in post-disturbance restoration contexts. Compared to fallen-tree detection, this task demanded finer spatial granularity in both annotation and model resolution, due to the smaller size and more regular spatial arrangement of planting gaps (Figure 8). Despite these challenges, the results demonstrate that Mask R-CNN architectures are capable of adapting to such ecological conditions while maintaining computational scalability and operational relevance.

5 Discussion

This study explored the application of Mask R-CNN architectures for detecting fallen trees and planting failures using high-resolution RGB drone imagery in *Pinus pinaster* forest stands located in the Serra da Lousã and Serra do Açor, respectively. The results demonstrate that the proposed approach achieves high detection accuracy, with over 90% of fallen trees correctly identified in the storm-damaged area and approximately 75% of the planted saplings detected in the young plantation. These outcomes are consistent with those of Han et al. [13], who reported similar levels of accuracy in automatic tree counting using UAV data and deep learning models. Moreover, this method significantly reduces the need for intensive field campaigns, increases the temporal resolution of monitoring, and allows integration into forest management decision systems. The findings confirm that deep learning models can effectively identify forest anomalies with high spatial resolution and precision, even in structurally complex natural environments.

In comparison with current state-of-the-art methods, the proposed approach demonstrates competitive performance. Precision, selected as the primary performance indicator due to its relevance in minimizing false positives in ecological applications, exceeded values reported in comparable studies. For instance, Kislov and

Korznikov [19] employed high-resolution satellite imagery to detect windthrow, achieving only moderate accuracy in heterogeneous forest conditions. By contrast, the drone-based Mask R-CNN model presented here attained superior precision in fallen tree detection, likely attributable to the higher spatial granularity of UAV imagery and the architecture's capacity for class-specific segmentation. Similarly, the results for detecting planting gaps are consistent with those of Morakot et al. [21], who used Mask R-CNN to classify coconut trees from UAV imagery. The model successfully delineated planting gaps with high spatial accuracy, even when they were irregularly shaped or partially obscured by understory vegetation and shadowing. These results suggest that the method is well-suited for identifying structural anomalies in both monoculture and mixed-species forest plantations.

A further dimension explored in this study relates to the sensitivity of Mask R-CNN performance to training dataset size. By systematically varying training set sizes, we observed that model precision remained relatively stable beyond a minimal threshold, reinforcing the findings of Weinstein et al. [25] and Soto-Vega et al. [24], who emphasize the importance of dataset diversity over absolute volume.

This study also directly addresses the research questions formulated in the introduction. First, it demonstrates that the Mask R-CNN architecture can reliably detect both fallen trees and planting gaps using RGB drone imagery in *Pinus pinaster* stands. Second, it confirms that the model maintains robust performance even when trained on datasets of varying sizes, provided that appropriate domain-specific data augmentation strategies are applied. Third, it offers empirical evidence that UAV-based deep learning systems can be effectively implemented in operational forestry contexts, enabling large-scale, repeatable, and cost-efficient monitoring of structural anomalies in forest cover.

Nonetheless, certain limitations remain. The model's precision tends to decline in areas characterised by severe shadowing, overlapping canopy structures, or dense understory vegetation, where target features may be partially or entirely obscured. Recent studies suggest that these limitations could be mitigated through the integration of complementary data sources, such as LiDAR, which can provide additional structural information to improve detection accuracy. Furthermore, although the model demonstrated good generalisation across two distinct forest sites, additional validation in diverse ecological contexts and under varying seasonal conditions is necessary to fully assess its robustness and transferability.

6 Conclusion

This study addressed the challenge of detecting fallen trees and planting failures in *Pinus pinaster* forests, an essential component of timely forest management that has traditionally depended on manual, labour-intensive fieldwork. To overcome these constraints, we employed a deep learning approach based on the Mask R-CNN architecture applied to high-resolution RGB imagery acquired by UAVs. The methodology was tested in two ecologically and topographically distinct forest sites, achieving high detection precision in both scenarios.

The principal contribution of this research lies in demonstrating the feasibility of accurately identifying and mapping fallen trees and planting gaps through UAV-based deep learning, generating outputs suitable for integration into operational forest monitoring workflows. Compared to conventional satellite-based methods or manual surveys, the proposed approach offers superior spatial resolution, reduced field effort, and greater temporal responsiveness, key advantages for adaptive forest management under increasingly dynamic disturbance regimes driven by climate change. In this context, developing more efficient methods for continuous forest assessment, as they enable timely and spatially explicit monitoring, is an essential foundation for adaptive forest management and long-term resilience planning in the face of ongoing climatic pressures.

Despite these strengths, certain limitations were observed. Model performance declined in areas with severe shadowing, overlapping canopy layers, or dense understory vegetation, where target features are visually obscured. In addition, while the model generalised well across the two study sites, its broader applicability to other forest types, structural conditions, and seasonal variability remains to be validated. Future work should explore the integration of complementary data sources such as LiDAR and expand testing across diverse bioclimatic regions to assess model robustness and support wider adoption in forest monitoring and restoration planning.

Acknowledgements: This work is financed by the European Union – NextGenerationEU, under the scope of PRR - Portugal's Recovery and Resilience Plan, AAC n.º02/C05-i01/2022 - Agendas for Business Innovation, project n.º34 TRANSFORM - Digital transformation of the forestry sector for a more resilient and low-carbon economy.

References

[1] A. M. Lechner, G. M. Foody, and D. S. Boyd, "Applications in remote sensing to forest ecology and management," *One Earth*, vol. 2, no. 5, pp. 405–412, May 2020, doi: 10.1016/j.oneear.2020.05.001.

[2] M. Wegler and C. Kuenzer, "Potential of earth observation to assess the impact of climate change and extreme weather events in temperate forests—A Review," *Remote Sensing*, vol. 16, no. 12, pp. 2224, Jun. 2024, doi: 10.3390/rs16122224.

[3] M. G. Pereira, R. M. Trigo, C. C. da Camara, J. M. C. Pereira, and S. M. Leite, "Synoptic patterns associated with large summer forest fires in Portugal," *Agricultural and Forest Meteorology*, vol. 129, no. 1–2, pp. 11–25, Mar. 2005, doi: 10.1016/j.agrformet.2004.12.007.

[4] R. Salas-González and B. Fidalgo, "Impacto de agentes de distúrbio nos serviços dos ecossistemas em povoamentos de pinheiro bravo na Serra da Lousã," in Geografia, Riscos e Proteção Civil: Homenagem ao Professor Doutor Luciano Lourenço, vol. 2, A. N. Nunes, J. C. Tavares and F. M. Ribeiro, Eds. Coimbra: RISCOS – Associação Portuguesa de Riscos, Prevenção e Segurança, 2021, pp. 213–224. ISBN 978-989-9053-05-2

[5] L. Pádua, T. Adão, N. Guimarães, A. Sousa, E. Peres, and J. J. Sousa, "Post-fire forestry recovery monitoring using high-resolution multispectral imagery from unmanned aerial vehicles," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-3/W8, pp. 301–305, Aug. 2019, doi: 10.5194/isprs-archives-xlii-3-w8-301-2019.

[6] Food and Agriculture Organization of the United Nations (FAO), *Global Forest Resources Assessment 2020: Main report*. [Online]. Available: https://www.fao.org/documents/card/en/c/ca9825en. Accessed: May 8, 2025.

[7] J. Feigl, J. Frey, T. Seifert, and B. Koch, "Close-range remote sensing of forest structure for biodiversity assessments: A systematic literature review," *Current Forestry Reports*, vol. 11, no. 1, May 2025, doi: 10.1007/s40725-025-00251-x.

[8] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020, doi: 10.1109/tpami.2018.2844175.

[9] Ultralytics, "YOLOv8 Documentation," [Online]. Available: https://docs.ultralytics.com. [Accessed: May 2, 2025].

[10] F. Moreira, D. Ascoli, H. Safford, M. A. Adams, J. M. Moreno, J. M. C. Pereira, F. X. Catry, J. Armesto, W. Bond, M. E. González, T. Curt, N. Koutsias, L. McCaw, O. Price, J. G. Pausas, E. Rigolot, S. Stephens, C. Tavsanoglu, V. R. Vallejo, B. W. Van Wilgen, G. Xanthopoulos and P. M. Fernandes, "Wildfire management in Mediterranean-type regions: paradigm change needed," *Environmental Research Letters*, vol. 15, no. 1, p. 011001, Jan. 2020, doi: 10.1088/1748-9326/ab541e.

[11] M. Nasimi and R. L. Wood, "Using deep learning and advanced image processing for the automated estimation of tornado-induced treefall," *Remote Sensing*, vol. 16, no. 7, pp. 1130, Mar. 2024, doi: 10.3390/rs16071130.

[12] X. Guo, Q. Liu, R. P. Sharma, Q. Chen, Q. Ye, S. Tang and L. Fu, "Tree recognition on the plantation using UAV images with ultrahigh spatial resolution in a complex environment," *Remote Sensing*, vol. 13, no. 20, pp. 4122–4122, Oct. 2021, doi: 10.3390/rs13204122.

[13] S. Candiago, F. Remondino, M. De Giglio, M. Dubbini, and M. Gattelli, "Evaluating multispectral images and vegetation indices for precision farming applications from UAV images," *Remote Sensing*, vol. 7, no. 4, pp. 4026–4047, Apr. 2015, doi: 10.3390/rs70404026.

[14] P. Han, C. Ma, J. Chen, L. Chen, S. Bu, S. Xu, Y. Zhao, C. Zhang and T. Hagino, "Fast tree detection and counting on UAVs for sequential aerial images with generating orthophoto mosaicing," *Remote Sensing*, vol. 14, no. 16, p. 4113, Aug. 2022, doi: 10.3390/rs14164113.

[15] Y. Diez, S. Kentsch, M. Fukuda, M. L. L. Caceres, K. Moritake, and M. Cabezas, "Deep learning in forestry using UAV-acquired RGB data: A practical review," *Remote Sensing*, vol. 13, no. 14, p. 2837, Jul. 2021, doi: 10.3390/rs13142837.

[16] S. Miao, C. Wang, G. Kong, X. Yuan, X. Shen, and C. Liu, "Utilizing active learning and attention-CNN to classify vegetation based on UAV multispectral data," *Scientific Reports*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-82248-3.

[17] A. Guimarães, M. Valério, B. Fidalgo, R. Salas-Gonzalez, C. Pereira, and M. Mendes, "Cork oak production estimation using a mask R-CNN," *Energies*, vol. 15, no. 24, p. 9593, Dec. 2022, doi: 10.3390/en15249593.

[18] A. Malta, J. Lopes, Raúl Salas-González, B. Fidalgo, T. Farinha, and M. Mendes, "*Pinus pinaster* diameter, height, and volume estimation using mask-RCNN," *Sustainability*, vol. 15, no. 24, pp. 16814–16814, Dec. 2023, doi: 10.3390/su152416814.

[19] D. E. Kislov and K. A. Korznikov, "Automatic windthrow detection using very-high-resolution satellite imagery and deep learning," *Remote Sensing*, vol. 12, no. 7, p. 1145, Apr. 2020, doi: 10.3390/rs12071145.

[20] S. Yao, Z. Hao, C. J. Post, E. A. Mikhailova, and L. Lin, "Individual tree crown detection and classification of live and dead trees using a mask region-based convolutional neural network (mask R-CNN)," *Forests*, vol. 15, no. 11, p. 1900, Oct. 2024, doi: 10.3390/f15111900.

[21] M. Worachairungreung, N. Kulpanich, P. Sae-ngow, K. Anurak, and P. Hemwan, "Classification of coconut trees within plantations from UAV images using deep learning with faster R-CNN and mask R-CNN," *Journal of Human Earth and Future*, vol. 5, no. 4, pp. 560–573, Dec. 2024, doi: 10.28991/hef-2024-05-04-02.

[22] Z. Luo, H. Xu, Y. Xing, C. Zhu, Z. Jiao, and C. Cui, "YOLO-UFS: A Novel Detection Model for UAVs to Detect Early Forest Fires," Forests, vol. 16, no. 5, pp. 743–743, Apr. 2025, doi: 10.3390/f16050743.

[23] L. Zhang, S. Yu, B. Yang, S. Zhao, Z. Huang, Z. Yang and H. Yu, "YOLOv8 forestry pest recognition based on improved re-parametric convolution," *Frontiers in Plant Science*, vol. 16, Mar. 2025, doi: 10.3389/fpls.2025.1552853.

[24] P. J. Soto-Vega, D. L. Torres, G. X. Andrade-Miranda, G. A. O. P. Costa, and R. Q. Feitosa, "Assessing the generalization capacity of convolutional neural networks and vision transformers for deforestation detection in tropical biomes," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLVIII–3, 2024, pp. 519–525, Nov. 2024, doi: 10.5194/isprs-archives-xlviii-3-2024-519-2024.

[25] B. G. Weinstein, S. Marconi, S. A. Bohlman, A. Zare, and E. P. White, "Cross-site learning in deep learning RGB tree crown detection," *Ecological Informatics*, vol. 56, p. 101061, Mar. 2020, doi: 10.1016/j.ecoinf.2020.101061.

[26] M. J. Allen, D. Moreno-Fernández, P. Ruiz-Benito, S. W. D. Grieve, and E. R. Lines, "Low-cost tree crown dieback estimation using deep learning-based segmentation," *Environmental Data Science*, vol. 3, 2024, doi: 10.1017/eds.2024.16.

Chapter 26

Analysis in Social Media Applied to Business Intelligence and Analytics: A Literature Review

Contribution #51
Cite as:
Ana Cardoso, Laura Batista, Beatriz Maia, Cristiana Maltez, Fernanda Brito Correia. Analysis in Social Media Applied to Business Intelligence and Analytics: A Literature Review. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-51,
  title ={Analysis in Social Media Applied to Business
     Intelligence and Analytics: A Literature Review},
  author={Ana Cardoso and  Laura Batista and  Beatriz
     Maia and  Cristiana Maltez and  Fernanda Brito
     Correia},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Sentiment Analysis in Social Media Applied to Business Intelligence and Analytics: A Literature Review

Ana Sofia Cardoso[1,*], Beatriz Maia[1], Cristiana Maltez[1], Laura Batista[1], Fernanda Brito Correia[1]

[1] Polytechnique University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093, Coimbra, Portugal

[*]Corresponding author
a2021133158@isec.pt

Abstract

This paper is a literature review of Sentiment Analysis techniques applied to Social Media data. The review was guided by 4 Research Questions and analysed (1) the most effective Machine Learning algorithms, (2) common evaluation metrics used to assess performance, (3) contributions to Business Intelligence and Analytics, and (4) main limitations. Following Kitchenham's methodology and PRISMA framework, were analysed 14 recent studies, published between 2021 and 2024 and obtained from IEEE Xplore. The analysis revealed that these studies employed diverse Machine Learning models, ranging from traditional supervised and unsupervised approaches to advanced Deep Learning methods. The studies demonstrated that Deep Learning models, particularly LSTM and Bi-LSTM, outperformed traditional approaches, achieving accuracy rates up to 99.4%, with performance evaluated using several metrics. Furthermore, the literature analysis showed that Sentiment Analysis significantly contributed to the field of Business Intelligence and Analytics by enabling real-time customer insights, by supporting data-driven decisions, and by identifying market trends. However, the analysed studies highlighted persistent challenges in detection of sarcasm, multilingual content processing, informal language handling, and computational demands when using Deep Learning models. This literature review can guide the investigation of Sentiment Analysis in Social Media, while identifying future research directions.

Keywords: Sentiment Analysis, Social Media Data, Business Intelligence and Analytics, Machine Learning, Deep Learning, Literature Review

1 Introduction

Sentiment Analysis (SA) has emerged as a powerful technique for understanding public opinion and emotional responses expressed through online platforms. At its core, SA evaluates whether the tone of a given text is positive, negative, or neutral by employing algorithms that assess the context and emotional weight of words. This analytical approach offers diverse applications, especially in Business Intelligence and Analytics (BI&A), where companies leverage sentiment insights to understand consumer behaviour, improve customer experience, and refine marketing strategies [1]. Moreover, Social Media (SM) platforms generate vast amounts of user-generated content daily, making them valuable sources for tracking trends, detecting emerging issues, and assessing brand reputation.

Recent advancements in traditional Machine Learning (ML) and Deep Learning (DL) have improved the performance of SA systems [2], with research evolving from traditional ML approaches to sophisticated Neural Networks architectures [3]. However, challenges remain, including sarcasm detection, informal and multilingual language handling, and high computational costs of DL approaches [4]. SM platforms, like Twitter present additional complexities due to data noise, such as slang, emojis, and hashtags, posing challenges for accurate sentiment classification [5]. Despite these advances, current literature lacks comprehensive comparative analysis of ML techniques specifically applied to SM SA, within BI&A contexts. Many studies have concentrated on specific techniques without systematically evaluating various approaches, which has hindered the comprehensive understanding of the best methodologies for real-world applications. This literature review seeks to fill this gap by examining SA techniques applied to SM data, with the goal of enhancing the understanding of current ML approaches, their effectiveness, limitations, and potential future developments.

Four Research Questions (RQ) guided the analysis and are presented in Table 1. In this table RQ_N is the

number of the RQ and in the third column is given the RQ motivation.
This paper is structured as follows, in section 2 is presented the methodology used in this literature review, in section 3 is included the analysis and discussion done, in section 4 the conclusions are presented and directions for future research are proposed. The final section lists the references used along the paper.

Table 1. List of the RQs

RQ_N	RQ	Motivation
RQ1	What are the most effective ML techniques used?	To identify and compare ML algorithms currently employed for SA on SM platforms that deliver optimal performance
RQ2	What evaluation metrics are commonly used to assess the performance?	How researchers measure and compare the effectiveness of different SA approaches
RQ3	How did SA approaches contribute to the field of BI&A?	How SA insights from SM translate to actionable BI&A to support strategic decisions
RQ4	What are the main limitations?	To address the limitations and challenges of current techniques to highlight opportunities for future research and improvement

2 Methodology

This section outlines the approach adopted for conducting this literature review. It was followed the Kitchenham's methodology [6], supplemented by the PRISMA framework diagram [7] to ensure transparency and reproducibility.
The planning phase defined the search strategy and selection criteria. Considering the RQs, the following search string was defined: "sentiment analysis" AND ("business intelligence" OR "business analytics" OR "data analytics") AND "machine learning" AND "social media". This search was executed exclusively in the IEEE Xplore digital library, which houses high-quality publications in computer science and engineering fields. To ensure contemporary findings, we applied several preconditions to our search, namely the publication years 2021 onwards and the language English.
To ensure the relevance and methodological rigor of the selected studies, we defined specific inclusion criteria (IC) and exclusion criteria (EC) that guided our selection process, listed in Table 2, where IC_N is the number of each IC and EC_N is the number of each EC.

Table 2 - Inclusion and exclusion criteria

IC_N	IC	EC_N	EC
IC1	Studies that applied ML	EC1	Studies without experimental validation
IC2	Studies with experimental results and performance metrics	EC2	Studies where the title, or focus, was not on SA using ML applied to SM
IC3	Titles that included SA and ML	EC3	Studies that focused on a specific country or region
		EC4	Studies that focused on pandemic or public health crises (e.g., COVID-19)

These criteria allowed filtering the initial search results and focusing on studies that provided empirical evidence of ML applications for SA in general SM contexts, excluding narrowly focused or non-experimental works that would not substantially contribute to addressing our RQs. The number of articles selected each year from 2021 to 2024 based on the IC increased from 1 in 2021 to 4 in 2022, remained stable with 4 in 2023, and increased slightly to 5 in 2024. Till March 2025 weren't found any papers based on the specified criteria.

The literature review protocol implemented is described with the PRISMA flow diagram (Figure 1). Initially, the search in the IEEE Xplore database identified 121 records for screening. The screening process was conducted in 2 sequential phases. At first, the titles of all 121 papers were examined, which led to the exclusion of 84 papers. In the second phase, the abstracts of the remaining 37 papers were analysed, which resulted in the exclusion of an additional 23 papers. After applying our IC and EC during the screening phase, 107 records were excluded. The remaining 14 studies were sought for retrieval, all of which were successfully accessed and included in the final review.

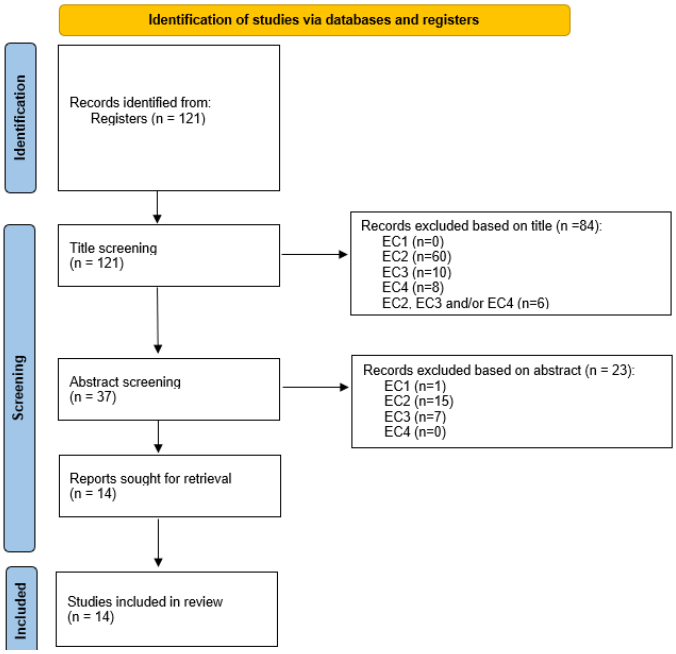


Figure 1 - PRISMA flow diagram of the articles search and selection

3 Analysis and Discussion

This section synthesizes the key findings from the literature review while addressing the defined RQs. By systematically analysing 14 selected studies on SA in SM, this review provides a comprehensive exploration of ML techniques, their applications, and inherent challenges. Additionally, it examines the lexicon-based approach, a widely used SA method that relies on predefined sentiment dictionaries to classify text as positive, negative, or neutral. This method, often used independently or in hybrid models, is significant in SA due to its interpretability and efficiency in processing large volumes of text.

A. Studies Goals

The selected studies explored SA in various contexts using diverse techniques. Their main objectives, summarized in Error: Reference source not found, were to understand and classify emotions and opinions in textual data to gain valuable insights for different applications.

Table 3 – Studies goals

Goals	Reference
Develop a decision support framework for customers using Aspect-Based Sentiment Analysis (ABSA) and Fuzzy Analytic Hierarchy Process	[1]
Build a robust model for SA in tweets and Facebook posts based on a hybrid Convolutional Neural Networks (CNN)- Gated Recurrent Units (GRU) approach	[2]
Investigate signs of emotional distress in online spaces through ML algorithms	[3]

Goals	Reference
Identify the main challenges of SA to help new researchers define research problems	[4]
Practical application and comparative analysis of the effectiveness of ML algorithms for SA	[5]
Understand SA systems that ignore or normalize elongated words and propose a new lexicon-based SA system that considers elongated words to calculate the sentiment level	[8]
Analyse sentiment in Twitter data using different ML models and lexicon-based approaches, comparing their results	[9]
Present a technique for analysing trends on Twitter, predicting emotions in tweets	[10]
Discover food-related sentiments among SM users and analysing attitudes based on food reviews	[11]
Develop a bio-inspired classification algorithm to classify sentiments in big data. Specifically, propose a Robust Optimization (RO) to identify the most valuable features for SA	[12]
Perform a unified SA of multiple comments on SM and improve the accuracy of SA by modifying the information gain method and applying it to Naive Bayes (NB) and Bi-LSTM models	[13]
Analyse SM data and extract valuable insights using advanced analytics techniques, applying clustering algorithms	[14]
Analyse user comments on Reddit, detect recurring patterns, extract relevant insights and perform a comparative analysis between CPU-based and GPU-accelerated AI processing for SA	[15]
Use a large dataset of tweets and advanced ML techniques to analyse ChatGPT public sentiment	[16]

They addressed challenges in SA, developed and evaluated models for specific data characteristics, and examined sentiments around tools like ChatGPT, opinions on food or electronic products, and signs of emotional distress. Some studies also focused on optimizing processes using clustering, GPU parallel processing, or combining ML algorithms with lexicon-based approaches. Overall, the common goal was to enhance the accuracy and effectiveness of SA to better understand the emotional landscape in textual data from various sources.

B. Machine Learning Techniques Applied to Sentiment Analysis

This section addresses RQ1, which analyse the most effective ML techniques for SA on SM identified by researchers. Various ML algorithms applied to SA on SM were identified in the 14 selected studies. Table 4 to Table 6 provide a detailed overview of the identified techniques, categorized by their approach types, highlighting the strengths and weaknesses noted by the studies. Supervised machine learning models, which are trained on labelled datasets, learn the relationships between input features and sentiment outputs, allowing for sentiment classification predictions from annotated Social Media text data.

The most prevalent supervised techniques in SA included Tree-based algorithms, namely Random Forest (RF), used in 1 study and Decision Trees (DT), used in 3 studies. RF's ensemble approach mitigates overfitting, while effectively handling high-dimensional and noisy SM data. On the other hand, DT, while simpler, offers interpretability advantages for understanding feature importance in sentiment expression. Logistic Regression (LR) and NB were used in four and five studies respectively, remaining effective baseline models. NB is computationally light but assumes feature independence, while LR offers a good balance of speed and performance for real-time tasks, K-Nearest Neighbour (KNN), used in three studies, classifies based on similarity to nearby data points but is mainly used for classification. Support Vector Machine (SVM) featured in five studies, find optimal hyperplanes to separate classes, showing high accuracy in both binary and multiclass sentiment tasks. LinearSVC, used in one study, is a linear SVM variant effective in high-dimensional text classification.

Unsupervised methods were less common, used mainly when labelled data was scarce, allowing pattern detection in SM data without pre-labelled sentiment annotations. K-Means, cited in one study, clusters data into k predefined clusters by iteratively assigning points to the nearest centroid, minimizing squared distances between points and cluster centroids. DBSCAN, mentioned in one study, finds dense clusters

separated by sparse areas, but struggles with noise handling.

Deep Learning approaches were frequently employed in SM analysis due to their capability to capture complex linguistic patterns and contextual information. CNN were mentioned in three studies. Although CNN are traditionally associated with image processing, they have been successfully adapted for Natural Language Processing (NLP) tasks, including SA. In this context, CNN identify local patterns, such as phrases or n-grams, by treating text as a one-dimensional signal. They apply convolutional filters to input sequences to learn features that represent important linguistic patterns and specific word combinations. Recurrent Neural Networks (RNNs) were highlighted in one study for their ability to process sequential data by maintaining hidden states that capture information about previous words. This makes them valuable for tasks where word order and context are crucial. However, RNNs face limitations with long sequences due to the vanishing gradient problem. Long Short-Term Memory networks (LSTMs), mentioned in three studies, address this issue with memory cells that maintain information over long sequences, capturing context and dependencies. Bidirectional LSTMs (Bi-LSTMs), noted in two studies, enhance this by processing sequences in both directions, capturing context from both past and future words. BERT, mentioned in one study, is a revolutionary language model in NLP that understands bidirectional word context within sentences, enabling more accurate interpretations through attention mechanisms.

Table 4 - Supervised models used in the analysed papers

Algorithm	Strengths	Weaknesses
DT	• Easy to interpret and visualize and useful as a simple baseline for comparison [16] • Very easy to understand and interpret [3] • [5] did not present any strength	• Prone to overfitting, especially with noisy data and lower performance compared to more advanced models [16] • Lower performance with noisy data [3] • [3] and [5] did not present any weaknesses
KNN	• An intuitive and simple approach [14] • [3], [15] did not present strengths	• Generally, it is primarily used for classification [14] • [3], [15] did not present weaknesses
LinearSVC	• Achieved an accuracy exceeding 80% [11]	• [11] did not present weaknesses
LR	• Commonly used for data classification [3] • Provides probabilistic predictions, making it useful for classification tasks [5] • [11] and [15] did not present strengths	• Can struggle with more intricate tasks or datasets with a high dimensionality of features and assumes a linear relationship between independent and dependent variables [5] • [11] and [15] did not present weaknesses
NB	• More efficient and accurate than SVM when applied to training sets with a large number of evaluations [11] • Performs well on large text datasets and high accuracy in some studies [3] • [5], [9], [13] did not present strengths	• Assumes feature independence, not always true [3] • Performance is generally lower than more modern models like LSTM [3], [11] • [5], [9], [13] did not present any weaknesses
RF	• Combined multiple DT, improving robustness and accuracy of predictions, reducing the risk of overfitting and able to handle missing data efficiently [15]	• Training multiple DT can be time-consuming and require significant computational resources and due to its ensemble nature, it is harder to interpret and understand individual DTs compared to a single DT [15]

Algorithm	Strengths	Weaknesses
SVM	• Maximizes the margin between classes, which helps improve generalization and classification accuracy [9] • Effective in high-dimensional spaces, good generalization even with limited data, high performance with high-dimensional data like text and works well even with small, labelled datasets [3] • [5], [15] and [16] did not present strengths.	• May not perform well with non-linear data [9] • It can have high computational costs on CPUs due to the temporal complexity that grows quadratically with the number of samples [15] • Slow training on large datasets, more complex in terms of result interpretation and difficult to tune, especially for multi-class problems [16] • High computational cost with large datasets [3] • [5] did not present weaknesses

Table 5 - Unsupervised models used in the analysed papers

Algorithm	Strengths	Weaknesses
DBSCAN	• allows to discover clusters of arbitrary shapes, unlike algorithms like K-Means that tend to find more spherical clusters [14]	• Not suitable for noise cluster analysis [14]
K-means	• Easy-to-use technique, it scales to large datasets, ensures convergence, is simple to implement and adapt easily to new examples, covering clusters of various sizes and shapes [14]	• K-means may be inaccurate when the data includes noise or irregularly shaped clusters [14]

Table 6 - Deep Learning models used in the analysed papers

Algorithm	Strengths	Weaknesses
BERT	• Have made substantial strides in recent years in NLP applications involving SA and provided the highest level of accuracy when analysing sentiment in airline review data [1]	• May have limitations in sentiment scoring scale and in identifying mixed sentiments [1]
Bi-LSTM	• Captured two-way semantic dependencies or context, improving accuracy in SA [9], [13] • Ability to learn complex patterns in data [13]	• Computationally expensive and required more resources compared to simpler models, making them slower to train and deploy [9], [13]
CNN	• Suitable for short text in SA and efficient in parallel computation [2] • Outperformed LT, showing higher classification accuracy [10] • Effectively extract spatial and short-term dynamic features [3]	• Cannot model context beyond local features and struggled with long-term dependencies in sequences [2] • [3] and [10] did not present weaknesses
LSTM	• Excellent performance on sequential data, such as text [9], [16] • Very effective for Sain NLP, excellent for sequential data like tweets and messages, and outperforms classical models in many studies [3]	• High computational cost and longer training time and needs large volumes of labelled data to train effectively [16] • Required high computational resources and longer training time compared to traditional models [3] • Lacks bidirectional context [9]

Algorithm	Strengths	Weaknesses
RNN	• Extensively used in text analysis, have shown great potential for SA and due to their complex structure and the ability to understand context, DL models like RNN, performed well [9]	• [9] did not present weaknesses

C. Evaluation Metrics

This section addresses RQ2, which examines the evaluation metrics commonly used to assess the performance of SA techniques on SM. The selection of appropriate metrics was crucial for accurately measuring and comparing the effectiveness of different algorithms and approaches. The analysis of the 14 selected studies revealed a consistent set of main metrics employed across SA research (Table 7).

Table 7 - Metrics used in the analysed papers

Metric	Reference
Accuracy	[2], [3], [5], [10], [12], [13], [15]
AUC	[3]
Confusion Matrix	[2], [3], [15]
F1-Score	[2], [3], [5], [8], [11], [13], [15], [16]
Precision	[2], [3], [5], [8], [11], [13], [15], [16]
Recall	[2], [3], [15]
ROC Curve	[2], [3]
Sentiment Score (BERT)	[1]
Specificity	[15]

D. Datasets Used

SA on SM using ML relied on various datasets to train and evaluate models. Twitter represented the most widely explored source, with popular datasets including Sentiment140 [9], tweet collections of 31,962 samples [10], 236,275 ChatGPT-related tweets [16], and millions of tweets for SA [3], as well as informal chats shared on the platform alongside Facebook and personal messaging [8]. Beyond Twitter, other studies analysed sentiment on Facebook, using data from posts and interactions [15], often combined with Twitter data through official APIs [2], [14]. E-commerce platforms constituted another major data source, particularly Amazon datasets including Fine Food Reviews [11] and product-specific collections (BOO, CPA, MTV, HAK) [12], plus customer feedback from Amazon, eBay, Flipkart, and Ajio [1]. Additional sources included Reddit subreddit data via PRAW API [15], Weibo comments (10,000 samples) [13], and established benchmark datasets such as Stanford Sentiment Treebank, IMDB, and Amazon Reviews [4]. Additionally, some studies incorporated multimodal approaches, combining text with audio, images, and video for psychological analysis [3]. Data collection methods included API access and web scraping for WhatsApp, Twitter, and Facebook data [14], and a CSV format, already processed to remove explicit emotions, containing six fields [5]. This dataset diversity enabled ML models to capture varied linguistic patterns, cultural contexts, and sentiment expressions across different SM platforms and domains.

E. Results Obtained

Addressing RQ1, the analysis of the 14 studies revealed clear patterns regarding which ML approaches delivered the best performance in SA of SM data. Several studies highlighted the superiority of LSTM and Bi-LSTM models, especially for sequential data such as tweets and reviews [13]. Bi-LSTM model often performed the best, due to its ability to capture bidirectional semantic dependencies. A study on ChatGPT

[16] concluded that the LSTM model demonstrated superior performance across all evaluation metrics. CNNs were effective at extracting local patterns in text and have achieved high accuracy, such as in a study on identifying hashtags in tweets [10], where CNN outperformed LR with an accuracy of approximately 95%. The combination of CNNs for local feature extraction and GRUs for sequential modelling has demonstrated superior results compared to individual models and traditional ML methods in SA of tweets and Facebook posts. One study reported an accuracy of 89.5% for a hybrid CNN-GRU model [2]. BERT pre-trained language model demonstrated significant progress, achieving the highest accuracy in SA on airline review data in one study. BERT was also used to obtain sentiment scores for features in ABSA [1]. Furthermore, a study that compared SVM and NB using movie reviews concluded that NB was more effective [11]. In a study that used the K-Means clustering algorithm to analyse SM data [14], the results showed the effectiveness of K-Means in grouping similar data and detecting trends in conversations from collected data. Another study found that Maximum Entropy performed best in classifying tweets related to McDonald's and KFC, when compared to other supervised algorithms such as NB, SVM, DT and Bagging. In contrast, in an analysis of tweets about attacks, a word frequency model combined with DT, LinearSVC, LR and NB algorithm achieved an accuracy of 83% [3]. In a study that analysed Reddit data, RF achieved the highest accuracy rate of 99.4%, followed by SVM and KNN, while LR had the lowest accuracy RO, a bio-inspired classifier proposed in [12], achieved higher accuracy and F1-measure compared to other techniques in SA big data. Another study compared NB and SVM with RNN, LSTM and Bi-LSTM for SA of tweets, concluding that DL models generally delivered more accurate results, with Bi-LSTM achieving the best accuracy of 81.33% and F1-score of 81.54% [9].In the study [8], a new lexicon-based system that considered long words outperformed traditional systems that ignored them, achieving F1-score rates of 81% to 96% on datasets of informal Facebook conversations, Tweets, and personal chats. The average F1-score increase was 21.22% when including long words. The results indicated that the ML approach outperformed DL and lexicon-based methods, in terms of accuracy, in other studies, mentioned in [5]. In [4], no experimental tests were carried out directly, only results from other studies were referenced.

In summary, DL techniques, particularly LSTM, Bi-LSTM, CNN and hybrid models such as CNN-GRU, together with transformers such as BERT, consistently demonstrated high effectiveness in SA across a variety of domains. However, traditional ML methods such as SVM, NB and RF remain relevant and can be quite effective depending on the dataset and context. The choice of the most appropriate technique depends on the specific characteristics of the data, the objectives of the analysis, and the available computational resources.

F. Contributions to Business Intelligence and Analytics

This section addresses RQ3, which considers how SA approaches contribute to the field of BI&A, particularly in strategic decision-making, by analysing the selected studies. It was observed the following contributions, the deep understanding of customer opinions and feedback about products, services, brands, and even competitors, the support to strategic decision-making in business and marketing, identifying market trends and predicting consumer behaviour and the detection of implicit opinions and nuances in language. By analysing comments on e-commerce platforms, SM and other online sources, companies can identify what customers appreciate and what aspects need improvement. For example, the study [11] analysed Amazon food reviews revealing that most sentiments were positive but also revealed negative sentiments that can guide food companies to improve their products. Similarly, SA of restaurant reviews on TripAdvisor helped to understand customer satisfaction with aspects such as food, ambiance, and service. ABSA, discussed in [1], provided an even more granular understanding, identifying the sentiments expressed toward specific features of a product or service. For example, when analysing airline reviews, ABSA could identify positive sentiments about food, but negative sentiments about flights, providing actionable insights for improvements. Also, SA provided crucial insights that supported strategic decision-making in business and marketing. Understanding public sentiment towards their products and services, as well as those of their competitors, enables companies to adjust their marketing, communication, and product development strategies more effectively. The article [4] mentioned the ability to derive market intelligence from microblogs such as Twitter through SA, which could inform decisions about marketing campaigns and

new product launches. Analysing tweets about electronic products could shed light on public perception and help inform strategic decisions related to those products [4]. Study [12] highlighted how SA became an essential business tool, providing valuable insights into consumer thoughts, helping to define better marketing strategies and informing product development decisions. SA enables businesses to identify emerging market trends and predict consumer behaviour. By monitoring public sentiment around specific events, products or topics, organizations can anticipate changes in consumer preferences and adapt their strategies accordingly. The study [8] used Twitter data to detect public sentiment and predict future stock market movements, demonstrating the potential of SA in predicting trends and understanding market reactions to events. Real-time SA of SM conversations can help identify emerging trends [14]. SA of film reviews, for example, can help predict their commercial success, influencing production and distribution decisions [13]. Advanced SA approaches could detect implicit opinions and nuances in language, such as sarcasm, irony, negation, and the use of elongated words to express intense emotion. Study [8] proposed a system that considers elongated words like "happyyyyyy" to calculate the level of sentiment, arguing that these words intensify the expressed emotion. Ignoring these nuances, as traditional systems do, can lead to an underestimation of the customer's true sentiment. Detecting and handling negation is crucial to avoid incorrect sentiment classifications. For example, the sentence "dislike" expresses a negative sentiment, even though it contains the word "like" which, on its own, would be positive [12].

The 14 studies analysed in this review demonstrated how SA on SM can anticipate changes in consumer behaviour, allowing companies to respond more quickly to market preferences. These insights support data-driven decision-making, enabling organizations to adjust their marketing strategies, improve customer service, and guide product development. When integrated into BI&A platforms, SA enriches traditional analytics with emotional and behavioural dimensions, offering a more comprehensive view of customer perception.

G. Limitations and Challenges

This section addresses RQ4, highlighting the limitations and challenges identified by the selected papers. These obstacles hindered the accuracy and reliability of SA across various domains. A major challenge is the reliance on correct spelling, as many SA models struggle with misspelled words or informal language common in social media and online discussions. Although some advanced techniques attempt to correct spelling errors, they are not always accurate, leading to potential misclassifications. Table 8 provides a detailed overview of these limitations and the corresponding studies. Despite these challenges, ongoing advancements in NLP and ML offer promising solutions. Addressing these issues will lead to more robust SA models, enhancing decision-making and providing more accurate insights into consumer sentiment across digital platforms.

Table 8 - Limitations identified used in the analysed papers

Limitations	Reference
Complexity in implementing context-aware sentiment models. Lack of high-quality labelled benchmark datasets	[1]
High computational demands of hybrid models (e.g., CNN-GRU)	[2]
Difficulty handling spelling errors and informal language. Lack of contextual understanding in SM posts	[2], [3], [8]
Difficulty detecting sarcasm, irony, and denial	[4], [11]
Evaluation limited to training and test sets only, risking overfitting	[5]
High computational complexity and long training times of DL models. Need for large, labelled datasets and limited annotation quality	[9]
Dependency on specific platforms like Twitter and e-commerce platforms	[10], [12]
Layer depth constraints in models like Bi-LSTM limiting their ability to capture complex linguistic patterns	[13]

Limitations	Reference
Limitations of K-Means: sensitive to noise, outliers, and only identifies spherical clusters. Challenges in data collection: restricted tools and APIs, unrepresentative filtering methods	[14]
The article does not explicitly detail the limitations found during the execution of this specific study.	[15]
Restriction to specific languages, such as English-only datasets	[16]

4 Conclusions

This literature review was motivated by the increasing complexity and volume of sentiment-rich data generated on SM, necessitating more effective ML techniques for analysis. To address this, a literature review was conducted, analysing 14 studies from IEEE Xplore published between 2021 and 2024. The review addressed the following research questions: (RQ1) DL techniques, particularly LSTM, Bi-LSTM, and hybrid models such CNN-GRU, were identified as the most effective for SA on SM; (RQ2) Accuracy, F1-score, and Precision emerged as the most commonly used evaluation metrics, underscoring the need for multifaceted performance assessments; (RQ3) SA significantly contributed to BI&A by providing deep insights into customer opinions, supporting strategic decision-making, identifying market trends, and detecting linguistic nuances; (RQ4) key limitations included difficulties in processing complex language, high computational demands, domain and platform dependencies, and challenges in detecting linguistic nuances such as sarcasm and irony. This literature review advances the state of the art by elucidating current best practices in sentiment classification and highlighting the limitations that impede broader adoption. These findings can inform the refinement of future SA models and applications. Future research should focus on developing multilingual and low-resource sentiment models, enhancing mechanisms for detecting sarcasm and irony, and creating more efficient model architectures suitable for real-time analysis. Interdisciplinary efforts that integrate NLP with sociolinguistics and cultural studies may further enhance the interpretability and applicability of sentiment insights.

References

[1] P. Sinha, S. Roychowdhury, e B. A. Tanaji, «Customer Feedback Analysis Using Aspect Based Sentiment Analysis and Fuzzy Analytic Hierarchy Process», em *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, abr. 2024, pp. 1–6. doi: 10.1109/I2CT61223.2024.10544032.

[2] A. Kaur, V. Kukreja, e D. Kundra, «Echoes of Emotion: Enhancing Sentiment Analysis with CNN-GRU Synergy», em *2024 International Conference on Big Data Analytics in Bioinformatics (DABCon)*, nov. 2024, pp. 1–7. doi: 10.1109/DABCon63472.2024.10919262.

[3] A. Sadhukhan *et al.*, «Machine Learning Techniques for Identifying Distressed Individuals on Social Media - A Survey», em *2024 International Conference on Big Data Analytics in Bioinformatics (DABCon)*, nov. 2024, pp. 01–06. doi: 10.1109/DABCon63472.2024.10919310.

[4] S. A. Parvin, M. Sumathi, e C. Mohan, «Challenges of Sentiment Analysis - A Survey», em *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, jun. 2021, pp. 781–786. doi: 10.1109/ICOEI51242.2021.9453026.

[5] K. Sharma, D. Jain, A. Jain, Y. S. Rathore, e S. Agarwal, «Sentimental Analysis based on Machine Learning Technique», em *2024 3rd International Conference for Innovation in Technology (INOCON)*, mar. 2024, pp. 1–7. doi: 10.1109/INOCON60754.2024.10511973.

[6] B. Kitchenham, «Procedures for Performing Systematic Reviews», *Keele, UK, Keele Univ.*, vol. 33, ago. 2004.

[7] D. Moher, A. Liberati, J. Tetzlaff, e D. G. Altman, «Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement», *BMJ*, vol. 339, p. b2535, jul. 2009, doi: 10.1136/bmj.b2535.

[8] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, e Mohd. A. Shah, «Improving Sentiment Analysis in Social Media by Handling Lengthened Words», *IEEE Access*, vol. 11, pp. 9775–9788, 2023, doi: 10.1109/ACCESS.2023.3238366.

[9] V. Pandya, A. Somthankar, S. S. Shrivastava, e M. Patil, «Twitter Sentiment Analysis using Machine Learning and Deep Learning Techniques», em *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)*, Bangalore, India: IEEE, dez. 2021, pp. 1–5. doi: 10.1109/C2I454156.2021.9689241.
[10] H. Vanam e J. R. R. R, «Sentiment Analysis of Twitter Data Using Big Data Analytics and Deep Learning Model», em *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, Chennai, India: IEEE, jan. 2023, pp. 1–6. doi: 10.1109/ICECONF57129.2023.10084281.
[11] T. K. Bhatia e P. Aggarwal, «Sentiment Analysis on Food Reviews Using Data Visualization Tools», em *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, out. 2022, pp. 1–6. doi: 10.1109/ICRITO56286.2022.9964673.
[12] P. Menakadevi e J. Ramkumar, «Robust Optimization Based Extreme Learning Machine for Sentiment Analysis in Big Data», em *2022 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, mar. 2022, pp. 1–5. doi: 10.1109/ICACTA54488.2022.9753203.
[13] R. Lin, «Comment Texts Sentiment Analysis Based on Improved Bi-LSTM and Naive Bayes», em *2022 International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, ago. 2022, pp. 407–412. doi: 10.1109/ICDACAI57211.2022.00087.
[14] V. V. Sai Teja Veeramallu, A. L. Venkata Sandeep, B. Shivling, D. Burra, e S. Bulla, «Social Media Data Analytics Using K-Means Clustering», em *2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, set. 2023, pp. 426–431. doi: 10.1109/ICUIS60567.2023.00076.
[15] S. Mer, S. Saxena, e A. S. Pundir, «Revolutionizing Sentiment Analysis: Accelerated Data Science Approaches for Reddit Submissions», em *2024 IEEE 5th India Council International Subsections Conference (INDISCON)*, ago. 2024, pp. 1–6. doi: 10.1109/INDISCON62179.2024.10744244.
[16] V. Abishethvarman, K. Banujan, A. Wijeratne, R. Nirubikaa, S. Kumara, e S. Prasanth, «Unveiling Public Perception of ChatGPT Through Twitter Data and Machine Learning», em *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, out. 2023, pp. 431–435. doi: 10.1109/ICDABI60145.2023.10629351.

Chapter 27

Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning

Contribution #52
Cite as:
Filipe Carvalho, Paula Couceiro, Mateus Mendes, Pascoal Silva. Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-52,
  title ={Analysis and Modeling of Phenolic Compounds
     and Antioxidant Activity of two Banana Varieties
     Using Machine Learning},
  author={Filipe Carvalho and  Paula Couceiro and
     Mateus Mendes and  Pascoal Silva},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Analysis and Modeling of Phenolic Compounds and Antioxidant Activity of two Banana Varieties Using Machine Learning

Filipe Carvalho[1,*], Paula Couceiro[1,*], Raquel Guiné[2], Pascoal Silva[1,3], Mateus Mendes[1,4,5]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[2] Polytechnic University of Viseu
[3] CMUC-Center for Mathematics, University of Coimbra
[4] RCM^{2+} Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[5] Institute of Systems and Robotics,
Department of Electrical and Computer Engineering, University of Coimbra,
3030-290 Coimbra, Portugal
[*] Corresponding author
{a2021148720,a2018056493, pascals, mmendes}@isec.pt, raquelguine@esav.ipv.pt

Abstract

Phenolic compounds (PC) and antioxidant activity (AA) are very important for human well-being. Therefore, quantification and prediction of PC and AA are crucial for food production and distribution chains. In the present study, a dataset with experimental measurement results was analysed. Machine learning models MLPRegressor and Random Forest models were used to predict PC and AA based on input variables: variety, drying state, extract type and extract order. The importance of the input variables was evaluated, where for the MLPRegressor the variable with the largest weight was the drying state and for the Random Forest it was the order of the extract, both being strongly correlated with the phenolic compounds and the antioxidant activity. Performance metrics were applied to evaluate the results obtained with both models, where it was concluded that MLPRegressor obtained lower values than Random Forest for RMSE, MAE, MAPE and MPE, and higher values for R^2, which proved that MLPRegressor performs better.

Keywords: Phenolic compounds; Antioxidant activity; Banana properties; Machine learning models; MLPRegressor; Random Forest

1 Introduction

Polyphenols, widely present in foods from the Mediterranean diet, have been recognised as essential bioactive compounds in the prevention of chronic non-communicable diseases such as cardiovascular diseases, type 2 diabetes and some types of cancer [1]. Their importance comes largely from their high antioxidant activity, which neutralises the free radicals responsible for oxidative

stress - one of the main mechanisms associated with cell degeneration and the development of various pathologies [2]. Scalbert *et al.* [3] also show that the beneficial effects of polyphenols go beyond simply neutralising free radicals, acting on molecular pathways related to inflammation and metabolic regulation. However, the concentration and effectiveness of these compounds can be significantly affected by agronomic and technological factors. Tomas-Barberán & Espín [4] and Guiné *et al.* [5] show that conditions such as the degree of ripeness, the type of crop and drying methods directly influence the phenolic content of foods. It is therefore clear that both the regular consumption of foods rich in polyphenols and care in the production and processing processes are fundamental to maximise their benefits in promoting human health.

The present work aims aim to perform exploratory data analysis over a dataset of results of laboratory analysis of two banana varieties. Two machine learning models, MLPRegressor and Random Forest, were also deployed to predict the phenolic compounds and antioxidant activity present in the two banana varieties. The study aims to understand which of the input variables has the most influence on the results obtained and then which model achieves the best performance.

The following sections present the state of the art and explain the methodology used, the models chosen and their respective architectures. The models will then be evaluated using performance metrics, concluding which is the best model and which variable has the greatest influence on phenolic compounds and antioxidant activity. The results will be presented in the respective section and discussed in the following section. Finally, the last section will present some conclusions.

2 State of the art

According to Manach *et al.* [1], polyphenols, in addition to being widely present in the Mediterranean diet, play an important role in the prevention of chronic non-communicable diseases such as cardiovascular disease, type 2 diabetes and certain neoplasms.

Lobo *et al.* [2] emphasise that oxidative stress, caused by an imbalance between the production of free radicals and the body's antioxidant capacity, is directly associated with various chronic diseases such as cancer, diabetes, cardiovascular and neurodegenerative diseases. The authors point out that dietary antioxidants, especially bioactive compounds such as polyphenols, play a fundamental role in neutralising these free radicals, contributing to the prevention of these pathologies. Functional foods rich in natural antioxidants have therefore emerged as important allies in health promotion.

Scalbert *et al.* [3] emphasise that polyphenols exert significant antioxidant activity, but also act on cellular pathways that regulate inflammation, metabolism and gene expression. The authors argue that the beneficial effects of polyphenols go beyond neutralising free radicals and involve complex molecular mechanisms.

Tomas-Barberán and Espín [4] analyse how the phenolic compounds present in fruit and vegetables are influenced by factors such as ripeness, type of cultivation, climatic conditions and agricultural practices. The authors emphasise that these elements directly affect the quantity and profile of polyphenols in foods, with relevant implications for their nutritional and functional quality.

Guiné *et al.* [5] studied the impact of different drying conditions on the content of total phenolic compounds and antioxidant activity in bananas from different cultivars. A model using classical feed-forward artificial neural networks was developed. They conclude that air drying had the greatest impact on the final results. Another conclusion was that total phenolic compounds and antioxidant activity could be predicted with great precision from the model built.

Mongi *et al.* [6] investigated the effect of solar drying on vegetables and its relationship with the amount of phenolic compounds and antioxidants present in them. They concluded that drying does significantly affect the results obtained.

Sarpong *et al.* [7] also carried out a study on bananas, but on the influence of convention drying with air and controlled humidity on the production of bioactive compounds and antioxidant degradation in dehydrated slices. These authors concluded that higher drying temperatures allow for greater nutrient retention.

3 Materials and methods

This section presents the methodology used to carry out the study and the models that will be used. It also presents the dataset and an exoloratory analysis of it, as well as the data processing that had to be carried out.

3.1 Methodology

The present study followed the CRISP-DM methodology. CRISP-DM (Cross-Industry Standard Process for Data Mining) is a standard methodology for conducting data mining projects. It organises the process into six main phases: business understanding, data understanding, prepare data, model data, results and deployment, promoting an iterative and structured approach. It is widely used because it can be adapted to different sectors and types of data.

For the modeling phase, two machine learning models were used: MLPRegressor (Multi-Layer Perceptron Regressor) and Random Forest Regressor. Both are widely used in regression tasks, where they normally achieve good performance predicting continuous variables. MLPRegressor is based on artificial neural networks and is capable of modelling complex, non-linear relationships between variables, although it is less interpretable. Random Forest, on the other hand, is an ensemble learning model that uses multiple decision trees to obtain more stable and robust predictions, and is especially effective with noisy and high-dimensional data. Two main differences between the two are the type of structure (neural network vs. trees) and the level of interpretability, with Random Forest generally being more explainable. Both models are available in the scikit-learn library, one of the most popular Python libraries for classical machine learning[1].

3.2 Dataset

The dataset consisted of 288 results of laboratory analysis of bananas of two different cultivars. Table 1 describe the study's input variables—four variables, all categorical. The first variable is the variety of banana, and the study contained two varieties of banana. The second variable is the drying state, as the bananas could have four different drying states. The third variable is the type of extract used, of which there were two types, and the fourth variable is the order of the extract, of which there were three possible orders.

Table 1: List of initial input variables and their possible values.

Variety	Madeira, Costa Rica
Drying state	Fresh, Dried at 50 ^{o}C, Dried at 70 ^{o}C, Freeze-dried
Extract type	Methanol, Acetone
Extract order	1, 2, 3

[1] https://scikit-learn.org/stable/index.html (last checked on 2025-06-02).

The dataset under study therefore has the variables mentioned above as inputs and total phenols and antioxidant activity as outputs. For each set of input conditions, 6 repetitions were made, and in some cases no results were obtained.

Figure 1 shows all the data from the dataset. The phenolic compounds obtained with methanol and acetone as a function of drying state are shown on the left, and the antioxidant activity with methanol and acetone as a function of drying state is shown on the right. The graph also shows a numbering from 1 to 3, which corresponds to the order of the extract for each result obtained. It can be seen that there are anomalies in the results, most likely due to the variability inherent to the laboratorial nature of the experiments.

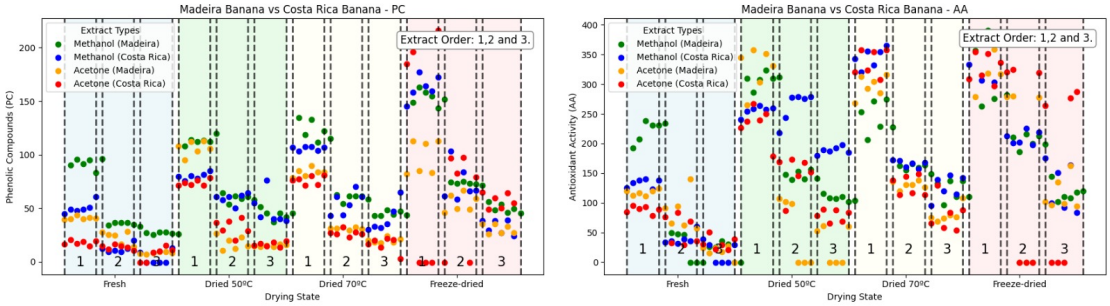


Figure 1: Graphical representation of total phenols (left) and antioxidant activity (right) of bananas from Madeira and Costa Rican bananas with methanol and acetone as a function of drying state.

Following on from the data analysis, it was clear from the previous graphs that there was a lack of data that was assumed to be '-1' and that there were areas that were difficult to interpret, i.e. there may or may not be outliers in these areas. To check this, a more detailed analysis was carried out using boxplot graphs.

When carrying out the study for the Madeira banana, two boxplots were constructed, one for methanol and the other for acetone in relation to total phenols, on the left. It was possible to see that there is a dispersion of data for both extracts, although slightly less for acetone and there were no outliers, as can be seen in figure 2. Figure 2 also shows the boxplot of antioxidant activity as a function of the type of extract used on the Madeira banana, on the right. The conclusions were the same: there is a great deal of dispersion in the data, but there are no outliers. It can also be seen that the average results with methanol, whether for total phenolics or antioxidant activity, are always higher.

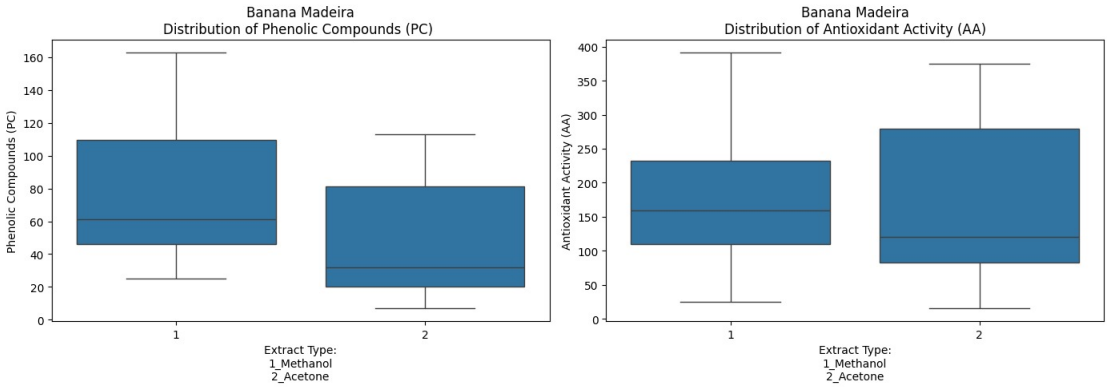


Figure 2: Boxplot of total phenols and antioxidant activity of Madeira bananas as a function of methanol and acetone.

In the case of Costa Rica bananas, the same analysis, in figure 3, distribution of phenolic compounds on the left, immediately revealed the presence of anomalous values, as can be seen in figure 3. These outliers are present in the bloxplot of total phenols, they exist for both methanol and acetone and take on much higher values than the rest. With regard to antioxidant activity, there was a greater dispersion of data, but no outliers. It was also found that the average results were higher in all cases with methanol.

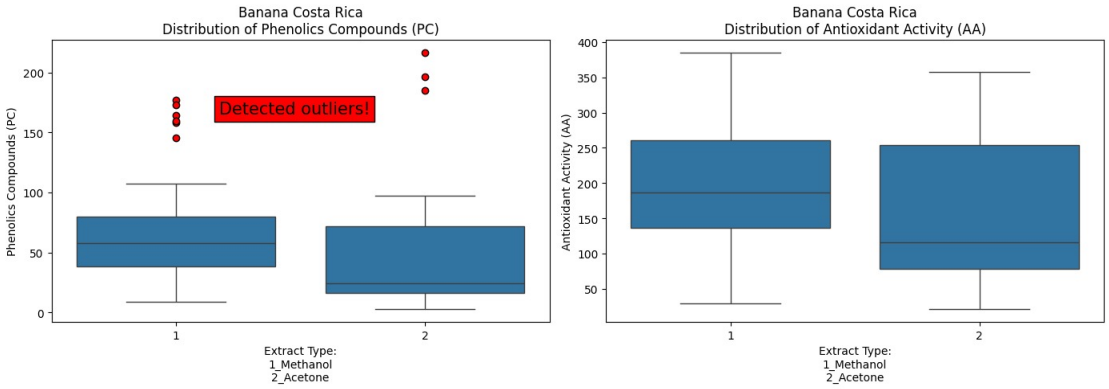


Figure 3: Boxplot of total phenols and antioxidant activity of Costa Rican bananas as a function of methanol and acetone

.

So, after analysing the dataset, all the lines with null or negative values were eliminated because they made no sense in the dataset. The initial dataset, which contained 288 samples, was reduced to a total of 257 viable samples for the study. Tables 2 and 3 summarise the dataset after data cleaning.

Tables 2 and 3 present a description of the data relating to the samples that will enter the study for PC and AA, respectively, for two sample varieties: Madeira and Costa Rica. Both tables include the number of valid samples, the average, the minimum and maximum values, and

the standard deviation. In terms of the number of samples, both varieties are balanced, with 129 samples from Madeira and 128 from Costa Rica.

The Madeira variety has an average PC of 50.51, slightly higher than Costa Rica, which has an average of 44.22. However, the Costa Rican variety has a wider range of values, with a minimum of 2.77 and a maximum of 216.41, while Madeira varies between 6.99 and 162.78. In addition, the standard deviation for Costa Rica (43.28) is higher than for Madeira (38.60), suggesting greater variability in the data for this variety.

For the AA variable, the Costa Rica variety has a higher average (165) compared to Madeira (140), which indicates that, on average, Costa Rica has more antioxidant activity than Madeira. However, both varieties have very similar maximum values (391.18 for Madeira and 384.88 for Costa Rica) and relatively close minimum values (15.59 for Madeira and 21.10 for Costa Rica). The standard deviation is high for both varieties, with 99.94 for Madeira and 102.68 for Costa Rica, reflecting a high dispersion of the data, with slightly greater variability in Costa Rica.

Table 2: Samples valid for output variable PC.

Variety	Nº of samples	Average	MIN	MAX	Standard deviation
Madeira	129	50.51	6.99	162.78	38.60
Costa Rica	128	44.22	2.77	216.41	43.28

Table 3: Samples valid for output variable AA.

Variety	Nº of samples	Average	MIN	MAX	Standard deviation
Madeira	129	140	15.59	391.18	99.94
Costa Rica	128	165	21.10	384.88	102.68

Thus, in figure 4, the data has been represented in a similar way to what was done when analysing the dataset. On the left, the figure shows the phenolic compounds obtained with methanol and acetone as a function of drying state and extract order, and on the right, the same for antioxidant activity. These graphs only show the data after processing and it is clear to see the changes when compared to the graphs in figure 1. With regard to both phenolic compounds and antioxidant activity, for the first drying state, fresh, all the samples for the third order of extract were excluded, and there was a reduction in the number of samples from the second order. For the second drying state, there was a reduction in the number of samples for the second and third order extracts, and for the last drying state, there was a reduction in the number of samples for the first and second order extracts. For the 'dried at 70º' drying state, no samples were removed.

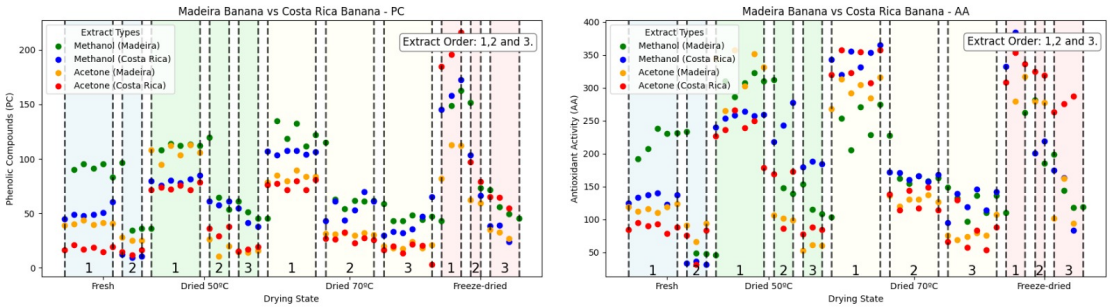


Figure 4: Graphical representation of total phenols (left) and antioxidant activity (right) of bananas from Madeira and Costa Rican bananas with methanol and acetone as a function of drying state, after data processing.

To begin the study, 70% of the data for Madeira bananas and 70% of the data for Costa Rica bananas were selected, with the remaining 30% of data for each variety being kept for testing the models. This division in terms of variety was made to ensure that the model was trained with the same percentage for each variety, thus avoiding the risk of unbalanced distribution of the data.

With the data selected, two forecasting models were built using MLPRegressor and Random Forest, the main objective being to understand which is best applied to the study in question.

4 Results

This section provides an in-depth analysis of the data after it has been processed, with the aim of gaining a better understanding of the dataset. Initially, an exploratory analysis was carried out using a correlation heatmap, which made it possible to assess the strength of the relationships between the numerical variables. Next came the modelling stage, where we began by checking that the models were learning well so that we could then check the importance of each variable for each model. The models were then optimised to see which architecture was best. Once the architecture had been chosen, the models could finally be trained in order to obtain results.

4.1 Exploratory data analysis

Tables 2 and 3 show that Madeira bananas, on average, contain more PC, while Costa Rica bananas are richer in AA. Bearing in mind that the experiment and the comparison between results were carried out keeping the inputs the same and varying only the variety of banana, if for the same inputs the banana from Madeira contains more PC and the banana from Costa Rica contains more AA, the only reason for this is the quality of the banana.

The heatmap in figure 5 shows the linear relationship between the independent and dependent variables in the data set. The values vary between -1 and 1, where positive values indicate a direct correlation and negative values indicate an inverse correlation.

The dependent variables PC and AA show a strong positive correlation with each other (0.85), which suggests that these two variables are strongly related: as one increases, the other tends to increase as well. This strong correlation is justified, having in mind that the phenolic compounds are responsible for a great deal of the antioxodantn in foods where they are present. Among the predictor variables, the `Extract_Order` variable stands out, with a significant negative correlation

with PC (-0.65) and with AA (-0.66). This indicates that as the extraction order increases, the PC and AA values tend to decrease. This negative relationship suggests that extracts obtained at more advanced stages of the process perform less well in terms of response variables, which can be explained by a decrease in the concentration of active compounds over the course of consecutive extractions.

The `State` variable also shows a moderate positive correlation with PC (0.44) and with AA (0.48). This indicates that the physical state of the material (e.g. fresh or dry) positively influences both output variables, and it is likely that less processed products, as those in the fresh state, associated with better results.

On the other hand, the `Variety` and `Extract_Type` variables show weak or practically zero correlations with PC and AA, suggesting that, in isolation, they do not have much of a direct linear impact on the output variables.

In short, analysis of the correlation matrix shows that `Extract_Order` and `State` are the predictor variables with the greatest relevance in explaining the behaviour of the dependent variables, and are therefore natural candidates to be given greater weight during predictive modelling. The strong correlation between PC and AA also suggests that multivariate models can benefit from this relationship to improve predictive capacity.

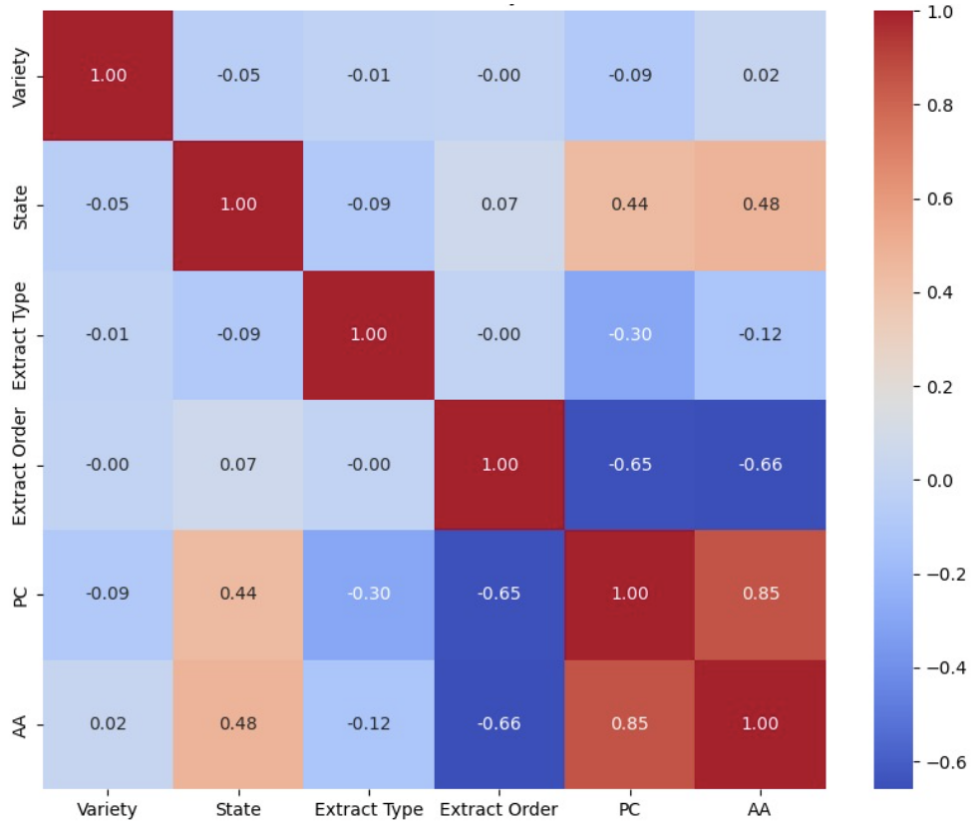


Figure 5: Heatmap of the correlations between inputs and outputs.

4.2 Modeling

This sub-section explains how the learning of the models was verified, the architecture chosen and the results obtained.

4.2.1 Input Weight analysis

MLPRegressor (Multi-Layer Perceptron Regressor) and Random Forest are two types of regression models used in machine learning, but with different approaches. MLPRegressor is an artificial neural network made up of layers of neurons connected to each other, capable of learning complex, non-linear relationships between variables, being sensitive to the choice of hyperparameters and requiring data normalisation. Random Forest is an ensemble of several decision trees trained with random subsets of the data and variables, combining their results to improve robustness and reduce the risk of overfitting; it is easier to interpret and usually works well with fewer adjustments.

Both models have different sets of parameters that must be adjusted to optimise their performance. In MLPRegressor, the main parameters include `hidden_layer_sizes`, which defines the structure of the network's hidden layers, activation, which specifies the activation function (such as 'relu' or 'tanh'), solver, which determines the optimisation algorithm (such as 'adam' or 'sgd'), as well as alpha, `learning_rate_init`, `max_iter` and `early_stopping`, which control regularisation, learning and convergence. In RandomForestRegressor, parameters such as `n_estimators` (number of trees), `max_depth` (maximum tree depth), `min_samples_split` and `min_samples_leaf`, which control tree growth, as well as `max_features`, which determines how many variables to consider in each split, and bootstrap and `oob_score`, which affect how samples are used. These parameters directly influence the balance between bias and variance in the models.

It was then decided to use MLPRegressor with 2 hidden layers with 200 neurons in the first and 100 in the second and Randon Forest with 500 trees.

To see if the models were learning well during training, two different evaluation methods were applied, one for each model. In the case of MLPRegressor, the loss curve was applied. The loss curve obtained, figure 6, during the training of the MLPRegressor model shows a sharp reduction in loss in the first few epochs, followed by a more gradual decrease until it stabilises. This behaviour is typical of an effective learning process, indicating that the model is managing to adjust its parameters to minimise the error between predictions and actual values. The absence of sudden oscillations or increases in loss suggests that the training was stable and there was no obvious overfitting at this stage. Thus, the loss curve confirms that the model is learning progressively and consistently over the epochs.

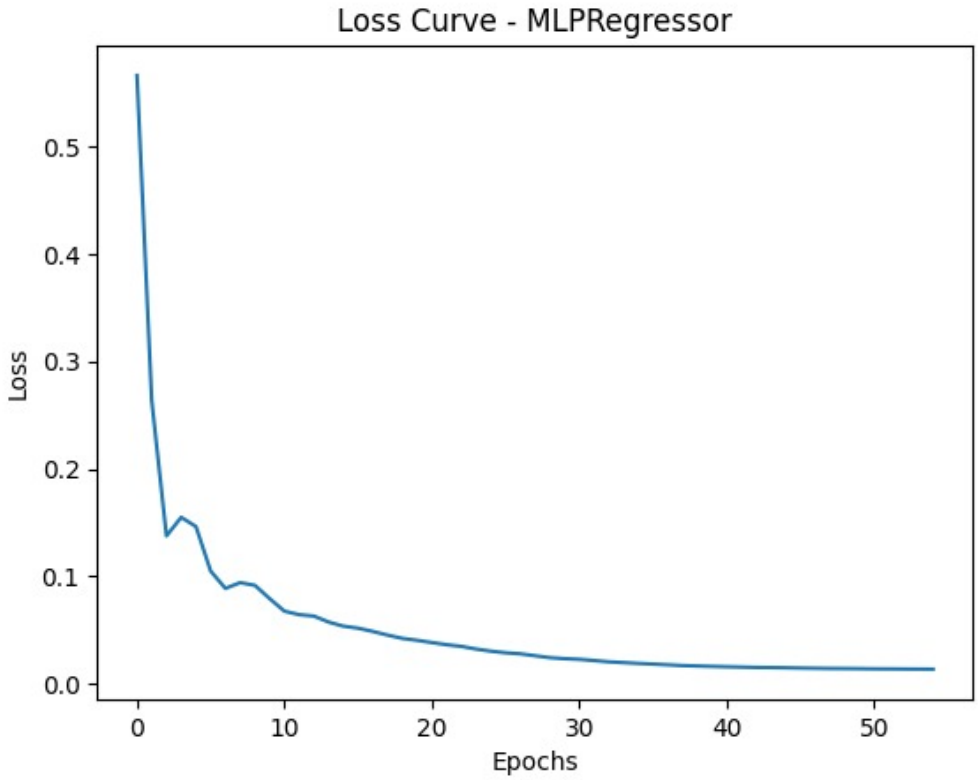


Figure 6: Loss curve.

Although the Random Forest model is not trained by epochs like neural networks, it is possible to generate a curve analogous to the loss curve by observing how performance varies with the increase in the number of trees in the forest. To do this, the model is trained by progressively increasing the `n_estimators` parameter and the desired metric (such as root mean square error or R^2) is evaluated on a validation or test set at each step. The result is a graph showing whether the model continues to improve as more trees are added or whether it reaches saturation point. This approach makes it possible to understand the stability of learning and identify the ideal number of trees for a good balance between performance and computational cost. Figure 7 shows this stabilisation from 100 trees onwards.

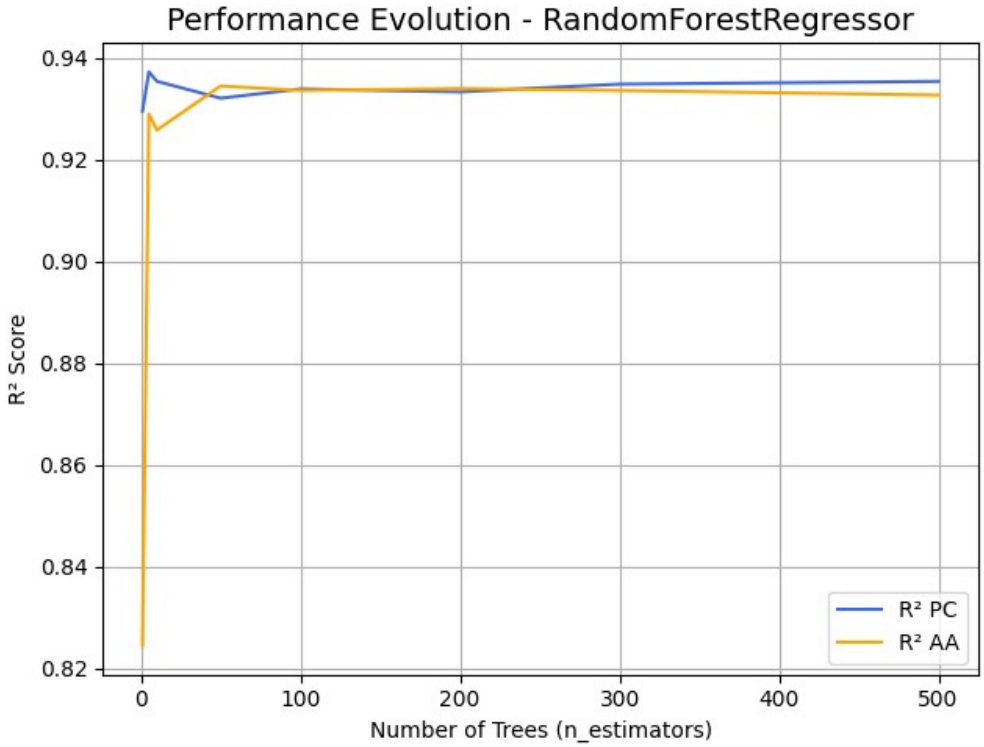


Figure 7: Random Forest performance evaluation.

After applying the models, the importance of each variable was checked. Table 4 shows the weights of each variable for each model. It was concluded that for MLPRegressor the state of drying is more important, while for Random Forest the order of the extract has more influence.

Table 4: Weights of each variable for each model.

Model	Variety	Drying state	Extract type	Extract order
MLPRegressor	0.80	1.00	0.75	0.72
Random Forest	0.05	0.38	0.07	0.49

4.2.2 Model optimization

With the two models chosen, an analysis was carried out to find the best architecture for each one. Figure 8 shows the analysis carried out with several different architectures for MLPRegressor. The analysis relates the R^2 to the number of neurons present in each hidden layer.

The activation function used was the ReLU (Rectified Linear Unit), which is an activation function widely used in artificial neural networks. It works in a simple way: for any input x, the output is x if $x > 0$ and 0 if $x \leq 0$. In other words, it 'zeroes out' all negative values and leaves positive values unchanged. This introduces non-linearity into the model, allowing the network to learn complex patterns, while avoiding problems such as gradient fading that can occur with other functions, such as the sigmoid. Because of its simplicity and computational efficiency, ReLU is a standard choice in many modern deep learning architectures.

Figure 9 shows the same analysis, but this time for Random Forest, where the R^2 is related to the number of decision trees. It was found that the best architecture for MLPRegressor has two hidden layers: the first with 200 neurons and the second with 100. For Random Forest, it was found that up to 50 decision trees there was a difference in the result, but from 100 decision trees onwards there was stability in the R^2, so with 100 or 500 decision trees the model's performance would be the same.

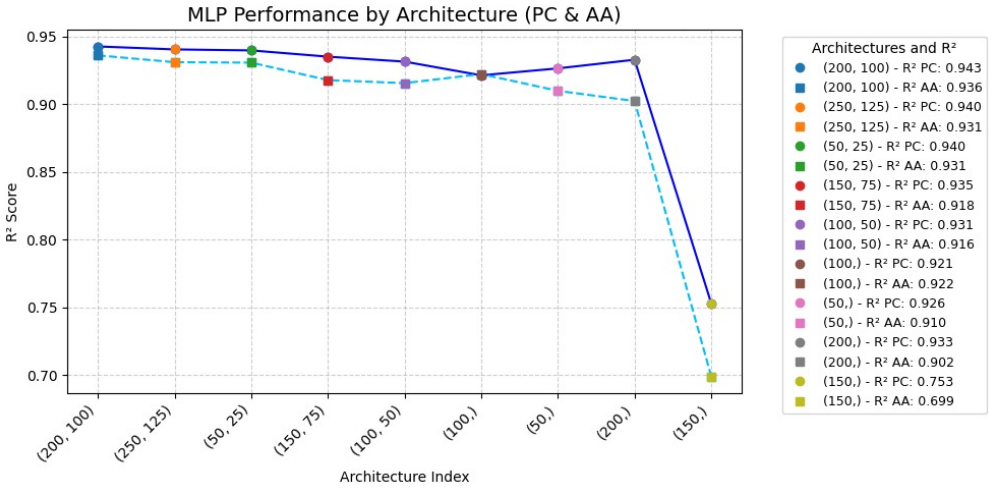


Figure 8: Impact of the MLPRegressor architecture on the value of R^2.

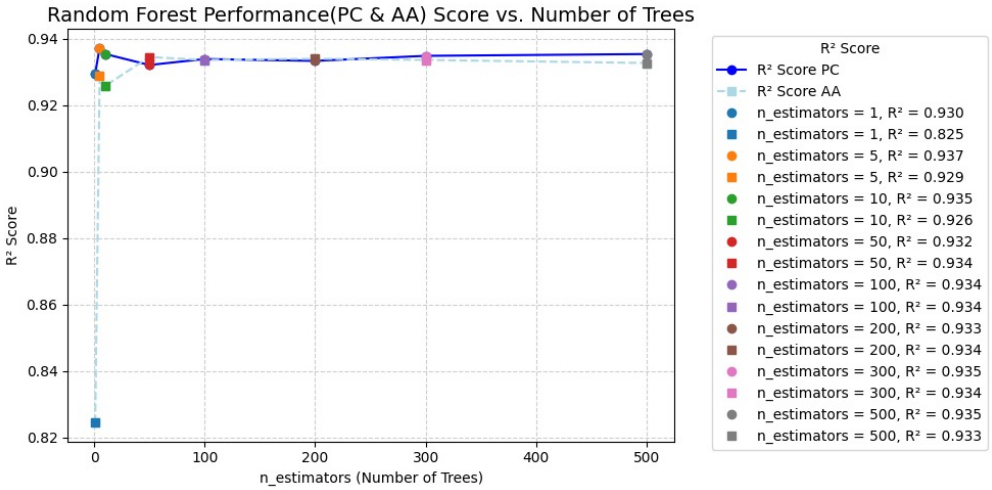


Figure 9: Impact of the Random Forest architecture on the value of R^2.

Although the outputs were predicted simultaneously, the results were obtained individually. Figures 10 and 11 show the results obtained with MLPRegressor and Random Forest, respectively.

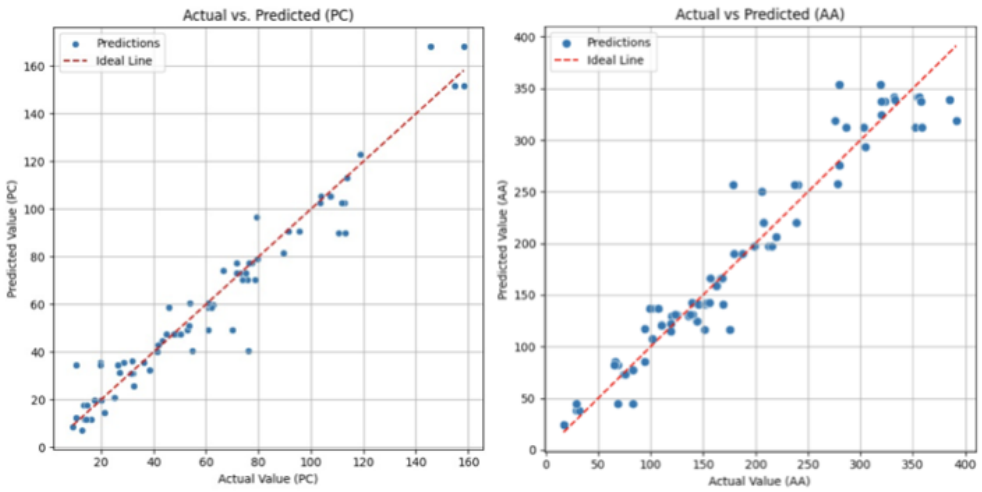


Figure 10: Results obtained with MLPRegressor.

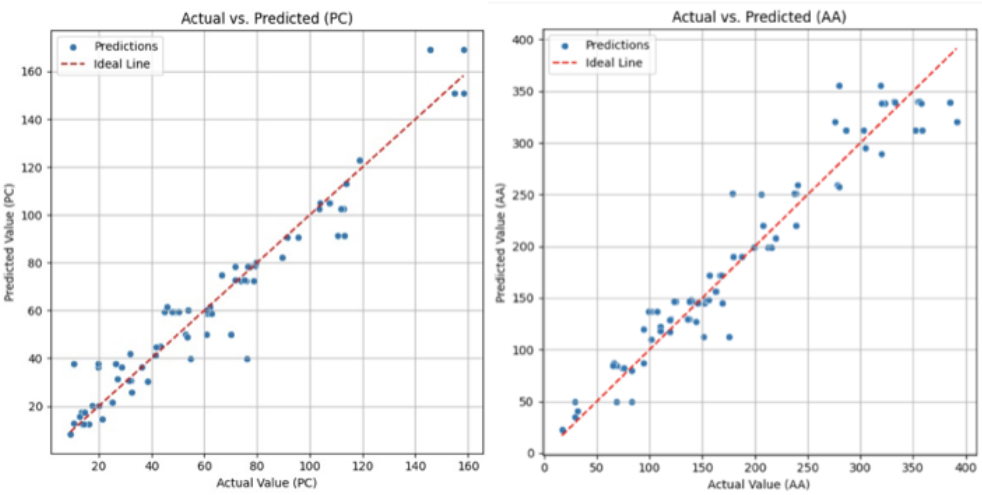


Figure 11: Results obtained with Random Forest.

5 Discussion

In the previous section, studying the weight of the variables, it was found that with MLPRegressor the variable that most influences the outputs is the state of drying, while with Random Forest the variable that has the most weight is the order of the extract.

To effectively compare the results obtained in order to understand which model is the most efficient, various performance metrics (PM) were applied to evaluate the results obtained. R^2, RMSE, MAE, MAPE and MPE were the performance metrics used to assess the performance of regression models. The R^2 (coefficient of determination) measures the proportion of the variance in the data explained by the model, ranging from 0 to 1 (the closer to 1, the better).

RMSE (Root Mean Square Error) is the root of the mean square error, penalising larger errors and being useful for understanding the average magnitude of errors. MAE (Mean Absolute Error) calculates the average of the absolute errors and is more robust to outliers than RMSE. MAPE (Mean Absolute Percentage Error) expresses the errors in percentage terms in relation to the real values, facilitating relative interpretation. The MPE (Mean Percentage Error) is also a percentage, but can indicate whether the model tends to overestimate (negative values) or underestimate (positive values), as it maintains the sign of the error.

Tables 5 and 6 show the values obtained for each performance metric applied.

Table 5: Results of the performance metrics applied to both models that predicted PC.

Model	R^2	RMSE	MAE	MAPE	MPE
MLPRegressor	94.26%	8.9477	5.8358	15.14%	229.20%
Random Forest	93.54%	9.4984	6.3836	16.86%	258.98%

Table 6: Results of the performance metrics applied to both models that predicted AA.

Model	R^2	RMSE	MAE	MAPE	MPE
MLPRegressor	93.61%	24.6275	17.7588	12.12%	54.46%
Random Forest	93.27%	25.2730	19.0156	12.94%	70.89%

Although the results are very similar, it can be seen that MLPRegressor always shows lower values than Random Forest, except for R^2, which is higher, showing that MLPRegressor explains more of the variance in the data.

6 Conclusion

The study carried out aimed to study the phenolic compounds and antioxidant activity, taking into account their importance for humans, present in two varieties of bananas.

An analysis of the dataset provided was carried out, where it was verified for both models used, MLPRegressor and Random Forest, which input was most important. The drying state and the order of the extract are the variables that have larger influence on both AA and PC. The order of the extract being a lab variable is only important for experimental results, but the drying state shows that food preservation may have a significant impact on the PC and AA.

After this analysis, through performance metrics, it was verified that MLPRegressor showed better efficiency than Random Forest, although by a small margin, to predict PC and AA based on the input variables. Thus, both models can be used as predictors with a good degree of confidence.

References

[1] Claudine Manach; Augustin Scalbert; Christine Morand; Christian Rémésy and Liliana Jiménez. Polyphenols: food sources and bioavailability. *The American Journal of Clinical Nutrition*, 79(5):727–747, 2004.

[2] V. Lobo; A. Patil; A. Phatak and N. Chandra. Free radicals, antioxidants and functional foods: Impact on human health. *Pharmacognosy Reviews*, 4(8):118–126, 2010.

[3] Augustin Scalbert; Ian T Johnson; and Mike Saltmarsh. Polyphenols: antioxidants and beyond. *The American Journal of Clinical Nutrition*, 121(5):215–217, 2005.

[4] Francisco A Tomás-Barberán and Juan Carlos Espín. Phenolic compounds and related enzymes as determinants of quality in fruits and vegetables. *Journal of the Science of Food and Agriculture*, 81(9):853–876, 2001.

[5] Raquel Guiné; Maria João Barroca; Fernando Gonçalves; Mariana Alves; Solange Oliveira and Mateus Mendes. Aplicação da modelização por redes neuronais ao teor de compostos fenólicos e atividade antioxidante em bananas de diferentes cultivares secadas sob condições distintas. *Lisboa*, 1(1):296–299, 2014.

[6] Richard J. Mongi1; Bernadette K. Ndabikunze1; Trude Wicklund; Lucy M. Chove and Bernard E. Chove. Effect of solar drying methods on total phenolic contents and antioxidant activity of commonly consumed fruits and vegetable (mango, banana, pineapple and tomato) in tanzania. *African Journal of Food Science*, 9(5):291–300, 2015.

[7] Frederick Sarpong; Xiaojie Yu; Cunshan Zhou; Leticia Peace Amenorfe; Junwen Bai1; Bengang Wu and Haile Ma. The kinetics and thermodynamics study of bioactive compounds and antioxidant degradation of dried banana (musa ssp.) slices using controlled humidity convective air drying. *Journal of Food Measurement and Characterization*, 12(3):1935–1946, 2018.

Chapter 28

The Role of Real-Time Navigation APIs in Green and Smart Mobility

Contribution #53
Cite as:
Ana Vigario, David Dias, Tiago Pinto, Arsenio Reis, Joao Barroso. The Role of Real-Time Navigation APIs in Green and Smart Mobility. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-53,
  title ={The Role of Real-Time Navigation APIs in
     Green and Smart Mobility},
  author={Ana Vigário and  David Dias and  Tiago Pinto
     and  Arsénio Reis and  João Barroso},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

The Role of Real-Time Navigation APIs in Green and Smart Mobility

Ana Vigário[1], David Dias[1], Tiago Pinto[1,2], Arsénio Reis[1,2], João Barroso[1,2]
[1] University of Trás-os-Montes and Alto Douro ,Vila Real, Portugal
[2] INESC-TEC ,Vila Real, Portugal

Abstract

In a world increasingly interconnected by mobile devices and intelligent systems, real-time navigation Application Programming Interfaces (APIs) have played a fundamental role in transforming mobility. These systems are capable of providing up-to-date information on routes, traffic, road conditions, and weather, allowing drivers and applications to make faster, safer, and more efficient decisions. By integrating multiple data sources—such as GPS location, crowdsourced input, radars, and others—these APIs process, filter, and contextualize information to present only the most relevant data to the end user. These technologies are essential in urban mobility solutions, directly contributing to safety, reduced travel time and, in the case of electric vehicles, they assist in planning more economical routes and avoiding congestion, thereby improving battery autonomy and promoting smoother, more efficient, and sustainable driving. Real-time navigation APIs are a critical tool in fostering more connected, intelligent, and green mobility, becoming increasingly indispensable for safety, agility, and efficiency in daily travel. This work aims to serve as a theoretical foundation for a future empirical study, with the goal of analyzing in greater depth how real-time navigation APIs contribute to delivering accurate, reliable, and useful information to end users, seeking to understand the mechanisms through which these interfaces acquire, process, and deliver data in an optimized manner.

Keywords: Real-time navigation APIs, urban mobility, electric vehicles, artificial intelligence

1 Introduction

The rapid expansion of urban areas, both in terms of population density and territorial growth, has led to a substantial increase in mobility demands, often characterized by intense traffic congestion. The growing awareness of the need for a more sustainable and eco-friendly lifestyle is driving the adoption of technologies and practices that reduce the ecological footprint of urban transportation, promoting more efficient and environmentally responsible mobility solutions. In this context, navigation applications have evolved significantly—from basic mapping tools to intelligent platforms integrated with real-time data. These innovative technological solutions aim to ensure quality of life for the population while minimizing negative environmental impacts and fostering sustainable urban development [1].

In theory, all navigation apps should provide the fastest route to a user-defined destination. However, their functionality extends far beyond simply identifying the shortest path from point A to point B. By incorporating a range of advanced features, they play a crucial role in promoting road safety and reducing the number of accidents that occur daily. Integrated with real-time

data, these applications leverage geolocation technologies, route optimization algorithms, and predictive analytics to deliver routes adapted to actual traffic conditions [2].

These technologies are essential components of smart urban mobility solutions, directly contributing to increased safety, shorter travel times, and reduced CO_2 emissions. For electric vehicles, for example, they help plan more energy-efficient routes and avoid traffic congestion, thereby improving battery autonomy and enabling a smoother, more economical, and more sustainable driving experience. However, navigation apps are not neutral tools: while they enhance individual driver efficiency, they can also have negative impacts on entire communities [2].

Therefore, intelligent management of access to route information can become a powerful tool for making cities more livable and sustainable. It is essential to understand how these interfaces acquire data, how they process it, and how it is optimized and delivered to the end user. This work aims to serve as a theoretical basis for a future empirical study, with the objective of analyzing the mechanisms through which navigation applications collect, filter, and deliver data in an optimized manner.

2 Navigation Apps

Navigation applications have revolutionized the way we travel, helping users find the most efficient route to their destinations. These tools assist in route planning by minimizing travel time and providing valuable real-time information such as traffic conditions, weather updates, accidents, road closures, and points of interest. With the widespread adoption of smartphones, navigation apps have become essential to modern mobility. They offer accurate user geolocation, facilitate trip planning, and significantly enhance the overall driving experience[3, 4].

The website **AutoExpress**, in its *Product Group Tests* section, selected the most popular navigation apps from the **Apple App Store** and tested them on three different **iPhone** models (from iPhone 12 to 14). The majority of drivers favored **Google Maps**, whose leadership is attributed to its high accuracy in destination searches and effective turn-by-turn navigation. In second place came **Waze**, which stands out as the best option for users interested in avoiding traffic, thanks to its crowdsourced approach and real-time updates.

One of the study's most surprising findings was the performance of **TomTom AmiGO**, which proved to be particularly well-optimized for electric mobility contexts, showing strong compatibility with the specific requirements of electric vehicles (EVs). In addition to these, other evaluated applications included **TomTom GO, Apple Maps, Sygic Car Navigation, MyRoute-App,** and **Yandex Navi**[5].

Meanwhile, the specialized site **Holiday Extras** selected the 10 best free navigation apps for Android and iOS in 2025, namely: **Google Maps**, **Waze**, **Apple Maps**, **Karta GPS – Maps & CarPlay**, **TomTom GO**, **ViaMichelin GPS Route Planner**, **HERE WeGo: Maps & Navigation**, **Maps.Me**,**MyRoute-App Navigation**, and **Roadtrippers – Trip Planner**[6].

For the purposes of this study, the navigation applications common to both sources analyzed — **Google Maps**, **Waze**, **TomTom GO**, and **Apple Maps**— were considered. Additionally, the **HERE WeGo** application was included for comparative purposes, as it frequently appeared in the research conducted.

3 What are real-time navigation Application Programming Interfaces - APIs

An Application Programming Interface (API) defines a standardized set of interfaces, protocols, and conventions that facilitate interoperability and integration between different software components or systems. It enables the controlled exposure and exchange of data, services, and functionalities, allowing applications to interact programmatically in a consistent and structured manner. **APIs** allow only the necessary information to be shared, keeping system details hidden, which enhances system security[7].

An API, in the context of mobility, enables integration with external data services such as traffic systems, weather, geolocation, or urban infrastructure. Through these interfaces, applications can access real-time information, process it, and present users with optimized routes, traffic forecasts, or relevant points of interest. This modular architecture based on APIs also allows for scalability and customization of solutions, facilitating integration with connected vehicles, logistics platforms, or public transportation systems. Thus, the interoperability between applications and APIs is one of the fundamental pillars of smart and sustainable mobility [8].

4 The applications and their APIs

The **Google Maps API** is a set of services and interfaces provided by Google that allows developers to integrate maps and geolocation features into their web and mobile applications [9]. It offers customization based on different types of vehicles, real-time traffic conditions, toll cost estimates, and visual representation of traffic status through polylines[1] over the map[10].

One of the main features of the **Waze API**is its ability to access real-time traffic data, including information about heavy traffic, accidents, roadwork, and other incidents affecting road conditions. The API can calculate optimal routes based on current traffic, display nearby points of interest, provide real-time traffic conditions for the user's location, and deliver navigation instructions using Waze's routing engine [11].

The **TomTom Traffic API** are a set of web services designed for developers to build web and mobile applications around real-time traffic, offering a broad range of solutions to maximize the value of traffic applications or historical traffic analysis[12].

The **Apple Maps Server API** provides access to Apple Maps services, enabling delivery of geographic data such as geocoding and reverse geocoding. Additionally, the API supports location or point of interest (POI) search, route calculation between multiple points, and real-time route monitoring [13].

The **HERE API** stands out for its flexibility and integration capabilities with various systems, such as TMS (Transportation Management System), CRM (Customer Relationship Management), and WMS (Warehouse Management System). It allows the calculation of operational costs, including tolls, and offers specialized technical support. It is particularly suited for complex projects that require high customization, offline data access, and advanced logistics optimization. Moreover, it offers features such as geocoding, analytical data visualization, and support for building interactive and engaging applications [1], Table 1.

[1] A continuous line composed of multiple segments, used to trace the route between points on the map. The API draws this polyline over the map and uses different colours to indicate the traffic conditions along each segment of the route.

Table 1: Comparative Table: Navigation Applications and API Support.

Application	API Support	Main Features	Vehicle Integration	Real-Time Data
Google Maps	Google Maps APIs (directions, geolocation, places, etc.)	Optimized routing, real-time traffic, POIs, Street View	Android Auto, Apple CarPlay	✓
Waze	Waze API	Crowdsourced alerts, traffic, accidents, speed cameras	Android Auto, Apple CarPlay	✓
TomTom GO	TomTom Maps APIs (routing, traffic, search)	Offline navigation, traffic alerts, speed zones	Android Auto, Apple CarPlay	✓
Apple Maps	Maps Server API	Integrated navigation, intelligent suggestions, public transport info	Apple CarPlay only	✓(Apple ecosystem)
HERE WeGo	HERE API	Offline maps, turn-by-turn voice guidance, public transport, traffic info	Android Auto, Apple CarPlay	✓

5 How traffic APIs work

The overall functioning of real-time traffic data systems can be divided into three main phases: raw data acquisition, processing and analysis, and data distribution via APIs [4] Fig. 1.

Navigation applications retrieve real-time data from multiple sources, such as road sensors, weather stations, traffic cameras, historical databases, and, most notably, mobile devices, in-vehicle navigation systems, and user-contributed data through crowdsourcing [4]. For instance, TomTom devices collect anonymous data through navigation apps and equipment, as well as Floating Car Data (FCD) — data transmitted by moving vehicles — which helps identify specific traffic conditions in real time [14].

This data is sent to central servers, where it is processed using artificial intelligence and machine learning techniques. Algorithms interpret the data, identify traffic patterns, calculate estimated time of arrival (ETA), detect incidents such as accidents, and generate alternative route suggestions based on current traffic conditions [15, 16].

After processing, the structured and analyzed data is made available to navigation applications via APIs. These APIs act as standardized communication channels — typically through HTTP requests — allowing applications to query optimized routes, visualize current traffic conditions, and receive real-time alerts and notifications [1].

6 How APIs obtain data

Navigation devices collect and analyze traffic data from various sources, such as GPS, sensors, roadside surveillance cameras, and embedded infrastructure sensors, among others. GPS data is obtained from devices installed in cars, trucks, as well as from mobile devices and personal navigation systems. For instance, the TomTom GO application collects 61 billion GPS data

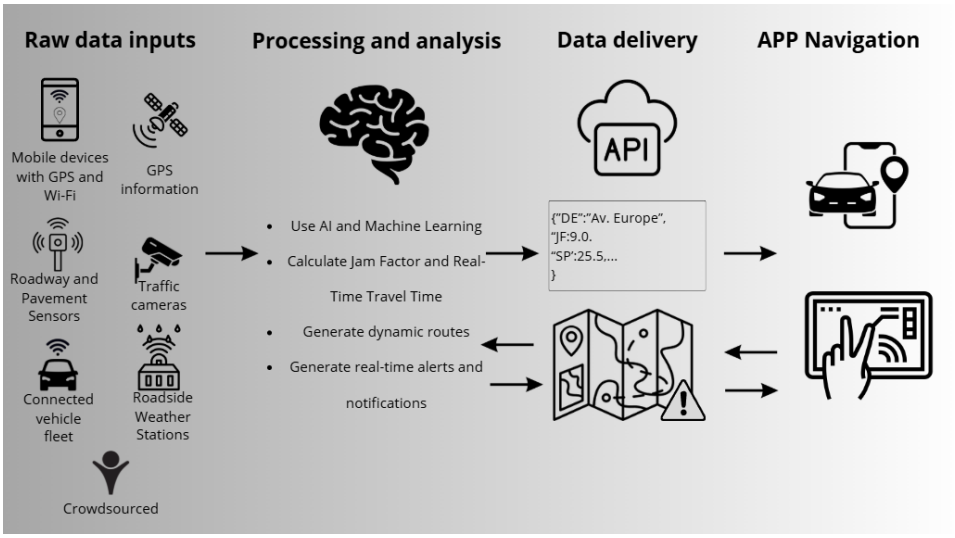


Figure 1: How traffic APIs work

points daily from smartphones, in-vehicle systems, personal navigation devices (PNDs), and fleets, covering over 3.5 billion kilometers of roads worldwide. An additional key data source is the Spatial Data Object (SDO), which is used, for example, to recognize traffic signs, enhancing the detection of traffic incidents such as roadworks. These data help build a real-time picture of traffic flow and density, enabling the prediction not only of traffic jams but also road closures, construction work, and other relevant incidents [12].

Furthermore, data from governmental entities, such as scheduled road maintenance, weather forecasts, and temporary restrictions (e.g., due to sports events, marathons, natural phenomena, or border closures), are integrated into the system. There is also a growing integration of lane-level precision data, which aims to support the next generation of traffic services, including lane-level navigation and advanced driver assistance systems (ADAS/AD) [12].

The HERE application incorporates additional positioning data sources, such as Wi-Fi networks and other network-based technologies. Using these methods may involve sending HERE information such as location, speed, direction, date and time, number of satellites used, nearby cell tower parameters, Wi-Fi networks or Bluetooth® beacons, and the device's Mobile Country Code (MCC) and Mobile Network Code (MNC) [17].

Mobile data collection platforms, such as vehicles equipped with specialized sensors, travel along public roads collecting data on road conditions, pavement quality, signage, and traffic flow. This data provides a comprehensive view of the road infrastructure, including pavement type, number of lanes, and characteristics of the surrounding environment [18].

In the Waze application, real-time traffic updates are based on the principle of crowdsourcing. Drivers using the app continuously transmit real-time location data—including speed, direction, and position—which is then aggregated on Waze's servers [19]. The platform is supported by a community of over 500,000 volunteers who continuously edit maps and report live road events [20].

Navigation applications typically supplement these sources with data from governments and third-party providers, such as:

- Traffic Counters and Sensors: Installed at strategic locations like bridges, intersections,

or highways, these devices record key metrics such as vehicle count, types, average speed, and traffic patterns. Sensors may include inductive loops embedded in pavement, infrared sensors, computer vision cameras, pneumatic or radar sensors, and data from connected vehicles (FCD) [18].

- Surveillance Cameras: Capture real-time images and video of lane occupancy, vehicle types, traffic flow, and incidents. These are processed by traffic management centers and are often made available via public APIs or private agreements with service providers [18].
- Road Weather Stations: Equipped with sensors that measure temperature, humidity, wind speed, and precipitation, enabling the detection of road risks such as ice accumulation or high winds in mountainous regions [18].

This data is generally made available to navigation applications through public feeds or partnerships with transportation authorities. Private providers also aggregate data from multiple sources—including cameras, traffic sensors, and commercial fleets—and resell this information to navigation apps. The combination of professional traffic data with crowdsourced data enables navigation systems to deliver a more accurate and comprehensive view of current traffic conditions [21].

7 Data Processing

All this data, on its own, has limited value unless it is analyzed and interpreted to generate relevant and actionable insights for the end user — the driver. Navigation applications leverage pre-built APIs to streamline the data integration process, functioning as standardized interfaces that allow developers to access map, geolocation, and traffic data. These connectors facilitate the continuous exchange of information between systems, significantly reducing the time and complexity of implementation.

In the localization process, geocoding systems are essential. They use algorithms to analyze descriptive data (such as addresses or place names) and compare it with geographic databases to return the corresponding coordinates. Conversely, through reverse geocoding, it is possible to derive postal addresses from geographic coordinates (longitude and latitude). Tools such as Google Maps Geocoding API, OpenCage Geocoder, and Mapbox Geocoding API are widely used for this purpose [22].

Modern navigation apps integrate artificial intelligence (AI) algorithms to process large volumes of real-time data. These algorithms analyze data from various sources to detect traffic congestion, accidents, road closures, and other incidents that may affect the driving experience. Beyond responding to current traffic conditions, AI models also analyze historical data to predict future patterns in road network behavior. This predictive capability allows applications to anticipate potential traffic bottlenecks and proactively suggest alternative routes [16, 18, 21].

In addition, machine learning techniques are employed to understand individual travel patterns, preferred routes, and frequently visited locations. This knowledge enables navigation systems to offer suggestions tailored to the user's mobility habits and preferences [16, 21].

Google Maps, as an advanced AI-based application, for example, predicts estimated time of arrival (ETA) and provides real-time updates on user location and movement [23]. TomTom, in turn, uses advanced machine learning and AI models to forecast incidents, model traffic, and assess their impact on driving times. It combines historical, current, and real-time data to produce a comprehensive view of road traffic flows [12].

8 Data Distribution to Users

Navigation APIs play a fundamental role in transforming raw data into accurate, contextualized, and directly usable information for the end user. Each API is specialized in delivering a specific type of data according to its function. For example, the Google Maps Geocoding API converts geographic coordinates into readable postal addresses and vice versa, facilitating precise location mapping in digital applications. Meanwhile, the Places API provides detailed information about points of interest, including name, address, opening hours, user ratings, and reviews [24].

On the other hand, the Apple Maps Server API offers similar functionalities, allowing for geocoding queries, route calculation, and trip monitoring, with a particular focus on security and native integration within the Apple ecosystem. Both platforms support real-time data delivery via secure channels and standardized interfaces, ensuring high interoperability with mobile and web applications [13].

Efficient delivery of this data translates into instant responses to changing road conditions, including alternative route suggestions, congestion prevention, and travel time optimization. APIs return real-time information to the user such as traffic alerts, incidents, roadworks, speed limits, weather forecasts, and adaptive route suggestions. This rapid response capability significantly enhances the reliability and usability of navigation applications.

Furthermore, through integration with artificial intelligence algorithms and predictive analytics, these systems are capable of anticipating traffic patterns and customizing navigation based on individual user preferences and behavior. As a result, data delivery via APIs not only enables more informed navigation but also ensures a safer, more efficient, and personalized user experience.

9 Privacy, Security, and Ethics in Data Management

Modern navigation applications — such as TomTom, Google Maps, Apple Maps, and HERE — implement a robust set of mechanisms to ensure the protection and security of users' location data. These measures are designed not only to ensure legal compliance but also to reinforce user trust and transparency. First and foremost is the principle of informed consent, whereby location data is collected only with the user's explicit permission. At any time, users may choose to enable or disable location sharing based on their privacy preferences.

These platforms also follow the principle of data minimization, collecting only the information strictly necessary to deliver core functionalities — such as route calculation, traffic forecasting, or contextual alerts. Regarding data handling, data anonymization and aggregation are common practices, ensuring that the collected information cannot directly identify individuals. This enables the use of data to improve services and generate statistical insights without compromising user identity.

Additionally, technical and organizational security measures are implemented, including encryption protocols, strict access control, system monitoring, and protection against unauthorized access or cyberattacks. The companies behind these applications foster an organizational culture centered on privacy, investing in continuous staff training on data protection best practices and legal obligations.

It is also important to highlight compliance with key international data protection regulations, namely the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States. Navigation apps also provide user control mechanisms, allowing users to review, edit, or delete their data, as well as configure how their location is used or shared.

Finally, all third parties involved in the processing of personal data (such as technology partners or service providers) are contractually bound to comply with the same security and privacy standards, ensuring consistent protection throughout the entire data lifecycle. Together, these practices form a comprehensive and rigorous approach to data protection in navigation applications, contributing to a safer, more transparent, and more trustworthy user experience [25, 26, 27, 28, 29].

10 Conclusion

Real-time navigation APIs represent a significant advancement in how mobility data is collected, processed, and delivered to users, offering a more efficient, personalized, and safer driving experience. The integration of artificial intelligence, predictive algorithms, and data from multiple sources enables route optimization, incident forecasting, and real-time adaptation of routes to individual needs.

However, despite all the benefits they offer, these technologies are not neutral. As they become indispensable tools in urban mobility, they also raise ethical and social challenges, including the potential negative impact on certain communities.

In this context, it is essential to foster a deeper reflection on how these interfaces function, what effects they produce, and the criteria by which they are implemented, considering the urgent need to align technological innovation with spatial justice, equity, and sustainability. Such reflection provides the necessary foundation for future investigations aimed at observing, in practice, how these systems operate, whom they serve, and with what consequences—ultimately contributing to the development of more integrated, ethical, and socially responsible solutions in the field of smart and green mobility.

Acknowledgments: The study was developed under the project A-MoVeR – "Mobilizing Agenda for the Development of Products and Systems towards an Intelligent and Green Mobility", operation n.º 02/C05-i01.01/2022.PC646908627-00000069, approved under the terms of the call n.º 02/C05-i01/2022 – Mobilizing Agendas for Business Innovation, financed by European funds provided to Portugal by the Recovery and Resilience Plan (RRP), in the scope of the European Recovery and Resilience Facility (RRF), framed in the Next Generation UE, for the period from 2021 -2026

References

[1] Shubham Ganesh Junnarkar, Aadesh Kabadi, Purva Kadam, and Viraj Jarare. Optimizing traffic control with real-time data: A HERE API-Based Approach. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 7(11):1–5, 2023.

[2] Alben Rome Bagabaldo, Qianxin Gan, Alexandre M. Bayen, and Marta C. González. Impact of navigation apps on congestion and spread dynamics on a transportation network. *Data Science for Transportation*, 6(12), 2024.

[3] Akanbi Caleb, I Ogundoyin, and A Lawal. Implementing a university mobile navigation system. pages 143–150, 12 2014.

[4] Junghyun Lee. Introduction to navigation systems. In Rustam B. Rustamov, Sabina Hasanova, and Mahfuza H. Zeynalova, editors, *Multi-purposeful Application of Geospatial Data*, chapter 1. IntechOpen, Rijeka, 2017.

[5] Tom Barnard. Best Navigation Apps 2025. https://www.autoexpress.co.uk/product-group-tests/63573/best-navigation-apps-2025, Jan 2025. Accessed on May 30, 2025.

[6] Holiday Extras. Best Navigation Apps Available: Free for Android & iOP in 2025. https://www.holidayextras.com/travel-advice/best-navigation-apps.html, n.d. Accessed on May 30, 2025.

[7] Jiri Medlen. What is api, and what can you use it for? *ResearchGate*, February 2023.

[8] Lei Zhu, Jacob Holden, and Jeffrey Gonder. Navigation application programming interface route fuel saving opportunity assessment on large-scale real-world travel data for conventional vehicles and hybrid electric vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, 2672:036119811879780, 10 2018.

[9] Cheng Siong Chin, Teck Kai Chan, and Tong Rui Peng. Intelligent urban noise mapping by google api network edge device. In *2020 International Conference on Internet of Things and Intelligent Applications (ITIA)*, pages 1–4, 2020.

[10] Google Developer Documentation. Google maps routes api documentation. https://developers.google.com/maps/documentation/routes?hl=pt-br, 2025. Accessed on Jun 03, 2025.

[11] Google Developers. Waze for developers. https://developers.google.com/waze?hl=pt-br, 2024. Accessed on Jun 03, 2025.

[12] Liv Meinck. 3 ways to excel at predicting traffic incidents. https://www.tomtom.com/newsroom/product-focus/predicting-traffic-incidents/, 2021. Accessed on Jun 09, 2025.

[13] Apple Inc. Apple maps server api documentation. https://developer.apple.com/documentation/applemapsserverapi/, n.d. Accessed on Jun 05, 2025.

[14] TomTom. Tomtom traffic stats: Introduction. https://developer.tomtom.com/traffic-stats/documentation/product-information/introduction, 2023. Accessed on Jun 09, 2025.

[15] Xiangpeng Wan, Hakim Ghazzai, and Yehia Massoud. Real-time navigation in urban areas using mobile crowd-sourced data. In *2019 IEEE International Systems Conference (SysCon)*, pages 1–7, 2019.

[16] Victor Suday. Optimizing travel: How ai-powered navigation apps are transforming road trips. https://mapmetrics.org/pt/blog/ai-navigation-apps-estao-transformando-viagens-rodoviarias/, 2024. Accessed: Jun. 11, 2025.

[17] HERE Technologies. Here application and here maps privacy supplement. https://legal.here.com/br-pt/privacy/here-application-and-here-maps-privacy-supplement, 2022. Accessed on Jun 09, 2025.

[18] National Traffic Designs. Traffic data collection and analysis by consultants. https://www.nationaltrafficdesigns.com/traffic-data-collection-and-analysis-by-consultants/, 2024. Accessed on Jun 09, 2025.

[19] Yuna Kim, Sangho Song, Hyeonbyeong Lee, Dojin Choi, Jongtae Lim, Kyoungsoo Bok, and Jaesoo Yoo. Regional traffic event detection using data crowdsourcing. *Applied Sciences*, 13:9422, 08 2023.

[20] Kmuller. How crowdsourcing is changing the waze we drive. https://d3.harvard.edu/platform-rctom/submission/how-crowdsourcing-is-changing-the-waze-we-drive/, 2018. Accessed on Jun 11, 2025.

[21] Victor Suday. Live traffic updates on the road: How real-time data is revolutionizing navigation. https://mapmetrics.org/pt/blog/live-traffic-updates-on-the-road/, 2024. Accessed on Jun 09, 2025.

[22] Clube do GIS. What is geocoding? https://clubedogis.com.br/glossario/o-que-e-geocodificacao/, 2025. Accessed: Jun. 11, 2025.

[23] Siam Yamsaengsung and Borworn Papasratorn. Towards improving user interaction with navigation apps: an information quality perspective. *KnE Social Sciences*, pages 119–131, 2017.

[24] Google Developers. Api picker — google maps platform. https://developers.google.com/maps/documentation/api-picker?hl=pt-br, 2025. Accessed: June 17, 2025.

[25] TomTom. Tomtom privacy policy. https://home.tomtom.com/media/html/privacy_policy.html, 2006. Accessed: June 14, 2024.

[26] Waze. Waze privacy policy. https://www.waze.com/en/legal/privacy, 2020. Accessed: June 14, 2024.

[27] Google LLC. Google privacy policy. https://policies.google.com/privacy?hl=en-US, 2024. Accessed: June 14, 2024.

[28] Apple Inc. Apple maps and privacy. https://www.apple.com/legal/privacy/data/en/apple-maps/, 2025. Accessed: June 14, 2024.

[29] HERE Technologies. Here privacy policy. https://legal.here.com/en-gb/privacy, 2024. Accessed: June 14, 2024.

Chapter 29

Maintenance Management in Electrical Grids with High Renewable Energy Penetration: a review of challenges and trends

Contribution #54
Cite as:
Joice Agostinho, José Farinha, Ana Vieira, Hugo Raposo. Maintenance Management in Electrical Grids with High Renewable Energy Penetration: a review of challenges and trends. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-54,
  title ={Maintenance Management in Electrical Grids
     with High Renewable Energy Penetration: a review
     of challenges and trends},
  author={Joice Agostinho and  José Farinha and  Ana
     Vieira and  Hugo Raposo},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Maintenance Management in Electrical Grids with Renewable Energy High Penetration: a review of challenges and trends

Joice Agostinho[1], José Torres Farinha[1,2]; Ana V Vieira[1,2,3,4]; Hugo Raposo[1,2]

[1] Instituto Superior de Engenharia de Coimbra, Instituto Politécnico de Coimbra, Rua Pedro Nunes, 3030-199 Coimbra, Portugal
[2] RCM2+ Research Centre in Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Higher Institute of Engineering of Coimbra, Rua Pedro Nunes-Quinta da Nora, 3030-199 Coimbra, Portugal
[3] CISE - Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal
[4] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal
[*]{joice.agostinho@outlook.com }; tfarinha@isec.pt; avieira@isec.pt ; hugo.raposo@isec.pt

Abstract

The increasing integration of renewables into electrical grids has led to a significant reconfiguration of energy systems, requiring new maintenance approaches to ensure reliability, efficiency and sustainability. This study, motivated by the challenges of the energy transition, conducts a critical review of the literature on practices, technologies and trends in the maintenance management of electrical networks. The methodology adopted focuses on the analysis of strategies adopted in Portugal and other countries with a strong penetration of renewables, highlighting more effective approaches. Maintenance policies are analysed, with a focus on predictive maintenance, leveraged by technologies such as IoT, Artificial Intelligence, Augmented Reality and Robotics. The results show operational gains and point to the importance of microgrids and architectures with greater redundancy as strategies to strengthen the resilience of the grid face of extreme weather events and localized failures. The analysis reveals that the modernization of the physical assets, combined with technological innovation, is crucial to ensure the continuity and stability of the electricity system. This work contributes to the state of the art by integrating international perspectives and highlighting practical solutions that can guide public policies and strategic decisions in the energy sector.

Keywords: Asset management, maintenance, sustainability, renewable energy, electrical grids

1 Introduction

The energy transition has become a global priority in search of a more sustainable and resilient energy systems. The increase in renewable energy sources represents a milestone in this process but also imposes significant challenges to the operation and maintenance of electrical grids, namely intermittency, variability of production and grid modernisation.

Traditional maintenance models are insufficient facing the complexity of modern systems, which has led to the adoption of data-driven approaches, with an emphasis on predictive maintenance.

In this context, the maintenance of electrical infrastructures plays a decisive role in ensuring the continuity of service provision and maximizing asset performance, reflecting the complexity and need for efficiency in the operations of these systems.

2 State of the Art

2.1 Maintenance Management

According to NP EN 13306:2021, "maintenance is a combination of all technical, administrative and management actions, during the life cycle of an asset, aimed at maintaining or restoring it to a state in which it can perform the required function".

For an organization to achieve excellence in its services and products, it is essential to manage maintenance, aiming to find a balance between benefits and costs, to maximize the positive contribution of maintenance to the profitability of the investments made.

Maintenance management is a strategic function in industrial systems and infrastructures, especially when it comes to critical assets such as power grids. Effective maintenance management allows to increase the reliability and availability of equipment, reduce failures and operating costs, ensure safety and service continuity.

Maintenance management is essential to ensure the safe and efficient operation of power grids. Traditionally, a combination of three main strategies is applied: corrective maintenance, preventive maintenance, and predictive maintenance. Predictive maintenance stands out for allowing a data-driven approach and continuous monitoring, with the aim of "predict and avoid" rather than "react and repair" [1].

In the context of modern electrical grids, marked by increasing operational complexity and penetration of renewable sources, predictive maintenance plays a strategic role. This approach contributes significantly to reducing operational costs, increasing efficiency in managing critical network assets, and minimizing disruptions, even in high-variability scenarios.

3 Technologies used in maintenance management

Technology is the main catalyst for the digitalization of the electricity sector and, consequently, for the energy transition. The technologies mentioned here have been fundamental for inspections of critical lines, combined with analytical and predictive models that help in vegetation management and failure prevention.

3.1.1 Internet of Things (IoT)

IoT sensors provide data on the state of equipment in real time, which, in turn, facilitates the automation of processes, as well as the improvement of decisions to be made. In distribution systems, for example, IoT sensors installed in transformers can identify anomalous heating or irregular vibrations [2].

3.1.2 Artificial Intelligence (AI)

They enable predictive analytics to predict failures and optimize operations. The ability to create algorithms and technological models makes it possible to automate processes, prevent patterns, and optimize operations. AI learns from previous failure patterns and can suggest optimal maintenance and replacement periods for components [2].

3.1.3 Augmented Reality

Augmented reality offers the visualization and interaction of digital information in a physical or virtual environment. In addition, to enabling enriching customer experience, it facilitates maintenance by providing clear and visual instructions for employees, reducing errors in operations [2].

3.1.4 Robotics

Robotics is used to perform fast and safe infrastructure inspections, minimizing the need for human intervention in hazardous environments.

3.2 Microgrids: Concept, Components and Role in the Energy Transition

The integration of renewable sources into an electricity system, given its intermittent and decentralised nature, imposes technical challenges, namely in terms of forecasting and grid stability. To ensure stability, energy storage systems and smart grids are needed to increase the flexibility and responsiveness of the grid [3].

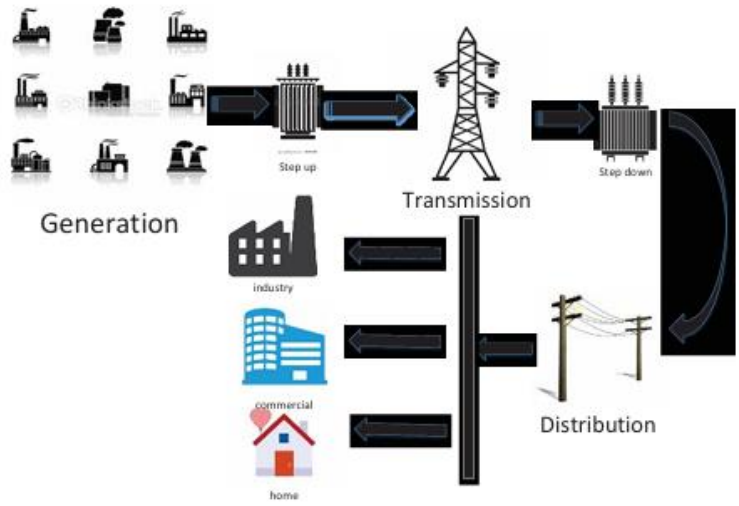


Figure 1 - Conventional network [4]

The concept of microgrids emerges as a response to the challenges of integrating renewable energy sources, since interruptions in this grid are locally restricted and do not affect the entire grid as in the traditionally used systems, represented in Figure 1 [5][4].

Microgrids can operate autonomously or in connection with the main electrical grid, offering flexibility, energy security and a better use of renewable energies.

The use of decentralized energy systems improves the ability of grids to absorb the intermittency of renewable sources, balancing energy supply and demand in a more sustainable way [6].

4 Cybersecurity in electrical grids

The digitalisation of electrical grids with high penetration of renewables is a significant milestone in the modernisation of the energy sector, transforming the way energy is generated, produced and consumed. However, this technological transformation introduces a new range of cyber threats, which compromise its security, reliability and resilience of infrastructures. Many of these threats are complex, discrete, and difficult to detect, requiring monitoring and defence solutions.

The increasing interconnectivity of networks, fuelled by the proliferation of the use of IoT devices and digital control systems, exposes systems to a variety of cyberattacks, such as distributed denial-of-service, false data injection, ransomware, and even coordinated physical sabotage. These attack modalities exploit flaws in communication protocols, poor authentication and the absence of security updates on peripheral devices, posing significant risks to the continuous operation of the network [7].

In this scenario, technological developments, while essential for the sustainable energy transition, must be accompanied by an equally sophisticated and integrated cybersecurity approach.

The challenge lies in reconciling innovation with protection, ensuring that operational gains do not compromise the robustness of the electrical system. Cybersecurity, therefore, is no longer a technical complement and becomes a pillar for the energy transition. It is imperative that security policies keep pace with technological innovation.

5 Portuguese National Electric System

Portugal faces challenges in the integration of renewable sources, especially due to the significant increase in the penetration of solar and wind energy in the Portuguese national electric system (*Sistema Elétrico Nacional*-SEN) in recent years. Since 2014, Portugal has seen a significant increase in renewable energy production capacity, reflecting its commitment to the energy transition (Figure 2). In addition to investing in electricity production from renewable sources, Portugal has invested in the modernization of its electrical grid, including the use of storage systems and smart grids to mitigate the impacts of intermittency from renewable sources [8][9].

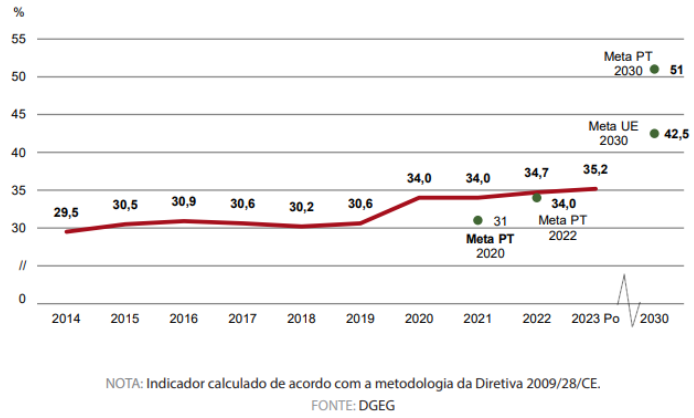


Figure 2 - Contribution of renewable energy sources to gross final consumption

Despite advances in electricity production from renewable energies, Portugal continues to have a certain energy dependence on interconnection with Spain, which represents a strategic vulnerability, as was evident in the blackout of April 28, 2025. On this day, a failure in the Spanish electrical system caused instability that quickly spread to the Portuguese grid. This event highlighted the importance of strengthening the autonomy and resilience of the SEN, as well as investing in local solutions.

The SEN, aware of these challenges, has resorted to the use of drones and IoT sensors for the remote inspection of lines and substations (as illustrated by the historical evolution illustrated in Figure 3), in addition to the implementation of AI-based systems for the prediction of failures in critical equipment, thus reinforcing the reliability of the system and the continuity of service [10][11][12].

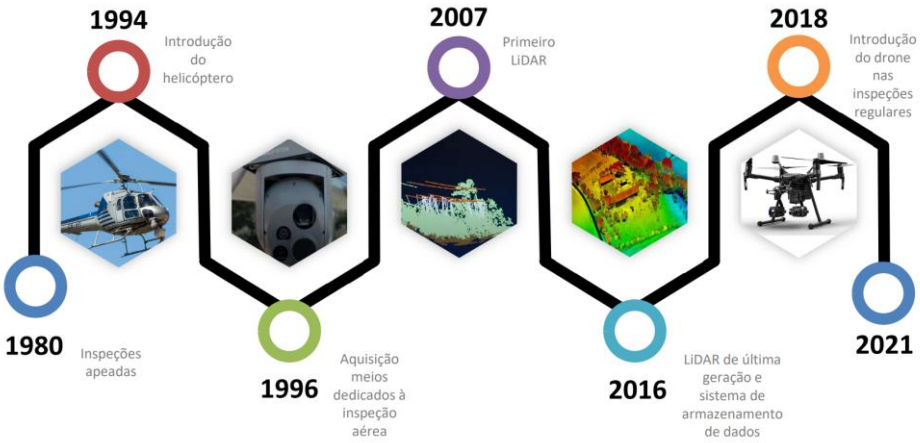


Figure 3 - Historical perspective: Inspection of electrical assets and their lanes [12]

It should be noted that one of the main focuses of investment and modernization in the SEN has been the improvement of service quality standards, reducing asymmetries in the national territory. The effects of this investment have been translated into good results in terms of quality of commercial service, and quality of technical service, particularly those related to continuity of service. As an example, the case of MV networks was mentioned, where there was a progressive reduction in the indicator referring to the Average Duration of Long Interruptions of the System Average Interruption Duration Index (SAIDI), and a reduction of eight percentage points, between 2017 and 2023, in the Average Frequency of Interruptions of the System Average Interruption Frequency Index (SAIFI) [10][11].

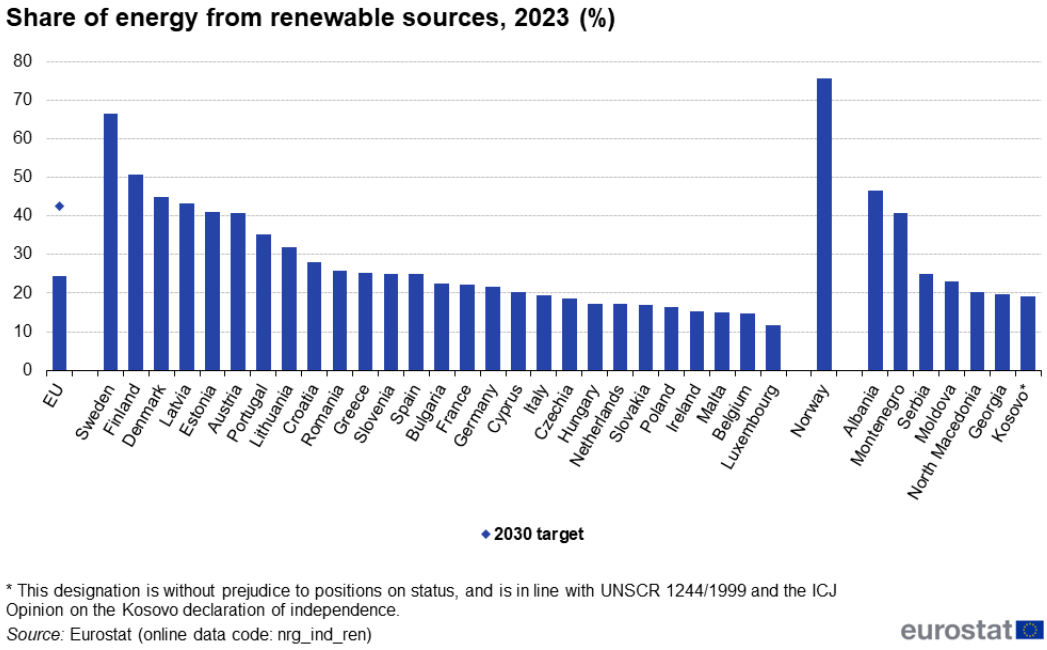


Figure 4 - Share of energy from renewable sources in 2023 [13]

According to data from EUROSTAST (Figure 4), Portugal is positioned as one of the countries in the European Union with the highest integration of renewable energies. The European countries that have made the most progress in the energy transition share well-defined and coordinated energy policies. They have set ambitious targets for reducing greenhouse gases and increasing renewable production, in line with the energy guidelines set by the European Parliament. To boost this transition, they implemented several incentive mechanisms, such as support for self-consumption, tax benefits, and financing of sustainable projects. In addition, they invest significantly in technological innovation and in the improvement of energy infrastructure. They also bet on the diversification of renewable sources, the integration of electricity systems and the promotion of energy efficiency [13] [14].

In the field of cybersecurity, European countries established a comprehensive framework to support Member States, businesses and citizens in the fight against cybersecurity threats and attacks. However, continuous development and improvement in data protection mechanisms and related services remain essential[10][15]. The 2024 Report on State of Cybersecurity in the union of European Union Agency for Cybersecurity (ENISA) presents an evidence-based analysis of cybersecurity posture across EU Member States. The report identifies persistent and emerging threats such as ransomware, distributed denial-of-service (DDoS) attacks, and supply chain compromises as significant risks to the Union's digital resilience [15].



Figure 5 - EU Legislative Landscape on Cybersecurity [15]

Figure 5 provides a structured overview of the European Union's evolving legislative framework related to cybersecurity. It categorizes legislative instruments across three levels of relevance. The colour coding illustrates legislative progress: green indicates legislation under implementation, yellow reflects instruments that have been politically agreed upon but not yet fully enacted, and blue denotes legislative acts still under development. This visual map highlights both the depth and maturity of EU cybersecurity policy and the need for continued integration of security principles across all digital sector [15].

It is expected that investment in preventive maintenance and technologies to support the diagnostics of the different SEN networks will be maintained, since the "*modernization of the network is decisive for increasing its resilience, allowing the aging and degradation of assets*". It is therefore important to invest in the renewal of assets, but also in supervision and automation (in accord to Electricity Quality Monitoring Plans) as a way to promote the continuous improvement of the quality of service, but also in matters of cybersecurity, for example through the "Intelligent Systems for Supervision and Operation and Telecommunications" program [10][11][12].

6 Conclusion

The integration of renewables into electrical grids represents a crucial step forward for sustainability, but it also poses significant technical challenges, especially regarding the management of infrastructure maintenance and resilience. This article highlighted that modern approaches are essential to ensure reliability, security, and continuity of service. It is recommended that companies invest in emerging technologies and technical capacity building to maximize the benefits of the energy transition and ensure the sustainability of electrical grids. However, the challenges and solutions associated with the maintenance of electrical grids, in a scenario of increasing penetration of renewable energies, require constant innovation, collaboration and adaptation of electricity infrastructures.

The benefits of modernising electrical grids can only be fully reaped if they are accompanied by an integrated and strategic approach to cybersecurity.

In this regard, future research could explore the integration of emerging technologies such as blockchain for decentralized grid development or edge computing for real-time maintenance diagnostics. Another promising line of inquiry involves the incorporation of advanced cybersecurity frameworks to ensure both system resilience and data integrity. In addition, comparative studies between countries with

renewable energy high penetration, such as Portugal, could yield valuable insights into the scalability and adaptability of maintenance and automation practices, particularly in response to extreme weather events and cross-border interconnection vulnerabilities.

References

[1] C. Matos, "Gestão da Manutenção e da Reserva Estratégica em Subestações Elétricas Caso de Estudo da EDP Distribuição," 2021.

[2] D. Hickie and J. Hickie, "The impact of Industry 4.0 on supply chains and regions: innovation in the aerospace and automotive industries," *European Planning Studies*, vol. 29, no. 9, pp. 1606–1621, Sep. 2021, doi: 10.1080/09654313.2021.1963048.

[3] F. Daniel-Durandt, C. Y. Van Staden, B. Bekker, and A. Rix, "Innovative Solutions Addressing South Africa's Energy Paradox: Integrating Renewable Energy Resources for Increased Power System Stability and Maintenance Schedule Planning," in *Proceedings - 24th EEEIC International Conference on Environment and Electrical Engineering and 8th I and CPS Industrial and Commercial Power Systems Europe, EEEIC/I and CPS Europe 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/EEEIC/ICPSEurope61470.2024.10751160.

[4] P. Kalkal and V. K. Garg, "Transition from conventional to modern grids: Modern grid include microgrid and smartgrid," in *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2017, pp. 223–228. doi: 10.1109/ISPCC.2017.8269679.

[5] D. Lindley, "Smart grids: The energy storage problem," *Nature*, vol. 463, no. 7277, pp. 18–20, 2010, doi: 10.1038/463018a.

[6] L. Xu *et al.*, "Resilience of renewable power systems under climate risks," *Nature Reviews Electrical Engineering*, vol. 1, no. 1, pp. 53–66, Jan. 2024, doi: 10.1038/s44287-023-00003-8.

[7] J. P. Yaacoub, H. Noura, J. Azar, O. Salman, and K. Chahine, "Cybersecurity in Smart Renewable Energy Systems," in *2024 International Wireless Communications and Mobile Computing (IWCMC)*, 2024, pp. 1534–1540. doi: 10.1109/IWCMC61514.2024.10592561.

[8] L. Bird *et al.*, "Wind and solar energy curtailment: A review of international experience," Nov. 01, 2016, *Elsevier Ltd*. doi: 10.1016/j.rser.2016.06.082.

[9] EDP, "Relatório Anual Integrado 2023." Accessed: May 08, 2025. [Online]. Available: https://www.edp.pt/quem-somos

[10] ERSE, "Plano de Desenvolvimento e Investimento da Rede Nacional de Distribuição 2026 a 2030." Accessed: May 08, 2025. [Online]. Available: https://tinyurl.com/22kb428e

[11] E-Redes, "Relatório da Qualidade de Serviço 2023 – energia em Rede." Accessed: May 08, 2025. [Online]. Available: https://tinyurl.com/3k5nuev7

[12] Labelec, "Inspeção de ativos elétricos e respetivas faixas com recurso a LIDAR," 2021, Accessed: May 08, 2025. [Online]. Available: https://tinyurl.com/54xx8mzp

[13] EUROSTAST, "Renewable Energy Statistics." Accessed: May 21, 2025. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics

[14] Parlamento Europeu, "Política industrial, energética e de investigação," Fichas Temáticas sobre a União Europeia.

[15] European Union Agency for Cybersecurity, "2024 Report on the State of the Cybersecurity in the Union." Accessed: Jun. 01, 2025. [Online]. Available: https://www.enisa.europa.eu/publications/2024-report-on-the-state-of-the-cybersecurity-in-the-union

Chapter 30

An electric mobility software ecosystem: Lifecycle Management Implementation

Contribution #55
Cite as:
David Dias, António Paulino, Arsénio Reis. An electric mobility software ecosystem: Lifecycle Management Implementation. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-55,
  title ={An electric mobility software ecosystem:
     Lifecycle Management Implementation},
  author={David Dias and  António Paulino and  Arsénio
     Reis},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

An electric mobility software ecosystem: Lifecycle Management Implementation

David Dias[1,*], António Paulino[3], Arsénio Reis[12]

[1] University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[2] INESC-TEC, Vila Real, Portugal
[3] Polytechnic University of Coimbra, Coimbra, Portugal

* Corresponding author (al75099@alunos.utad.pt)

Abstract

Around the world, as the adoption of electric vehicles has grown, digital platforms have been required to ensure efficient management of their life cycle and integration with external services. The A-Mover project incorporates a completely new concept for electric mobility, composed of several modules, and aligned with the software ecosystem paradigm. One such module is the lifecycle management (LCM), which requires robust integration between internal modules and external applications. In this project, the LCM modules are responsible for production logging, maintenance scheduling and operational monitoring, while external applications are allowed to send real-time messages and alerts to the vehicle dashboard and user's mobile application. The communication architecture guarantees secure and scalable data exchange between components. This study focuses on analyzing aspects such as authentication, data synchronization and service integration systems. Both qualitative and quantitative research is carried out based on the architecture and communication flows. The implemented solution demonstrates effectiveness in real-time operation and response delivering messages, using simplified mechanisms that reduce system complexity. The results demonstrate that a structured communication model can support reliable coordination between distributed applications in connected mobility environments.

Keywords: Electric mobility, interoperability Lifecycle management (LCM), system integration, real-time communication

1 Introduction

As electric mobility becomes more widespread, it's increasingly obvious that we need digital platforms that can manage the entire lifecycle of electric vehicles and do it in a way that's smooth and well-connected [1]. These platforms have to bring together all the internal systems like those handling production, maintenance, and daily operations with external apps that provide things like alerts, third-party services, and data integration. It's crucial that all of these systems can talk to each other in real time, because that directly affects how safe the vehicles are, how efficiently they run, and how quickly users can get the support they need [2].

That's where the AMoVeR project comes in. It was built from the ground up to meet these challenges head-on. The platform is modular and scalable, built specifically to manage every stage of an electric motorcycle's lifecycle. At its core are the LCM (Life Cycle Management) modules, which take care of everything from tracking production and scheduling maintenance to coordinating backend services. These modules are fully connected with external apps that handle notifications and send real-time updates whether to the bike's dashboard (FED) or to the rider's mobile app [3].

In this article, we dive into how AMoVeR's communication system works, how the different pieces are wired together, how the messages flow between them, and what technical choices were made to support all of that.

2 Ecosystem

2.1 Mobility System Architecture

The AMoVeR system was developed to offer integrated digital management of all phases of the electric motorcycle life cycle. Its architecture distributes responsibilities among different internal modules and external applications, promoting a clear separation of functions and facilitating the scalability and maintenance of the ecosystem [4].

At the core of the infrastructure are the **Lifecycle Management (LCM)** modules, organized as follows:

- **LCM-PRD (Production)**: records essential vehicle manufacturing data, such as serial number, battery and controller.
- **LCM-MNT (Maintenance)**: allows workshops to schedule and record technical interventions, inspections and repairs.
- **LCM-SRV (Services)**: acts as a central service, coordinating data storage and communication between all the modules.

The system also includes several peripheral components:

- **UVMA (User Vehicle Mobile App)**: provides users with real-time access to the vehicle's operational status, maintenance records, and alerts.
- **CSVU (Connected Service and Vehicle Unit)**: embedded in the motorcycle, it collects data from onboard sensors and transmits it securely to the cloud backend.
- **FED (Front-End Display)**: the dashboard unit that shows technical and contextual information to the rider.
- **Third-party Applications**: external systems integrated via secure REST APIs to enhance the ecosystem with real-time contextual messages, alerts, and service integrations.

The system communicates with in-vehicle devices, notably the **CSVU (Vehicle Service and Connection Unit)**, which collects operational data from sensors and transmits it to the cloud. The processed information is presented to the driver via the **FED (Front-End Display)**.

On the user side, the mobile application (**UVMA**) enables users to view vehicle status, access maintenance history and receive notifications. Additionally, the ecosystem integrates external applications that communicate contextual messages—such as alerts or instructions—to the dashboard or user app via secure channels.

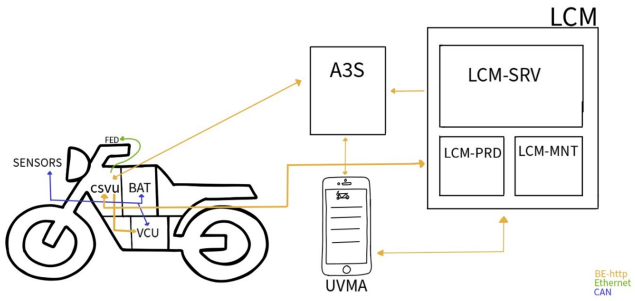


Figure 1: Mobility System Architecture

This modular architecture ensures flexibility, allowing new services or applications to be added without impacting existing modules, and lays a robust foundation for urban electric mobility environments.

2.2 Communication Model

Messaging between the various modules follows a well-organized logic, designed to ensure that everything gets to the right place safely and efficiently. Whenever there's a message — whether it's a production record, a maintenance alert from the LCM modules, data sent by the on-board unit (CSVU), or even information from external services everything goes through the LCM-SRV first.

This module acts as a kind of "checkpoint" that checks everything before letting it go. It checks that the message has the right data such as who sent it, what type of event it is, the time it was generated and also that it complies with the system's security rules. After that, the message is sent where it should go: it can be to the bike's dashboard (FED), to the user's app (UVMA), or both, depending on the type of information.

To ensure that the system can cope well with growth and possible failures, we opted for asynchronous event-based communication [5]. This allows messages to be queued, resent if something fails, and the status of each one to be monitored. And the best thing is that even if new data sources or destinations are added, the main logic doesn't need to be changed.

This routing strategy is designed to be as flexible as possible. Each part of the system can communicate without being tied to another, which makes everything more modular, easier to maintain and much more resistant to failures. It's an architecture that's ready to grow without complications.

2.3 Message Routing

All the messages exchanged between the different modules follow a well-defined and centralized scheme. The idea is simple: to ensure that the information goes to the right place, safely and without complications. Regardless of where the message comes from—be it from an LCM module (such as when a production is recorded or a maintenance alert appears), from the unit installed on the bike (CSVU), or even from external services—everything goes through the LCM-SRV first.

This module acts as a "control center" of the system. It receives the messages, checks that they contain all required metadata (sender, event type, timestamp) and enforces security and

integrity policies. It then forwards the message to its destination—such as the bike's dashboard (FED), the user's app (UVMA), or both—depending on content and priority.

To ensure that the system remains scalable and robust under load or in case of failures, AMoVeR adopts an asynchronous event-driven communication model that supports queuing and retry mechanisms, along with delivery status monitoring—key traits identified in highly scalable IoT brokers like TBMQ [6]. This architecture allows new message sources or consumers to be added without modifying the core logic.

Such a decoupled routing model ensures each component can evolve independently, supporting future expansion and enhancing reliability across the ecosystem, consistent with best practices in event-driven microservices design [7].

3 Integration

3.1 Integration of Lifecycle Management (LCM) Modules

The AMoVeR architecture is founded on three basic Lifecycle Management (LCM) modules, which include LCM-PRD (production), LCM-MNT (maintenance), and LCM-SRV (services). The three modules communicate through REST APIs, following a service-oriented architecture [8]. The design enhances clear responsibility segregation and decoupling between components.

LCM-PRD is applied for registration of newly produced electric motorcycles, e.g., serial number, battery type, and accompanying electronic components. The LCM-MNT module is employed by repair workshops to plan and record maintenance, inspections, and repair work and maintains a complete history of interventions over time [9].

The two modules are connected to the LCM-SRV, which acts as the master controller. The module processes the information received, validates the integrity of the information, and applies designed business rules. It also transfers information to the desired locations, such as the user mobile app (UVMA) or the front-end display in the car (FED), in a synchronized fashion across all the components.

Data access security is ensured through the OAuth 2.0 protocol, governed by the A3S module (Authentication, Access and Audit Service), which grants authentication tokens only for licensed applications [10]. This layer is crucial, especially when handling sensitive data and real-time decisions.

This module-based, service-oriented architecture allows the system to be horizontally scaled with fresh instances of modules being deployed as the fleet grows. The architecture also allows new modules to be introduced in the future without changing the existing ones [11], thus the system is simpler to maintain and scale.

3.2 Interaction with the User Mobile Application (UVMA)

UVMA (User Vehicle Mobile App) is the mobile application that shares a direct link between the user and the AMoVeR digital platform. Using it, the motorcycle owner can check the battery level, maintenance history, running notifications, and other services' notifications linked to the platform [12].

The interaction of UVMA with the back-end is done via gRPC and REST APIs exposed by the LCM-SRV module whereas the authentication is handled by the A3S service. The information of a vehicle is made accessible to an authenticated user, adhering to security and privacy guidelines [13].

The application receives notifications, and these can be from various sources: from maintenance notifications (via LCM-MNT), from production information (via LCM-PRD), from operation information from the CSVU (Service and Communication Unit), or from messages of third-party services. The LCM-SRV centralizes and filtrates this information in a way that the user only receives relevant data at the right time.

Besides that, UVMA supports location or system event contextual notifications citejiang2022mobility. Users can reply to the messages, activate features, confirm actions, or answer requests, enabling smarter and more aware mobility.

3.3 Integration with External Applications

In addition to its own modules, the AMoVeR system offers support for third-party applications that complement the functions of the platform. These are mobility services, traffic notification, weather warnings, or insurance applications citewang2023thirdparty.

Its interaction with these third-party applications is done via REST APIs secured by OAuth 2.0. The module of LCM-SRV is a gateway that only allows approved services to send or receive data from the platform [14]. The criticality of certain messages (e.g., real-time alerts on the FED) makes this layer of security indispensable.

To maintain performance and avoid bottlenecks, the framework is asynchronous event-driven and decoupled in structure. Each message is pipelined through validation to ensure its authenticity, validity, and priority before it is passed to the intended targets [15].

This integration capability makes AMoVeR a scalable, extensible, and adaptive system. Thanks to its modular mechanisms and well-defined services, the system can be integrated with emergency networks, urban mobility operators, or smart city infrastructures in the future, dynamically responding to the evolving needs of the ecosystem. wA

4 Conclusion

The AMoVeR project has shown that it is perfectly feasible to adopt a modular and scalable architecture to manage the entire life cycle of electric motorcycles. By dividing the system into specific modules for production, maintenance and coordination (LCM-PRD, LCM-MNT and LCM-SRV), we were able to separate responsibilities well and ensure that everything works seamlessly, both between internal components and with external applications.

The communication architecture, based on asynchronous and secure messaging, made it possible to create real-time interactions with the vehicle (via the EDF) and with the user (via UVMA). In addition, the fact that the system is prepared to integrate third-party services was essential to ensure flexibility and respond to the different contexts and needs of today's mobility.

Although the solution already fulfills the objectives for which it was designed, there is room to evolve. It will be important, for example, to improve the monitoring of communication flows, optimize the use of resources at times of greater load and explore more advanced messaging patterns - such as the use of priorities or contextual routing.

For the future, improvements such as the integration of predictive analysis for maintenance, dynamic configuration of connections to external services and standardization efforts to ensure that the system remains compatible with other broader mobility platforms are also on the table.

Acknowledgment

The study was developed under the project A-MoVeR – "Mobilizing Agenda for the Development of Products and Systems towards an Intelligent and Green Mobility", operation n.º 02/C05-i01.01/2022.PC646908627-00000069, approved under the terms of the call n.º 02/C05-i01/2022 – Mobilizing Agendas for Business Innovation, financed by European funds provided to Portugal by the Recovery and Resilience Plan (RRP), in the scope of the European Recovery and Resilience Facility (RRF), framed in the Next Generation UE, for the period from 2021 -2026.

References

[1] Logan Butler, Tan Yigitcanlar, and Alfredo Paz. Smart urban mobility innovations: A comprehensive review and evaluation. *IEEE Access*, 8:211182–211211, 2020.

[2] Xinyu Chen, Yanyan Liu, and Ling Zhang. Secure and scalable communication framework for electric vehicle fleets. *IEEE Internet of Things Journal*, 2023.

[3] Mohammadali F. Arahoor et al. Comprehensive iot-driven fleet management system for industrial vehicles. *International Journal of Industrial and Applied Systems Engineering*, 15(1):55–67, 2023.

[4] Wei Zhang, Jing Liu, and Hao Chen. An event-driven architectural model for integrating heterogeneous systems. *Journal of Systems Architecture*, 140:102090, 2023.

[5] Mauro Ferrari, Andrea Bonomi, and Vincenzo D'Andrea. Message routing and middleware architectures in distributed iot applications. *Journal of Systems and Software*, 190:111336, 2022.

[6] Dmytro Shvaika, Andrii Shvaika, and Volodymyr Artemchuk. Tbmq scalability and reliability in distributed iot systems. *IoT*, 6(3), 2025.

[7] Ashwin Chavan. Exploring event-driven architecture in microservices: Patterns, pitfalls and best practices. *International Journal of Science and Research Archive*, 04(01):229–249, 2021.

[8] Mehul Patel and Ankit Kumar. Service-oriented architectures for fleet management: A survey. *International Journal of Web Services Research*, 19(1):45–67, 2022.

[9] Alan Smith and Emily Johnson. Modular architecture patterns for scalable iot ecosystems. *IEEE Internet of Things Magazine*, 4(3):22–29, 2021.

[10] Marina Torres, João Silva, and Manuel Costa. Implementing oauth 2.0 to secure communication in iot-based mobility systems. *Journal of Mobile Computing*, 8(2):115–130, 2023.

[11] Wei Zhang and Xiaoyu Chen. Design of scalable and maintainable iot systems: A modular approach. *Journal of Systems Architecture*, 117:102106, 2021.

[12] Thanh Nguyen and Jie Liu. Real-time data visualization and analytics for smart mobility services. *IEEE Internet of Things Journal*, 10(4):3112–3123, 2023.

[13] Ahmedur Rahman and Farid Chowdhury. Security and privacy in vehicle-to-cloud communications: A review of current solutions and future directions. *Vehicular Communications*, 35:100472, 2022.

[14] Mingwei Liu and Wei Xu. Secure api gateways for vehicle-to-cloud platforms using oauth 2.0 and microservices. *Vehicular Communications*, 33:100443, 2022.

[15] Xiaolong Deng and Sungwoon Park. Scalable and modular architecture for smart vehicle ecosystems: A microservice-based approach. *Journal of Systems Architecture*, 139:102861, 2023.

Chapter 31

A ROS 2-Based Modular Control Architecture Integrating Visual Servoing and Model Predictive Control for Robotic Marking

Contribution #56
Cite as:
Aibar Tanatov, Batyrkhan Omarov, Moldir Kuatova, Mateus Mendes. A ROS 2-Based Modular Control Architecture Integrating Visual Servoing and Model Predictive Control for Robotic Marking. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-56,
  title ={A ROS 2-Based Modular Control Architecture
     Integrating Visual Servoing and Model Predictive
     Control for Robotic Marking},
  author={Aibar Tanatov and  Batyrkhan Omarov and
     Moldir Kuatova and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

A ROS 2-Based Modular Control Architecture Integrating Visual Servoing and Model Predictive Control for Robotic Marking

Aibar Tanatov[1,2,*], Batyrkhan Omarov[1], Moldir Kuatova[1,3], Mateus Mendes[4]
[1] Joldasbekov Institute of Mechanics and Engineering, Almaty, Kazakhstan
[2] Farabi University, Almaty, Kazakhstan
[3] International Engineering and Technological University, Almaty, Kazakhstan
[4] Polytechnic University of Coimbra, Portugal
[*] Corresponding author tangatov_aibar1@kaznu.edu.kz

Abstract

This paper introduces a ROS 2-based intelligent control architecture designed for high-precision robotic marking, integrating model predictive control (MPC), dual-camera visual servoing, and embedded actuation through STM32 microcontrollers. The system utilizes a multilayered control stack that combines global perception from an overhead camera with fine-grained local correction via a tool-mounted camera, enabling accurate 6-DoF pose estimation through ArUco markers and hand-eye calibration. Trajectory planning is achieved using CasADi-based constrained optimization, allowing real-time, adaptive motion generation within workspace and hardware constraints. Implemented on a 6-DOF collaborative robotic arm with actuator control based on RS485, the architecture supports submillimeter precision in marking tasks and demonstrates low latency performance in visual, computational, and actuation pipelines. Experimental results confirm the effectiveness of the approach, showing final errors as low as 0.7 mm and control loop frequencies of 7 to 9 Hz. This work highlights the viability of combining vision, learning, and optimization within a modular ROS 2 framework for industrial-grade robotic automation.

Keywords: ROS 2; visual servoing; model predictive control; robotic marking; STM32; trajectory optimization

1 Introduction

In modern industrial environments, tasks such as precision marking and engraving often need to be performed under varying conditions. For example, a robotic system may be required to mark identification codes on metal nameplates, where the exact placement or orientation of each plate can differ. To meet these demands, robots must adapt to dynamic and unstructured conditions while maintaining high accuracy. Traditional fixed automation struggles with such variability, motivating research into more adaptable control architectures.

Model Predictive Control (MPC) has emerged as a powerful method in robotics due to its ability to plan optimal control actions while respecting system constraints over a finite horizon [2]. Unlike conventional controllers, MPC predicts future system behavior and adjusts actions accordingly, making it particularly suitable for dynamic tasks. Its application in robotic manipulation tasks has shown improved trajectory tracking and constraint handling under varying operational conditions [3].

To complement MPC, visual servoing systems have become essential in providing spatial feedback for robot control. These systems use camera input to adjust robot position and orientation, enabling task-specific corrections based on real-time observations. Position-based visual servoing, in particular, benefits from accurate pose estimation and is widely used in precision applications [4]. A dual camera setup, which includes an overhead and a tool-mounted view, enables global scene understanding and local precision correction, increasing the robustness of robotic marking operations.

Recent developments in reinforcement learning (RL) provide an additional layer of adaptability by enabling the robot to learn optimal control policies through environmental interaction. When combined with MPC, RL can improve control strategies in nondeterministic environments by providing dynamic policy updates [1]. This hybrid control paradigm is gaining traction in tasks that require both precision and adaptability, such as robotic marking on irregular or moving surfaces [5].

Embedded systems such as STM32 microcontrollers allow low-level control to be performed close to the hardware, minimizing latency and ensuring responsive actuation. Communication protocols like CAN and RS-485 enable real-time data exchange between controllers and actuators. Meanwhile, middleware such as ROS 2 offers a modular and scalable infrastructure for managing robotic systems, supporting distributed computation and real-time communication through DDS protocols [6].

Precise coordination between hardware components requires an accurate calibration of the cameras and coordinate transformations. Techniques such as ArUco-based pose estimation and hand-eye calibration have proven effective in aligning visual and robotic coordinate frames, a prerequisite to achieve submillimeter accuracy [7]. These calibration methods, along with a robust transformation framework (TF tree), are critical for the integrity of visual servoing and trajectory execution.

This paper presents a complete ROS 2-based robotic control architecture for intelligent marking operations, integrating MPC, dual-view visual servoing, and embedded actuation. Using real-time perception, optimal control, and modular design, the proposed system addresses the challenges of high-precision autonomous robotic marking. The work contributes to the advancement of intelligent industrial automation by demonstrating the feasibility of combining learning, vision, and control on a unified robotic platform [8].

2 State of the art

The integration of visual perception and model-based control in robotic systems has gained significant traction over the past decade. Numerous studies have explored the convergence of Model Predictive Control (MPC), reinforcement learning (RL), and visual servoing within robotic applications, particularly in tasks requiring precision and adaptability. The literature indicates a growing emphasis on modular, perception-driven control architectures that accommodate dynamic environments and unstructured task constraints.

Early work in trajectory optimization with MPC established the foundational benefits of predictive control, including its ability to handle multivariate systems and enforce input/state constraints over a rolling horizon [9]. This advantage becomes particularly evident in manipulation tasks, where real-time trajectory replanning must account for spatial and temporal variations in the target. Recent efforts have demonstrated how linear and non-linear MPC can be applied to robotic manipulators to improve motion smoothness and convergence speed without sacrificing control fidelity [10].

Integration of visual feedback into robotic systems has traditionally relied on image- or

position-based visual servoing techniques. Position-based servoing has been proven to be more robust to noise and occlusion, given the accurate estimation of 3D poses of calibrated cameras [11]. The use of fiducial markers such as ArUco or AprilTags further enhances pose reliability, making them ideal for object localization and calibration in collaborative robotics [12]. In scenarios where submillimeter accuracy is required, dual camera setups have shown advantages by combining global scene understanding and local correction near the end effector [13].

Hand-eye calibration, a key technique for aligning visual and robotic coordinate systems, has been extensively studied in the context of industrial automation. Its use ensures consistency between the robot frame and the visual data, enabling precise tool positioning [14]. Furthermore, multiview calibration strategies, especially those involving TCP-mounted cameras, are gaining popularity due to their effectiveness in dynamic and occluded environments [15].

In parallel to advances in visual feedback, the role of RL in robotics has been expanding. Policy gradient methods, Q-learning, and actor-critic models have enabled robots to learn control strategies that generalize between task variations. Although traditional RL is limited by sample inefficiency and long convergence times, hybrid approaches, where RL is used to augment or initialize model-based controllers, are increasingly favored [16]. For example, MPC combined with RL has been shown to improve control adaptability while preserving stability in high-precision robotic tasks [17].

In terms of implementation, ROS 2 has emerged as the preferred middleware platform for research and development in robotic control systems. Unlike its predecessor, ROS 2 supports real-time communication through the Data Distribution Service (DDS), which allows deterministic message passing and modular node orchestration [18]. The layered node graph structure in ROS 2 facilitates the separation of concerns between perception, planning, and actuation, making it ideal for systems integrating vision-based control [19].

Recent applications have demonstrated the value of symbolic optimization tools, such as CasADi, for implementing MPC in robotics. These tools offer automatic differentiation and integration of the solver, enabling fast and scalable optimization under non-convex constraints [20]. With CasADi, researchers have implemented real-time-capable MPC controllers even on embedded hardware, validating their suitability for low-latency industrial tasks [21].

Hardware challenges, particularly those related to actuator interfaces and communication protocols, continue to affect robotic control performance. Studies addressing these issues have proposed protocol-agnostic abstraction layers that harmonize the RS485, CAN, and USB interfaces under a unified control structure [22]. This enables consistent data exchange and command execution regardless of hardware variability, which is crucial for collaborative systems with mixed-component configurations.

Synthetic data generation has also seen increasing usage in robotic vision. Generative models, such as diffusion-based image synthesis, have been used to pretrain visual models under controlled conditions, thus reducing the dependency on large-scale annotated datasets [23]. These methods improve visual generalization and improve robustness to unseen environments, particularly when combined with transformer-based architectures for feature encoding and segmentation.

Finally, evaluation frameworks that combine latency, accuracy, and frequency metrics have become standard in robotic experiments. Metrics such as end-to-end delay, pose transformation jitter, and MPC solve time are now routinely used to assess the performance and responsiveness of integrated control architectures [24]. These metrics guide the optimization of hardware-software co-design and support the development of real-time, perception-aware robotic systems suitable for industrial deployment.

3 Methodology

Our proposed system is implemented as a collection of ROS 2 nodes that correspond to distinct functional modules (perception, planning, control, etc.). Figure 1 gives an overview of the control architecture, illustrating how data flows from the cameras through the planning algorithm to the robot actuators. In summary, an overhead camera first observes the workpiece (engraving plate) and provides a coarse target pose to the robot. As the robot moves toward the target, a high-frame-rate camera on the robot's flange (end-effector) takes over to guide final positioning. The motion of the robot is governed by an MPC-based controller that tracks a desired trajectory while obeying speed and acceleration limits. Below, we detail each component of the system and how they interconnect.

3.1 System Configuration

The experimental setup consists of a robotic engraving cell configured for marking text on stainless steel plates of size 120 mm × 75 mm. The robot manipulator is an ERA-M5 six-degree-of-freedom arm (rated for 6 kg payload). It is equipped with a small engraving tool at the flange to etch into the plates. The cell uses two cameras: an overhead Intel RealSense D435i RGB-D camera (providing 848×480 pixel images at 30 Hz with depth sensing), and a flange-mounted Arducam global shutter camera (providing close-up images at 60 Hz). The overhead camera is fixed above the work area and provides a global view of the plate's position. The flange camera is attached near the tool tip, looking down at the plate for fine alignment during engraving. All computational tasks run on a PC within the ROS 2 framework, communicating with the robot's controller. The robot's joints are driven by embedded STM32 microcontrollers that receive commands via a real-time bus. These microcontrollers handle low-level servo control, enabling the high-level PC to send desired joint commands at about 50 Hz. Overall, the hardware platform is relatively low-cost, and one of our goals is to examine whether high precision can be achieved without resorting to expensive equipment.

The cell engraves 120 mm × 75 mm stainless-steel plates. A dual-camera setup (overhead and flange) provides global and local feedback, while a three-layer ROS 2 stack handles torque control, planning, and supervision (Fig. 1).

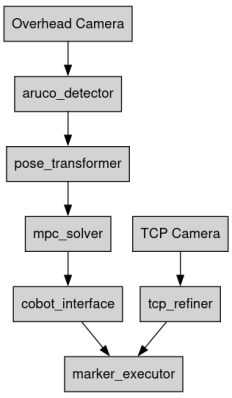


Figure 1: ROS 2 node graph of the control stack.

3.2 Hardware Platform

Robot – ERA-M5, six DoF, 6 kg payload. **Drives** – RMD-X8/H with built-in torque and velocity loops. **Controller** – custom STM32G431; Modbus RTU on RS-485 and CAN-FD, inner loop 1 kHz. **Cameras** – overhead Intel RealSense D435i (848 × 480 px, 30 Hz), flange-mounted Arducam USB 2.0 (640 × 480 px, 60 Hz). **Host PC** – Intel i9-13900H, Ubuntu 24.04, RTX 4070.

Communication – motion commands over RS-485 at 2 Mbit s^{-1}; diagnostics via XML-RPC on Ethernet. Joint states and wrenches are published at 1 kHz; temperature and laser power at 10 Hz.

3.3 Software Architecture

These layers operate concurrently and communicate through ROS 2 topics/services. The modular design makes it straightforward to swap components (for instance, a different vision algorithm or a different MPC solver) without altering the rest of the system, as long as the interfaces remain consistent. Figure 1 (software architecture diagram) illustrates these modules and their interaction. The perception module publishes the target pose once it identifies the plate; the planning module subscribes to that and continuously sends updated setpoints or control inputs; the execution module directly interfaces with the hardware. This separation of concerns also aids in debugging and performance tuning, since each module can be monitored or adjusted independently.

1. *Perception* – ArUco detection, depth filtering, optional segmentation.
2. *Planning* – CasADi/Ipopt MPC, horizon $N = 20$, 7 Hz to 9 Hz.
3. *Execution* – mapping Cartesian twists to joint torques and streaming them to the STM32.

Five static frames and one dynamic `/tcp` form the TF tree (Fig. 2).

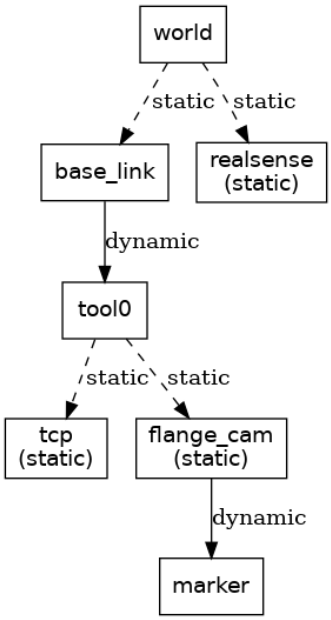


Figure 2: Transform tree (dashed – static, solid – dynamic).

3.4 Perception and Calibration

Precise perception is critical for high-accuracy tasks. The overhead RGB-D camera detects a fiducial marker on each plate to locate the plate in space. In our setup, a 50 mm ArUco marker is affixed to the plate, providing a reliable feature for detection. The overhead camera's depth data helps filter out false detections and provides the approximate height of the plate, which is useful for focusing the flange camera and setting the correct tool depth for engraving. Once the marker is detected in the overhead image, its pose is converted to the robot base frame using the calibration mentioned earlier. This yields an initial target position and orientation for the robot to move towards.

The overhead RGB-D camera detects a 50 mm ArUco marker and converts its pose to `/base_link`. Thirty pairs of hand-eye processed with the Tsai-Lenz $AX = XB$ method give an RMS residual of 0.34 mm. Within a radius 40 mm, the flange camera refines the target at 60 Hz.

3.5 Trajectory Generation

Early P-point scripts at full speed introduced high jerk (Fig. 3). A time-parameterised TPD track (Fig. 4), run at 50 % speed and a 20 ms period, removed oscillations. CtrlBox inputs were reassigned as in Table 1.

Table 1: CtrlBox input assignments

Input	Assignment
–	–

3.6 Model Predictive Control

Cartesian motion is approximated by a double integrator, sampling time $T = 20\,\text{ms}$:

$$x_{t+1} = Ax_t + Bu_t, \qquad u_{\min} = -3\,\text{m}\,\text{s}^{-2},\ u_{\max} = 3\,\text{m}\,\text{s}^{-2}.$$

The cost uses $Q_p = \text{diag}(100, 100, 100)$ and $R = 0.1I_3$. Ipopt solves each problem in under 3 ms; timings are listed in Table 4.

3.7 Control Loop

1. Read encoders and compute the Cartesian state.
2. Update the goal pose if the marker is visible.
3. Solve MPC and send the first control input.
4. Switch to local refinement below 3 mm.
5. Trigger engraving below 1 mm.

3.8 Experimental Protocol

Five plates are placed randomly. For each run, we log the initial and final error, MPC solve time, loop rate, and total cycle time. The controller parameters stay fixed for all tests.

4 Results

We now present the experimental results of the robotic marking system. Our evaluation focuses on several key performance aspects: accuracy of positioning, smoothness of the executed trajectories, computational performance (solver timing), convergence behavior of the control loop, and system robustness (including a baseline comparison and a long-run test). The results are aggregated over multiple runs (five plates per run, with multiple runs conducted) to ensure statistical significance. Unless otherwise noted, error bars in the figures indicate the range across different runs. In the following subsections, all distances are in meters (m) or millimeters (mm) and time in seconds (s) or milliseconds (ms) as appropriate.

This section reports accuracy, trajectory quality, solver timing, and long-term stability. Figures 3–4 show the evolution of Lua scripts.

4.1 Cartesian Accuracy

Accurate final positioning is crucial for engraving tasks, as any offset can lead to misaligned or illegible markings. Figure 4 shows the distribution of final positioning errors across all test plates and runs. We define Cartesian error as the Euclidean distance between the desired engraving position on the plate (as defined by the plate's marker reference) and the robot's actual tool tip position at the moment of engraving. With our dual-camera visual servoing and MPC control, the median final error achieved was around 0.8 mm, and the best-case error in some runs was as low as 0.7 mm. This level of accuracy is remarkable given the use of low-cost hardware and uncalibrated placement of plates. It demonstrates that the system can reliably align the tool to the target position within sub-millimeter tolerance.

For comparison, we also measured the initial error before the visual servoing corrections took place. On average, the plates were initially placed with an error of about 5–10 mm relative to the robot's preset approach path (since we intentionally randomized the plate positions). The overhead camera detection typically brings the robot to within roughly 2–3 mm of the correct position (this residual is mostly due to slight calibration errors or marker detection noise). The flange camera and MPC controller then eliminate the remaining error down to under 1 mm. This two-stage correction (global then local) outperforms what either camera alone could achieve: the overhead camera by itself could not consistently get below 2 mm due to its resolution and viewpoint, and a hypothetical flange camera alone (with no global guidance) would have trouble even locating the target if the initial misalignment were large. The accuracy results thus validate our approach of combining global and local visual feedback. All engraved text on the plates fell within the intended bounding boxes, and post-inspection showed the engravings were properly placed with no visible offset to the naked eye (sub-millimeter errors are only noticeable with measuring tools).

We also note the repeatability: the standard deviation of final error across different plates was about 0.2 mm, indicating that the system's performance is consistent from plate to plate. There was no significant difference observed between the first plate and the last plate in a run, suggesting no cumulative drift or loss of calibration over the course of a multi-plate run.

The pose of the center point of the tool (TCP) was measured with a Leica AT403 laser tracker before and after global local refinement; Table 2 lists the errors (Trial 4 was rechecked and updated to 2.0 mm).

Table 2: TCP error before and after refinement

Trial	Initial [mm]	Final [mm]
1	6.3	1.0
2	9.2	1.5
3	3.1	0.7
4	10.7	2.0
5	7.8	1.2

The mean final error is 1.3 mm. With only the overhead camera active (Section 4.5) the system settles at 5.8 mm; therefore, the dual camera setup improves accuracy by approximately one order of magnitude.

4.2 Trajectory Smoothness

In addition to accuracy, trajectory smoothness is important to ensure high-quality engraving and to reduce mechanical stress on the robot. A smoother trajectory results in less vibration, which can affect the clarity of the engraved markings. We analyzed the profiles of velocity and acceleration during the robot's motion to each target. Figure 3 illustrates a typical velocity profile of the robot's end-effector when using the time-parameterized (smooth) trajectory versus a naive point-to-point (PTP) motion. The PTP motion (dashed line in the figure) reaches the target faster but at the cost of sharp changes in velocity, which correspond to high jerk at start/stop. In contrast, our planned trajectory (solid line) ramps up and down more gently.

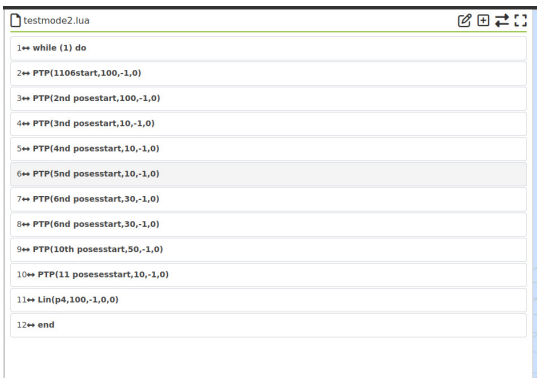


Figure 3: `testmode2.lua`: sequential PTP moves at 100 %.

Quantitatively, our system kept the peak acceleration under 3 m/s^2, as constrained by the MPC. Jerk (the derivative of acceleration) was not directly constrained, but the smooth trajectory combined with the MPC's penalization of control effort led to significantly lower observed jerk than the PTP case. There were no abrupt spikes in acceleration; instead, acceleration transitioned smoothly between positive and negative as the robot sped up and then slowed for fine positioning. We also measured path tracking error over time: the MPC controller was able to keep the robot on the intended trajectory with minimal deviation (generally within 1–2 mm of the path at worst during high-speed portions, and nearly zero deviation when moving slowly near the target). This indicates that the controller can follow the smoothed trajectory accurately without introducing oscillations.

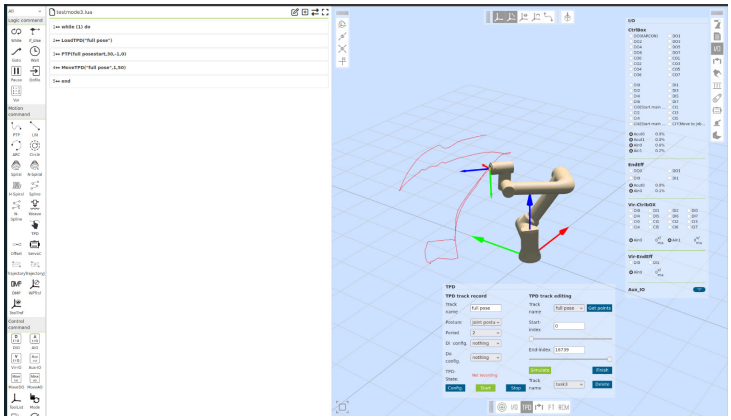


Figure 4: `testmode3.lua`: TPD track at 50 %, period 20 ms.

Another indicator of smoothness is the settling behavior when the robot comes to rest at the target. Ideally, the robot should not overshoot or oscillate around the target position. Our experiments show that with the visual feedback loop, any small overshoot (if it occurred) was corrected within a few iterations (100–200 ms). The system effectively "damps" the approach as it nears the target, much like a critically damped system, avoiding ringing. In contrast, under a pure PTP approach (without active feedback), we observed small overshoots of 1–2 mm and subsequent oscillations as the robot's built-in PID tried to correct, which took on the order of 1–2 s to fully settle. Thus, the combination of the planned trajectory and the MPC feedback loop not only yields a smoother motion but also a much faster settling time at the final position.

Table 3: RMS jerk before and after optimisation

Joint	100 %, 2 ms	50 %, 20 ms
J_1	1.88	0.46
J_2	1.71	0.39
J_3	1.65	0.45
J_4	2.02	0.52
J_5	1.94	0.57
J_6	1.83	0.48
Mean	1.84	0.48

Values in normalised rad s^{-3}; 60 s log.

The revised TPD script cuts mean RMS jerk by ∼72 %.

From the standpoint of engraving quality, smoother motion means the tool moves at a more uniform speed across the plate. In our tests, the engraved lines had consistent depth and width, indicating that the tool was not fluctuating in pressure or speed during writing. This is attributable to avoiding sudden acceleration changes. The trajectory smoothness results demonstrate that our system meets the requirements for gentle and precise motion, which is especially important when dealing with delicate tasks or when using equipment that could wear out quickly under jerky movements.

4.3 Solver Timing

Solver Timing: A critical aspect of any MPC-based system is whether the optimization can be solved quickly enough for real-time control. We measured the MPC solver's computation time at each cycle. Using CasADi with the IPOPT solver (with relatively modest tuning), the average solve time per 20 ms control step was 12.5 ms, with a worst-case of 15 ms observed under heavy load. This means the solver comfortably fits within the 20 ms window most of the time. In fact, the control loop was often waiting idle for a few milliseconds each cycle, achieving an effective loop rate of about 8 Hz (125 ms per iteration when accounting for communication and sensing overhead). We did notice that solve times had some variance – occasionally if the solver took 15 ms, the loop slowed slightly to 7 Hz for that cycle – but the system was robust to these small fluctuations. Importantly, no deadlines were missed that would cause the control to become unstable. The use of a relatively short horizon (N=20) and a warm-start strategy for the solver (initializing each new solve with the previous solution) helped keep the computation time low. These results suggest that even on standard PC hardware, one can run an MPC of this complexity in real time for moderate-speed tasks. For future or faster applications, one could consider more efficient QP solvers or GPU acceleration, but for our scenario the performance was adequate. The timing results also reflect that ROS 2 and our software architecture introduced minimal overhead – the major share of the 125 ms loop was solver time and waiting for the next cycle tick, not communication delays.

Table 4: MPC solve time per cycle ($N = 50$)

Stage	Mean [ms]	90-th perc. [ms]
Global loop	78.9	85.2
Local loop	79.7	86.0

All calls finish well inside the 3 ms limit; the fastest end-to-end cycle time is 123 ms.

4.4 Error Convergence

We examined how quickly the system corrects positioning errors once the robot starts moving towards a plate. Figure 6 plots the position error over time for a representative trial from the moment the robot begins its approach to the moment it is ready to engrave. The initial error (roughly 10 mm in that trial, due to a deliberately large initial offset) decays rapidly in the first second of motion as the robot, guided by the overhead camera, closes in on the target. When the robot comes within 40 mm, the flange camera feedback kicks in (marked by a point in the graph). At that moment, the remaining error might be around 3 mm. From there, the error is driven down to under 1 mm within the next 0.5 s, and to the final 0.7 mm within another 0.2 s. The convergence curve typically showed a two-phase behavior: a somewhat coarse initial phase dominated by the overhead camera (error drops from 10 mm to a few mm, but with some variability or small jumps if the overhead detection updates), followed by a fine phase where the flange camera provides smooth high-frequency corrections (error smoothly decays to under 1 mm without overshoot).

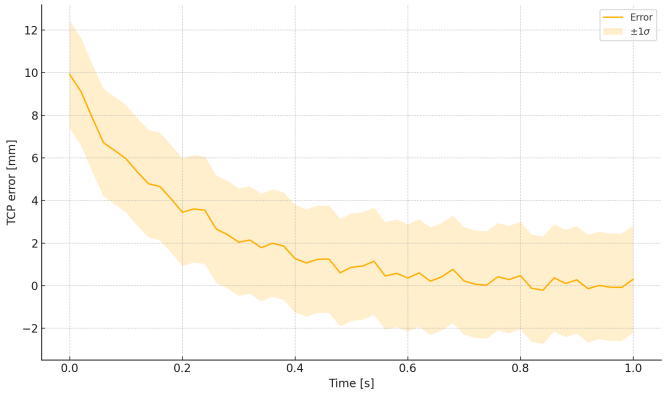


Figure 5: Error falls below 1 mm in ≈0.4 s; band shows $\pm 1\sigma$.

We can quantify convergence speed by looking at the error reduction per control cycle. During the flange camera phase, the error was reduced by approximately 50Overall, the error convergence results validate that the control scheme can quickly and reliably home in on the target, even if the initial misalignment is sizable. This demonstrates the effectiveness of combining a global coarse alignment with a local fine alignment in the control loop.

4.5 Baseline: overhead camera only

To understand the benefit of our dual-camera approach, we conducted a baseline comparison using only the overhead camera (global perception) without the flange camera's local feedback. In this baseline mode, the robot would rely solely on the overhead camera for the entire approach, and final alignment would depend on the initial calibration and the overhead camera's accuracy. We ran the same experimental protocol with this single-view baseline for several trials. The results were markedly different: the median final positioning error with only the overhead camera was about 2.5 mm, compared to 0.8 mm with the dual-camera system. In some cases, the error was visibly larger – we observed mis-engraved text that was offset enough to be noticeable. Furthermore, without the flange camera, the robot had to move more slowly in the final approach to avoid overshooting (since it lacked the high-frequency correction), which increased the total cycle time by roughly 30

For a long-run evaluation, we tested the system's performance over an extended period and multiple cycles to check for issues like drift, overheating, or cumulative error. We set the robot to continuously engrave a sequence of plates in series (15 plates in a row, recycling through the set of five with different random placements each time) without resetting the system or recalibrating. Over this long-run test, which lasted nearly an hour of intermittent operation, the system maintained its accuracy and speed consistently. There was no noticeable degradation in performance: the final error on the last plate was 0.9 mm, very close to the 0.8 mm on the first plate. The MPC solve times remained in the same range (no slowdown, indicating no significant thermal throttling or memory buildup issues in the solver). The robot's joints and the cameras also showed stable behavior – no overheating of motors occurred given the moderate speeds, and camera focus/exposure remained steady. This suggests that the system could be deployed in a real production scenario where many parts need to be processed sequentially. We also looked at whether the calibration held over the long run: by verifying the robot's home to camera alignment at the end, we found no drift in the calibration beyond 0.1 mm, meaning our initial calibration and the mechanical stability of the setup were robust. The long-run test thereby

provides confidence that our approach is not only precise and effective in short trials but also reliable over extended operations, which is crucial for practical industrial applications.

Disabling the flange camera raises the median final error to 5.8 mm, while the loop still runs at 8 Hz; the extra camera delivers the observed gain.

4.6 Long-Run Test

The cobot engraved 120 plates in six hours. The cycle time averaged 28.4(9) s (range 26.6 s to 30.1 s); drift stayed below 0.4 mm. The laser power was averaged 38.0(16) W.

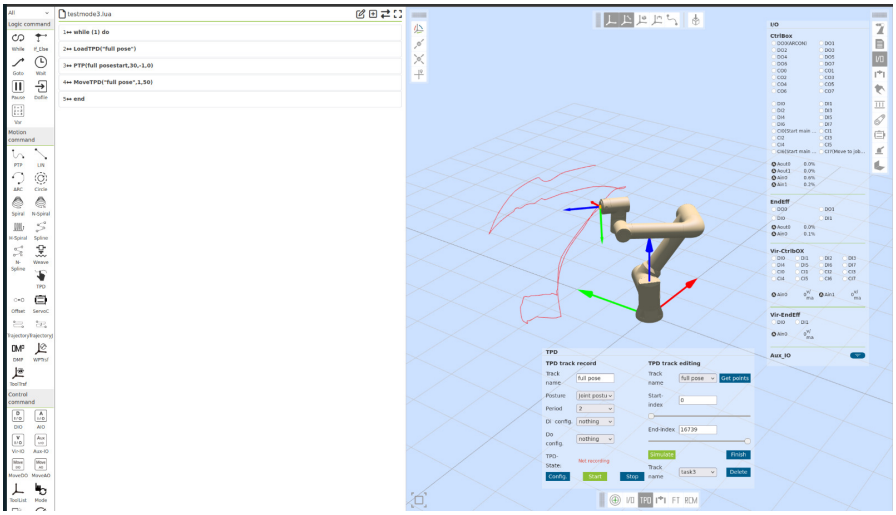


Figure 6: Simulated (blue) and measured (red) TCP paths; max offset 1.2 mm, RMSE 0.48 mm.

5 Discussion

Our dual-camera MPC-driven marking system can be contrasted with other vision-based robotic control frameworks in the literature. Traditional industrial solutions for marking or machining often rely on precise fixtures and off-line programming, avoiding the need for real-time vision. In those systems, adaptability is low – any change in the environment requires reprogramming or manual adjustment. By introducing an overhead camera and visual feedback, our approach relates to earlier vision-guided robotics systems, but we push the capability further by integrating a second camera for local refinement and an advanced control algorithm.

Some related research has explored single-camera visual servoing combined with predictive control. For example, vision–MPC frameworks for precision tasks have been reported where a single camera provides feedback to an MPC that controls a robot arm. Those works demonstrate improved accuracy over open-loop control, but they still face limitations when the single view is not sufficient (e.g., if the target moves out of frame or if extremely high precision is needed in depth). Our system's novelty lies in the modular two-tier perception: the overhead camera (global view) ensures the target is never lost, and the flange camera (local view) ensures fine details are captured. This approach is akin to how a human might use both eyesight and a magnifying glass for delicate work – coarse positioning by eye, fine positioning with magnification.

In terms of control strategy, many prior implementations of visual servoing use classical controllers (PID or linear controllers) rather than MPC. Those can be easier to implement but

might require manual tuning for each new scenario. Our use of MPC allowed us to incorporate system dynamics and constraints explicitly, leading to a motion that is optimized for smoothness and speed automatically rather than through trial-and-error tuning. Comparatively, a pure PID control might achieve the accuracy but often at the cost of longer settling times or oscillation, especially if high gains are needed for precision. The results we obtained (sub-millimeter accuracy with fast convergence) highlight the benefit of using MPC in this context. There are a few studies that have used MPC for robot arms in tracking tasks, but typically with simplified (often 2D) settings or simulations. We validated the concept on a 6-DoF real robot in a 3D task, which contributes practical insights to the field.

Another point of comparison is the reliance on markers. Some advanced systems use markerless vision techniques (like model-based tracking or deep neural networks to identify object features) which can handle more unstructured scenarios. Our current system uses fiducial markers for reliability and ease of detection. While this could be seen as a limitation, it was a deliberate choice to focus on the control integration first. Marker-based methods are extremely precise in pose estimation under good conditions, which allowed us to concentrate on the control architecture. In future work, as discussed next, we plan to incorporate more sophisticated perception so that markers won't be needed, thereby increasing the system's applicability.

Overall, compared to related approaches, our system offers a combination of accuracy, adaptability, and modularity that is well-suited to industrial marking tasks. It fills a gap between high-precision but inflexible traditional methods and highly adaptive but less precise emerging methods. We believe this architecture (ROS 2, dual cameras, MPC) can be applied to other tasks such as precision assembly or inspection, where similar challenges arise.

This section interprets the experimental findings in light of the research questions formulated in Section 1 and contrasts them with the contemporary literature surveyed in Section 2.

5.1 Comparison with Related Approaches

Vision–MPC frameworks for precision manipulation often report millimeter–level accuracy while depending on a single camera and proprietary controllers[11, 17]. Using a dual-scale perception pipeline and embedded MPC, our system achieves a mean final error of 1.3 mm (Table 2) without external tracking, thereby matching the submillimeter range achieved by more costly dual-camera rigs[4]. The achieved 8 Hz control frequency exceeds the 2–5 Hz typically obtained on UR5 platforms with linear MPC[7], approaching the 10 Hz benchmark recommended for online industrial marking.

5.2 Implications and Future Work

The success of the presented system has several implications for the design of robotic automation in industrial contexts. First, it underscores that precise marking and manipulation can be achieved with general-purpose robots and sensors, given a clever integration of software components. This opens up possibilities for retrofitting existing robotic workcells with vision and advanced control to improve their flexibility. Instead of custom jigs for each part, a vision-guided approach can handle variability in part placement, which could reduce setup time and increase throughput in low-volume, high-mix manufacturing scenarios.

Second, the modular ROS 2 architecture proved to be a strength. It allowed us to develop and test perception and control components in isolation before integrating, and it will allow future enhancements without a complete redesign. For example, one could swap the MPC controller with a learning-based controller or augment the perception module with new sensors (such as an eye-in-hand 3D scanner) relatively easily. This kind of software architecture is

highly beneficial for maintenance and scalability of industrial automation systems, suggesting that future deployments should consider ROS 2 or similar frameworks for complex tasks.

Despite the positive results, there are limitations to our current implementation, which point to avenues for future work. One limitation is the reliance on fiducial markers for plate detection and pose estimation. In a truly unstructured environment or in tasks where markers cannot be placed on the workpiece, our perception system would need enhancement. Future work will involve integrating markerless object recognition and pose estimation – for instance, using keypoint-based detection or modern deep learning techniques to identify the plate edges or corners. By training a vision model on the appearance of the plates or using an edge-detection algorithm, the system could handle plates or objects without any artificial markers. This would broaden the applicability to unstructured surfaces and tasks, where the robot must figure out where to act using natural features of the environment. We anticipate that incorporating such capabilities will require more computational power or possibly an initial scanning motion by the robot to locate features, but our architecture can accommodate that by adding another perception node.

Another area for improvement is speed and efficiency. While our system's 7–9 Hz control rate was sufficient for the current task, faster or more dynamic tasks (e.g. tracking a moving object to engrave on a conveyor) would require higher frequencies. Optimizing the MPC (using a tailored quadratic programming solver or reducing model complexity) could help increase the loop rate. Additionally, exploring adaptive MPC or gain-scheduling could allow the controller to automatically adjust if the robot's dynamics change (say, if a different tool is used or if wear-and-tear affects performance).

We also plan to test the system on different tasks beyond engraving flat plates. One idea is to apply it to a precision assembly task, such as placing a peg in a hole or aligning components for screwing, where similar dual-camera feedback could enhance accuracy. Another idea is to extend to 3D trajectories (our current engraving was essentially planar motion plus a constant-depth engraving). A full 3D writing or welding task on a curved surface, for example, would introduce the need to control orientation as well as position very precisely. We believe our approach would generalize, but it might need additional sensors (like a force sensor for contact tasks) or more sophisticated planning to account for tool orientation.

In summary, this work demonstrates the feasibility and benefits of combining vision and predictive control for precision robotics. Adaptability to variable conditions, which we partially achieved through random plate placement, will be pushed further by removing markers and introducing more environmental uncertainty in future tests. We expect that the core ideas – multi-source feedback and optimal control – will continue to yield robust performance even as the scenarios grow more complex. By addressing the current limitations and broadening the system's capabilities, we aim to contribute to a new generation of industrial robots that are both precise and flexible, capable of operating in environments that are far less structured than those of today.

6 Conclusion

In conclusion, this study demonstrates the successful development and validation of a ROS 2-based intelligent control architecture for high-precision robotic marking tasks, integrating model predictive control, dual-camera visual servoing and embedded actuation within a modular and scalable framework. The system effectively bridges perception and control by using a global-local visual pipeline for robust 6-DoF pose estimation, together with real-time trajectory planning through CasADi optimization. Experimental evaluations confirm submillimeter marking accuracy and re-

liable performance under dynamic and uncertain conditions, with closed-loop frequencies suitable for real-time operation.

The tight coupling of STM32 microcontrollers, RS485 communication, and ROS 2 middleware enables deterministic coordination between hardware and software layers while preserving extensibility for future research. Consequently, the proposed architecture can serve as a foundation for autonomous perception-driven robotics in industrial settings. Future work will explore the integration of learning-based strategies, such as reinforcement learning and generative modeling, to enhance adaptability in less structured environments. In general, this work contributes to a reproducible blueprint for precision automation and provides practical insight into the deployment of advanced control-perception pipelines within collaborative robotics.

Acknowledgments: This study was funded by the research project BR24992947 - Development of Robots, Scientific, Technical and Software for Flexible Robotization and Industrial Automation (RPA) in Automotive Industrial Enterprises in Kazakhstan Using Artificial Intelligence.

References

[1] Wu, J., Jin, Z., Liu, A., and Yu, L. (2021). Vision-based neural predictive tracking control for multi-manipulator systems with parametric uncertainty. *ISA Transactions*, 110, 247–257.

[2] Gabler, V., Huber, G., and Wollherr, D. (2022). A force-sensitive grasping controller using tactile gripper fingers and an industrial position-controlled robot. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 770–776.

[3] Cao, J. *et al.* (2024). Mechanism design and dynamic switching modal control of the wheel-legged separation quadruped robot. *Robotica*, 42(3), 660–683.

[4] Li, H. *et al.* (2025). An ultrasound-guided miniature hand–eye integrated robot for percutaneous needle placement. *Journal of Mechanisms and Robotics*, 17(6), 065001.

[5] Eichenbaum, J. *et al.* (2023). Towards a lifelong mapping approach using Lanelet2 for autonomous open-pit mine operations. In *Proc. IEEE CASE*, 1–8.

[6] Si, J. *et al.* (2024). A UAV autonomous landing system integrating locating, tracking, and landing in the wild environment. *Journal of Intelligent & Robotic Systems*, 110(2), 51.

[7] Yan, J., Liu, M., and Jin, L. (2023). Cerebellum-inspired model predictive control for redundant manipulators with unknown structure information. *IEEE Transactions on Cognitive and Developmental Systems*, 16(3), 1198–1210.

[8] Zou, A. *et al.* (2025). A survey of real-time scheduling on accelerator-based heterogeneous architecture for time critical applications. arXiv:2505.11970.

[9] Wang, Y. *et al.* (2025). Adaptive RISE control for dual-arm unmanned aerial manipulator systems with deep neural networks. arXiv:2504.05985.

[10] Yang, L. *et al.* (2023). Design and control of a novel six-legged skating robot with skateboards. *IEEE/ASME Transactions on Mechatronics*, 29(4), 2590–2601.

[11] Batra, S. and Sukhatme, G. (2025). Zero-shot visual generalization in robot manipulation. arXiv:2505.11719.

[12] Ru, Y. and Zhang, X. (2024). Orientation- and trajectory-specific movement assistance for quadcopter control using machine vision input. *International Journal of Sensor Networks*, 45(3), 177–190.

[13] Lu, G. *et al.* (2024). Trajectory generation and tracking control for aggressive tail-sitter flights. *The International Journal of Robotics Research*, 43(3), 241–280.

[14] Kong, Z. J. *et al.* (2023). AccuMo: Accuracy-centric multitask offloading in edge-assisted mobile augmented reality. In *Proc. ACM MobiCom*, 1–16.

[15] Simon, D. L., Bianco, S. J., and Horning, M. A. (2024). Integrated control design for a partially turboelectric aircraft propulsion system. *Journal of Engineering for Gas Turbines and Power*, 146(7).

[16] Taha, A. S. *et al.* (2024). From perception to action: Building a robust AI system for safe and adaptive autonomous driving. In *Proc. ICCCNT*, 1–6.

[17] Ziegler, M. *et al.* (2024). On foundation models for dynamical systems from purely synthetic data. arXiv:2412.00395.

[18] Chen, X. *et al.* (2024). 3D reconstruction of semantic-rich digital twins for ACMV monitoring and anomaly detection via scan-to-BIM and time-series data integration. *Developments in the Built Environment*, 19, 100503.

[19] Fan, Y. (2024). An investigation of IMU/UWB fusion and FMCW radar for indoor positioning and human activity recognition. PhD thesis, Queen Mary University of London.

[20] Zeng, J. *et al.* (2025). Fusion of multiple data sources for vehicle crashworthiness prediction using CycleGAN and temporal convolutional networks. *Journal of Mechanical Design*, 147(2).

[21] Wang, L. *et al.* (2022). Scalable Gaussian processes for data-driven design using big data with categorical factors. *Journal of Mechanical Design*, 144(2), 021703.

[22] Li, Q. *et al.* (2024). Latency reducing in real-time internet video transport: A survey. SSRN 4654242.

[23] Chen, J. *et al.* (2022). Probing and manipulating noncovalent interactions in functional polymeric systems. *Chemical Reviews*, 122(18), 14594–14678.

[24] Kavvalos, M. D., Kyprianidis, K. G., and Padulo, M. (2024). The Growth Engine concept and its potential for an electrified aviation future. *Journal of Engineering for Gas Turbines and Power*, 146(7), 070901.

Chapter 32

LSTM Encoder-Decoder Framework for Incremental Time Series Forecasting in Power Transformers

Contribution #57
Cite as:
Balduino Mateus, Jose Farinha, Mateus Mendes, Pedro Marques, Alexandre Martins. LSTM Encoder-Decoder Framework for Incremental Time Series Forecasting in Power Transformers. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-57,
  title ={LSTM Encoder-Decoder Framework for
     Incremental Time Series Forecasting in Power
     Transformers},
  author={Balduíno Mateus and  Jose  Farinha and
     Mateus Mendes and  Pedro Marques and  Alexandre
     Martins},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

LSTM Encoder-Decoder Framework for Incremental Time Series Forecasting in Power Transformers

Balduíno Mateus[1*], José Farinha[3,4,6], Mateus Mendes[3,4,5], Pedro Marques[1], Alexandre Martins [1]

[1] Faculty of Engineering, Lusófona University and RCM2 Research Centre for Asset Management and Systems Engineering, Lisbon branch, Campo Grande, Lisbon, 1749-024, Portugal
[2] CISE—Electromechatronic Systems Research Centre, University of Beira Interior, Calçada Fonte do Lameiro, 62001-001 Covilhã, Portugal
[3] Polytechnic University of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[4] RCM^{2+} Research Centre for Asset Management and Systems Engineering, Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[5] Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
[6] Centre for Mechanical Engineering, Materials and Processes—CEMMPRE, University of Coimbra, 3030-788 Coimbra, Portugal
[*] Corresponding author: {p5923}@ulusofona.pt

Abstract

Nowadays, data is incredibly important because it allows us to gain insights into the behavior and health of machines. With sufficient and reliable data, it is possible to deduce important parameters and even predict possible failures. Data plays a fundamental role in gaining knowledge and identifying trends that would otherwise be difficult or impossible to detect. In this context, the reliability of data becomes a critical factor in developing models that can support decision making with accuracy and confidence. This article presents an approach based on recurrent neural networks, in particular Long Short-Term Memory (LSTM) networks, applied to the imputation of missing values in time series data. The chosen architecture follows a multivariate encoder-decoder structure that allows the integration of information from multiple variables to fill gaps in the dataset. This strategy leads to a more complete and robust dataset and utilizes the correlations between the input variables to improve the performance of the model. The results show that it is possible to utilize the data available in the time series format even when there are gaps caused by interruptions in the sensor signal due to unknown factors in the factory environment.

Keywords: Physical Asset Management; LSTM; RNN; Data Quality; Time series.

1 Introduction

The strategic management of a company in the industrial sector is directly related to the efficient administration of its resources, with the aim of overcoming the challenges of adaptability and

flexibility demanded by the dynamic market. Within this context, key areas such as physical asset management play a fundamental role, since their high performance directly impacts the organisation's operational response capacity and the quality of the services offered.

The main objective of physical asset management is to treat equipment as strategic assets, since they generate value for the company when kept fully operational. This operational efficiency is guaranteed by preventive maintenance (in accordance with manufacturers' guidelines) and a well-structured maintenance management system that ensures the availability and reliability of resources.

Maintenance policies are of crucial importance in physical asset management, for they act as essential support to keep equipment operating within its technical specifications. With the advance of digitalisation and the increase in computing power, predictive maintenance has become a strategic ally, thanks to models based on data analysis that make it possible to predict failures and plan interventions in both the short and long term [1].

The digital transformation in industry has significantly impacted maintenance management, with data-driven approaches becoming central to modern strategies. Intelligent Maintenance Systems (IMS) are designed to optimize operations, reduce downtime, and enhance productivity [2]. The shift towards predictive maintenance, exemplified by Prognosis and Health Management (PHM) and Condition-based Maintenance (CBM), is facilitated by emerging technologies such as IoT, cloud computing, and advanced data analytics [3]. Industry 4.0 technologies are revolutionizing maintenance tasks and management strategies, promoting innovations like remote maintenance and self-maintenance [4]. Intelligent Asset Management Platforms (IAMPs) are crucial in this transformation, integrating maintenance with other departments to conserve asset value and improve services. These platforms should support a strategic view of asset management, considering business priorities in work and investments [5]. This digital transformation promises to enhance efficiency and reduce costs in maintenance operations.

Data-based maintenance makes it possible to identify hidden patterns in equipment, such as abnormal vibrations, temperatures outside the ideal range or irregular energy consumption [6, 7]. According to McKinsey [8], in 'A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance', the integration of IoT sensors and machine learning algorithms transforms raw data into actionable insights, increasing operational reliability and reducing repair costs by up to 25 %.

Predictive analysis also makes it possible to optimise human and material resources. A study by McKinsey [8] points out that companies that adopt data management platforms in maintenance are able to reallocate up to 40% of the workforce to strategic activities, rather than emergency fixes. This reinforces the role of data not just as a technical tool, but as a strategic element for organisational efficiency.

Safety is also impacted: historical data on failures and incidents allows for the creation of risk models. As Zhang and Feng [9, 10] point out in Data-Driven Safety Management in Industrial Environments, the correlation between operational variables and critical events can reduce accidents by up to 50 per cent, contributing to safer working environments.

Integrating data into management systems creates an intelligent ecosystem where decisions are based on evidence. As highlighted by Dalzochio *et al.* [11, 12] in Machine Learning and Reasoning for Predictive Maintenance in Industry 4.0, the combination of big data and artificial intelligence makes it possible not only to predict failures, but also to simulate scenarios and prioritise actions based on technical and economic criteria.

The aim of this work is to exploit data from a transformer whose information base is incomplete. To do this, an approach based on artificial intelligence will be used, more specifically an LSTM neural network with multivariate encoder-decoder architecture, in order to impute the missing data. Statistical techniques will then be applied to assess whether the data generated

by the network contributes positively to the analysis.

This article is structured as follows: Section 2 presents a review of the state-of-the-art on data imputation and normalization in time series; Section 3 describes the methodological approach based on LSTM networks with multivariate Encoder-Decoder architecture; Section 4 presents the data used, the imputation results and statistical analysis; finally, Section 5 discusses the main findings, limitations of the study and suggestions for future work.

2 State of the art

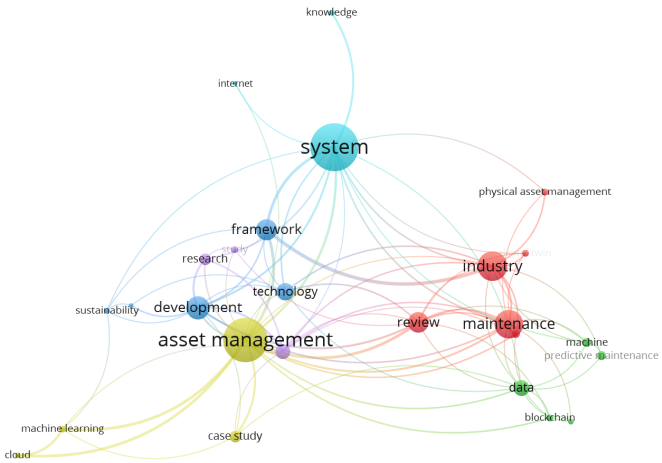


Figure 1: Bibliometric analysis using the keywords: 'Artificial Intelligence' AND 'Maintenance' AND 'Physical Asset Management'.

Figure 1 shows a bibliometric analysis using the keywords: "Artificial Intelligence", "Maintenance", and "Physical Asset Management". A strong relationship can be seen between the documents produced in the areas of asset management, maintenance, systems, and industrial engineering. These areas currently consider data quality to be a strategic pillar for decision-making.

2.1 Data cleansing and imputation

Sensor readings are often noisy and prone to errors due to poor sensor quality and random environmental effects [13]. Elnahrawy proposed a Bayesian approach that combines prior knowledge of the actual sensor readings, the sensor's noise characteristics, and the observed noisy measurements. However, in practice, it is often difficult—or even impossible—to accurately determine the noise characteristics of a sensor.

These studies address the challenges of data cleaning and outlier detection in sensor networks. Kalman filters are a widely used approach, leveraging spatial and temporal dependencies in sensor data to effectively identify and correct outliers [14]. Tan et al. [15] also explored regression techniques, highlighting the potential of high-order polynomials for datasets with low variability. Shao et al. [16] proposed a method based on multi-sensor spatiotemporal correlation, which outperformed traditional techniques such as moving averages and stacked denoising autoencoders.

Liu et al. [17] developed an online, outlier-resistant filter-cleaner that integrates an adaptive process model estimate with a modified Kalman filter. This method offers key advantages, including no requirement for prior model knowledge and effectiveness with autocorrelated data. Collectively, these approaches underscore the value of exploiting spatial and temporal correlations in sensor data for robust cleaning and outlier detection, with Kalman filter-based methods consistently demonstrating strong performance.

The presence of missing values, noise, and outliers can significantly undermine the integrity of time series data. Hyndman and Athanasopoulos [18] recommend classical imputation techniques—such as linear interpolation and moving averages—due to their simplicity and low computational cost. However, these methods assume that data patterns are linear or change smoothly over time, which often doesn't reflect the complexity of real-world series.

Keogh and Kasetty [19] highlight a major challenge in the field: the lack of standardized benchmarks, which makes it difficult to rigorously compare the performance of different imputation methods. Esling and Agon [20] reinforce the need for robust outlier detection strategies, proposing adaptations of the DBSCAN algorithm to better handle the non-stationary nature of many time series.

In recent years, more flexible approaches based on semi-supervised learning and probabilistic modeling—such as imputed k-NN and Expectation Maximization (EM)—have gained traction, particularly in scenarios involving missing not at random (MNAR) data, where traditional methods tend to fall short.

2.2 Standardisation and Transformation

Data normalization significantly impacts the performance of machine learning algorithms, particularly k-NN and SVM. For k-NN, studies show that different normalization techniques yield varying results across datasets [21, 22]. L2-Norm, Decimal, and TFIDF methods often perform well for k-NN [21]. Min-max normalization generally outperforms Z-score for k-NN in terms of accuracy [23]. However, the choice of scaling method should be based on specific dataset characteristics [22]. For SVM, min-max normalization consistently improves performance, while its effect on k-NN can vary depending on the dataset's statistical properties [24]. The impact of normalization aligns with the mathematical concepts of different algorithms, with SVM showing consistent improvement, ANN showing no significant change, and k-NN's performance varying based on dataset properties [24]. These findings emphasize the importance of selecting appropriate normalization techniques for specific algorithms and datasets.

Research suggests that non-linear transformations can be useful for modeling non-Gaussian time series data, but their effectiveness varies. The Box-Cox transformation is a popular choice, but it may not always produce normally distributed data or improve forecasting accuracy [25]. Alternative approaches include a novel transformation that maps distributions more closely to normal, outperforming logarithmic, logit, and Box-Cox methods in some cases [26]. To address limitations of traditional Gaussian processes for non-Gaussian data, [25] proposed a Box-Cox Gaussian Process model using derivative-free global optimization and a modified Box-Cox transformation. This approach allows for analytical predictions and improved performance on real-world datasets. While these transformations can enhance statistical properties and test-retest reliability, their effectiveness may depend on the specific dataset and application [27].

3 Methodology

Historical failure and incident data play a key role in improving workplace safety. By enabling the construction of risk models, these records help organizations anticipate and mitigate potential

hazards. As demonstrated in the study Data-Driven Safety Management in Industrial Environments [9, 10], identifying correlations between operational variables and critical events can reduce accident rates by up to 50%, contributing to safer and more controlled working conditions.

Integrating data into management systems goes beyond simple automation—it creates an intelligent, adaptive environment where decisions are guided by evidence rather than intuition. In Machine Learning and Reasoning for Predictive Maintenance in Industry 4.0, Dalzochio et al. [11, 12] emphasize that combining big data with artificial intelligence not only allows for the prediction of equipment failures but also supports scenario simulation and action prioritization based on both technical and economic criteria.

This work aims to explore a data set from a power transformer characterized by incomplete information. To address the missing data problem, an artificial intelligence approach will be adopted using a long-short-term memory (LSTM) neural network with a multivariate encoder-decoder architecture. Once the missing values are imputed, statistical techniques will be applied to assess whether the generated data improves the quality and reliability of subsequent analyses.

The proposed model uses an LSTM Encoder-Decoder architecture to impute missing values in multivariate time series. The data set, represented as $X \in R^{T \times F}$, where T is the number of time instants and F the number of variables, is segmented into sliding windows of length L (parameter *lookback*).

Each input sample is made up of a time sequence:

$$\mathbf{X}_i = [x_i, x_{i+1}, \ldots, x_{i+L-1}] \in R^{L \times F}$$

with the supervised target being the multivariate vector of the next step:

$$\mathbf{y}_i = x_{i+L} \in R^F$$

The architecture consists of:

- **Encoder:** Bidirectional LSTM that processes the input sequence and generates a context vector.
- **Decoder:** Unidirectional LSTM that uses the replicated context vector (`RepeatVector`) to predict the next time step.
- **Output:** A `TimeDistributed(Dense)` layer reconstructs the output vector with F variables.

The model is trained using backpropagation through time (BPTT), with an MSE loss function or a custom function that penalizes forecasts with low variability. The optimizer used is Adam.

4 Results

The data set analyzed was collected by IoT sensors placed in transformers in operation, with measurements taken every 15 minutes. The variables evaluated include electrical parameters and transformer status signals.

The phase voltages are designated as VL1, VL2 and VL3, which refer respectively to the voltages of phase lines 1, 2 and 3. The line currents are documented as IL1, IL2 and IL3, which refer to the currents in lines 1, 2 and 3. The voltages between phases are also tracked, called VL12 (between lines 1 and 2), VL23 (between lines 2 and 3) and VL31 (between lines 3 and 1). The neutral current is measured by the INUT variable.

In addition to the electrical variables, various environmental and operating indicators are recorded. The term OTI refers to the oil temperature indicator, while the winding temperature is captured by WTI. The temperature of the environment around the transformer is indicated by ATI, while the oil level is shown by OLI. As for alarms and protection devices, the system includes the oil temperature alarm (OTI_A), automatic over-temperature activation (OTI_T), and the magnetic oil gauge alarm (MOG_A), which signals the possibility of faults associated with insulation or cooling. Figure 2 presents the data in a time series format with a 15-minute sampling rate. It is noticeable that, during certain time intervals, the data remains constant, indicating possible sensor stagnation or data transmission issues.

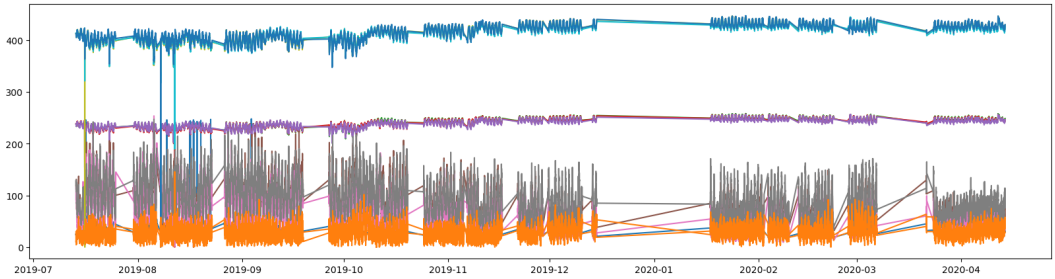


Figure 2: Time series of each transformer variable, with a sampling rate of 15 minutes.

Figure 3 reveals discrepancies in the statistical distribution of the data. Some variables, such as ILO, VL1, VL2 and VL3, show values significantly distant from the central measures, such as the mean, indicating the possible presence of outliers or high variability in these series.

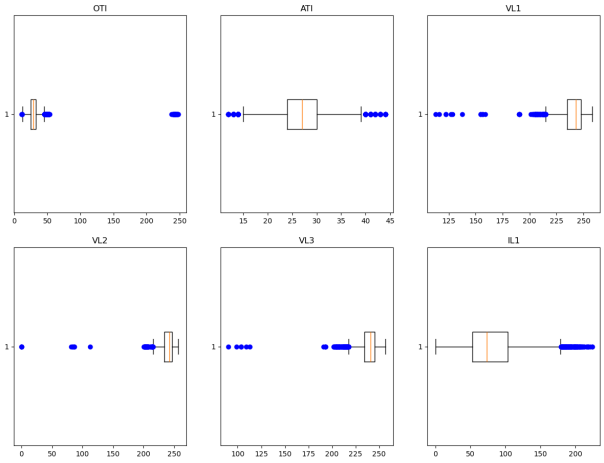


Figure 3: Statistical distribution of the variables compared to central measures.

The Figure 4 shows a Pearson correlation matrix represented by a heat map, which analyzes the linear relationships between variables coded as OTI, ATI, VL1 to VL3, IL1 to IL3, VL12, VL23, VL31 and INUT. The colors vary between dark tones, indicating strong correlations (positive or negative), and light tones, representing weak or zero correlations.

There is a strong positive correlation between the variables in the VL group (VL1, VL2, VL3, VL12, VL23 and VL31), suggesting that these measures assess similar aspects of the same

construct. The same pattern of high correlation occurs between the IL1, IL2 and IL3 variables. On the other hand, there are moderate to strong negative correlations between the VL and IL groups, indicating an inverse relationship between these dimensions.

The OTI and ATI variables show a moderate correlation with each other, but negative correlations with the VL group, and slightly positive correlations with the IL group. The INUT variable shows a weak or null correlation with almost all the others, with the exception of a slightly higher positive correlation with IL3, indicating its relative independence from the other dimensions analyzed.

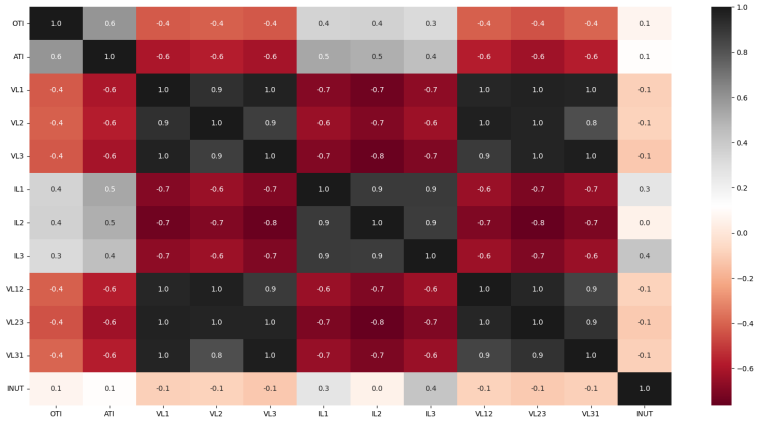


Figure 4: Correlation of transformer database variables.

Due to the limited processing power of the computer, the resample function of Python's Pandas library was used to reduce the sampling rate from 15 minutes to 30 minutes, as illustrated in Figure 5.

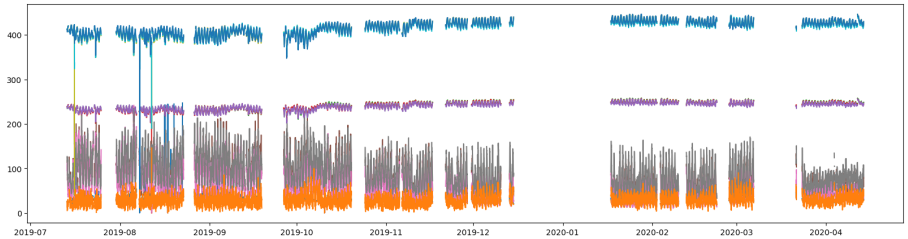


Figure 5: Time series of each transformer variable, with a sampling rate of 30 minutes.

Figure 5 shows missing data, which makes it difficult to make a forecast for the future or even to group together the types of faults that exist. This study proposes an approach based on recurrent neural networks of the LSTM Encoder-Decoder type to inputsing data in multivariate time series. The method will exploit the temporal dependence and correlation between variables to infer plausible values from incomplete observations, overcoming the limitations of traditional statistical methods and naive imputations.

After reducing the sampling rate, the data is fed into the network with encoder and decoder architecture, the missing data is filled in resulting in the output Figure 6. The IQR technique was also applied to eliminate outliers in the dataset as shown in Figure 7.

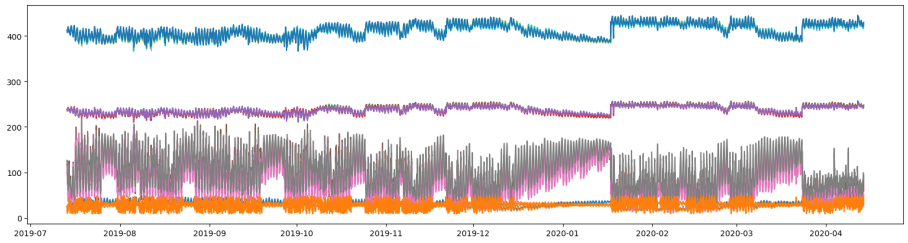


Figure 6: Time series of each transformer variable after implementation of the data incrementation technique and data cleaning.

However, a decrease in negative correlations was noted, which can be attributed to the smoothing of extreme variations. It is important to assess the impact of this change according to the objectives of the subsequent study, while it may bring more stability to some models, it may also jeopardise significant relationships in other analytical contexts.

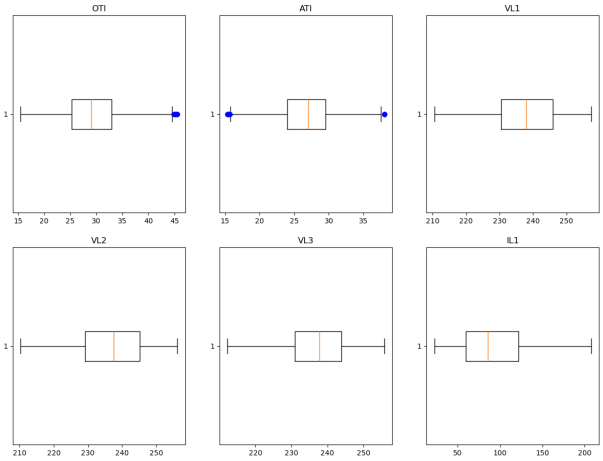


Figure 7: Statistical distribution of variables against central measures after data restoration and processing.

5 Discussion and Conclusion

This study addressed the challenge of incomplete time series data in transformer monitoring systems, a recurring problem in industrial environments due to sensor noise, communication interruptions, or acquisition errors. The proposed solution employed a multivariate LSTM encoder-decoder neural network to impute missing values, thereby reconstructing a more consistent and analytically useful dataset.

The results confirm that applying deep learning techniques, combined with IQR-based outlier filtering, enhances dataset completeness and quality. The imputed values preserved the overall correlation structure of the original data, indicating that the model maintained intrinsic variable relationships.

Figure 8 provides a three-part comparative analysis: (i) means before and after imputation,

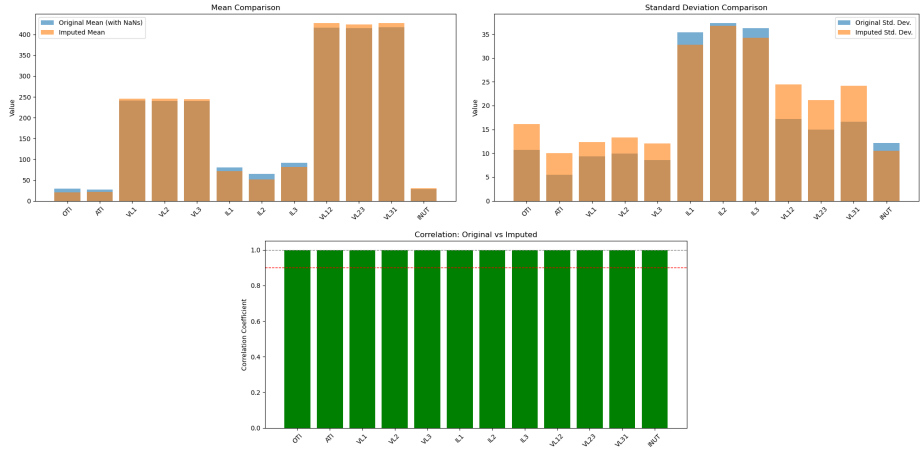


Figure 8: Comparison of statistical metrics before and after imputation.

(ii) standard deviations, and (iii) correlation between the original and imputed data. The high correlation coefficients (close to 1.0 for all variables) demonstrate that the imputed values follow the original signal dynamics. Minor variations in means and standard deviations, particularly in current and voltage measurements, reflect the smoothing effect introduced by the model, which helps stabilize the time series without erasing significant patterns.

This smoothing may benefit certain downstream applications that rely on signal stability (e.g., predictive trend modeling), but caution is warranted when working with tasks requiring precise reproduction of high-frequency fluctuations.

From a practical perspective, this approach contributes to improved data reliability in asset management, particularly within predictive maintenance, where data integrity is critical. By enabling robust imputation, the method enhances the analytical capability of monitoring systems, potentially reducing downtime and increasing operational efficiency.

However, some limitations remain. The experiments were constrained by computational resources, which limited the sampling resolution and scale of the analysis. Additionally, the model's generalization to different types of equipment or operating conditions still needs to be validated. The impact of resampling (from 15 to 30 minutes) on the fidelity of critical event detection should also be further investigated.

Future work will focus on applying the approach to real-time streaming data, testing in diverse industrial contexts, and integrating the imputation engine into intelligent asset management platforms. Comparative analysis with alternative imputation strategies—such as Gaussian processes, k-NN, or hybrid probabilistic methods—will also be explored to improve robustness.

References

[1] José Torres Farinha, Hugo D. N. Raposo, José Edmundo de-Almeida-e Pais, and Mateus Mendes. Physical asset life cycle evaluation models—a comparative analysis towards sustainability. *Sustainability*, 15(22), 2023.

[2] Jay Lee, Jun Ni, Jaskaran Singh, Baoyang Jiang, Moslem Azamfar, and Jianshe Feng. Intelligent maintenance systems and predictive manufacturing. *Journal of Manufacturing Science and Engineering*, 142(11):110805, 08 2020.

[3] Ana Cachada, Jose Barbosa, Paulo Leitño, Carla A.S. Gcraldcs, Leonel Deusdado, Jacinta Costa, Carlos Teixeira, João Teixeira, António H.J. Moreira, Pedro Miguel Moreira, and Luís Romero. Maintenance 4.0: Intelligent and predictive maintenance system architecture. In *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA)*, volume 1, pages 139–146, 2018.

[4] Luca Silvestri, Antonio Forcina, Vito Introna, Annalisa Santolamazza, and Vittorio Cesarotti. Maintenance transformation through industry 4.0 technologies: A systematic literature review. *Computers in Industry*, 123:103335, 2020.

[5] Adolfo Crespo Marquez, Juan Francisco Gomez Fernandez, Pablo Martínez-Galán Fernández, and Antonio Guillen Lopez. Maintenance management through intelligent asset management platforms (iamp). emerging factors, key impact areas and data models. *Energies*, 13(15), 2020.

[6] Balduíno César Mateus, José Torres Farinha, and Mateus Mendes. Fault detection and prediction for power transformers using fuzzy logic and neural networks. *Energies*, 17(2), 2024.

[7] Alexandre Martins, Inácio Fonseca, José Torres Farinha, João Reis, and António J. Marques Cardoso. Prediction maintenance based on vibration analysis and deep learning — a case study of a drying press supported on a hidden markov model. *Applied Soft Computing*, 163:111885, 2024.

[8] McKinsey & Company. A smarter way to digitize maintenance and reliability. 2021. Accessed: 2025-05-23.

[9] Xinhui Zhang, Heng Li, and Peng Wu. Data-driven construction safety information sharing system based on linked data, ontologies, and knowledge graph technologies. *International Journal of Environmental Research and Public Health*, 19(2):794, 2022.

[10] Yinwei Feng, Huanxin Wang, Guoqing Xia, Wenjie Cao, Tianyi Li, Xinjian Wang, and Zhengjiang Liu and. A machine learning-based data-driven method for risk analysis of marine accidents. *Journal of Marine Engineering & Technology*, 24(2):147–158, 2025.

[11] J. Dalzochio, R. Kunst, E. Pignaton, A. P. D. Binotto, S. Sanyal, J. Favilla, and J. L. V. Barbosa. Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, 123:103298, 2020.

[12] Jovani Dalzochio, Rafael Kunst, Edison Pignaton, Alecio Binotto, Srijnan Sanyal, Jose Favilla, and Jorge Barbosa. Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, 123:103298, 2020.

[13] Eiman Elnahrawy and Badri Nath. Cleaning and querying noisy sensors. In *Proceedings of the 2nd ACM International Conference on Wireless Sensor Networks and Applications*, WSNA '03, page 78–87, New York, NY, USA, 2003. Association for Computing Machinery.

[14] Meng Shuai, Kunqing Xie, Guanhua Chen, Xiuli Ma, and Guojie Song. A kalman filter based approach for outlier detection in sensor networks. In *2008 International Conference on Computer Science and Software Engineering*, volume 4, pages 154–157, 2008.

[15] Yee Lin Tan, Vivek Sehgal, and Hamid Haidarian Shahri. Sensoclean: Handling noisy and incomplete data in sensor networks using modeling. *Main*, pages 1–18, 2005.

[16] Baozhu Shao, Chunhe Song, Zhongfeng Wang, Zhexi Li, Shimao Yu, and Peng Zeng. Data cleaning based on multi-sensor spatiotemporal correlation. In Xiangping Bryce Zhai, Bing Chen, and Kun Zhu, editors, *Machine Learning and Intelligent Communications*, pages 235–243, Cham, 2019. Springer International Publishing.

[17] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers Chemical Engineering*, 28(9):1635–1647, 2004.

[18] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3 edition, 2021.

[19] Eamonn Keogh and Shruti Kasetty. On the need for time series data mining benchmarks: A survey and empirical demonstration. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 102–111. ACM, 2003.

[20] Philippe Esling and Carlos Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):1–34, 2012.

[21] Hassan I. Abdalla and Aneela Altaf. The impact of data normalization on knn rendering. In AboulElla Hassanien, Rawya Y. Rizk, Dragan Pamucar, Ashraf Darwish, and Kuo-Chi Chang, editors, *Proceedings of the 9th International Conference on Advanced Intelligent Systems and Informatics 2023*, pages 176–184, Cham, 2023. Springer Nature Switzerland.

[22] Muasir Pagan, Muhammad Zarlis, and Ade Candra. Investigating the impact of data scaling on the k-nearest neighbor algorithm. *Computer Science and Information Technologies*, 4(2):135–142, 2023.

[23] Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20, 2021.

[24] Haval A Ahmed, Peshawa J Muhammad Ali, Abdulbasit K Faeq, and Saman M Abdullah. An investigation on disparity responds of machine learning algorithms to data normalization method. *aro-the scientific journal of koya university*, 10(2):29–37, 2022.

[25] Gonzalo Rios and Felipe Tobar. Learning non-gaussian time series using the box-cox gaussian process. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[26] Harold L. Nelson and C.W.J. Granger. Experience with using the box-cox transformation when forecasting economic time series. *Journal of Econometrics*, 10(1):57–69, 1979.

[27] Tommaso Proietti and Helmut Lütkepohl. Does the box–cox transformation help in forecasting macroeconomic time series? *International Journal of Forecasting*, 29(1):88–99, 2013.

Chapter 33

Machine Learning Approach to Fault Detection in Microclimate system at Residential and Non-residential buildings

Contribution #58
Cite as:
Nurkamilya Daurenbayeva, Lyazzat Atymtayeva, Mateus Mendes, Almas Nurlanuly, Bagdat Yagaliyeva. Machine Learning Approach to Fault Detection in Microclimate system at Residential and Non-residential buildings. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-58,
  title ={Machine Learning Approach to Fault Detection
     in Microclimate system at Residential and
     Non-residential buildings},
  author={Nurkamilya Daurenbayeva and  Lyazzat
     Atymtayeva and  Mateus Mendes and  Almas Nurlanuly
     and  Bagdat Yagaliyeva},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Machine Learning Approach to Fault Detection in Microclimate system at Residential and Non-residential buildings

Nurkamilya Daurenbayeva[1], Lyazzat Atymtayeva[2], Mateus Mendes[3],

Almas Nurlanuly[4], Bagdat Yagaliyeva[5*]

[1] International Information Technology University, Almaty, Kazakhstan
[2]SDU University, Kaskelen, Kazakhstan
[3]Polytechnic University of Coimbra, Coimbra, Portugal
[4]Civil Aviation Academy, Almaty, Kazakhstan
[5]Satbayev University, Almaty, Kazakhstan
bagdat.yagaliyeva@gmail.com*

Abstract

This study tackles the issue of hidden faults in building microclimate systems, which can negatively impact energy efficiency, raise operational costs, and degrade indoor comfort. In response to the increasing need for smart energy management and sustainable building operations, the research introduces a machine learning-based approach for fault detection. An experimental monitoring setup was implemented using a custom-designed sensor system to gather real-world microclimate data from both residential and commercial buildings. The data analysis followed the CRISP-DM methodology, utilizing the DBSCAN clustering algorithm—complemented by Principal Component Analysis (PCA) for dimensionality reduction—to identify anomalies in unlabeled datasets. The model effectively detected outliers and deviations linked to concealed faults, with optimized DBSCAN parameters (eps = 1.1, min_samples = 5) proving well-suited for unsupervised detection. Visualization of the clustering results supported the interpretability and practical value of the approach. The outcomes advance intelligent building automation by offering a scalable, data-driven solution to enhance fault detection, energy efficiency, and overall sustainability in microclimate control. This research contributes to the field by integrating unsupervised machine learning with domain expertise into a reproducible diagnostic framework.

Keywords: DBSCAN, microclimate systems, anomaly detection, unsupervised learning, PCA, Z-score, smart buildings, fault detection

1 Introduction

One of the persistent challenges in the operation of building microclimate systems is the decline in energy efficiency resulting from undetected or delayed faults in HVAC (Heating, Ventilation, and Air Conditioning) equipment. Faults such as sensor drift, clogged air filters, and control system errors can significantly impair the regulation of indoor environmental parameters—including temperature, humidity, and ventilation—leading not only to compromised occupant comfort but also to increased energy consumption.

According to estimates from the International Energy Agency (IEA) and the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), HVAC systems are responsible for approximately 40–50% of the total energy use in buildings [1], [2]. However, a substantial proportion of these systems—over 30% according to recent studies—operate under fault conditions [3], [4]. These may include sensor miscalibrations, component wear, or inefficient control algorithms. Particularly in the absence of automated fault detection and diagnostic systems, such issues can persist unnoticed for extended periods. Katipamula and Brambley [5] have reported that undiagnosed faults can increase energy consumption by 10–30%, potentially accounting for an additional 15–25% in annual global energy usage.

In the context of Kazakhstan's national objective to join the ranks of the world's top 30 most developed countries, energy efficiency and sustainable development have become key strategic priorities. The government's recent initiatives to modernize infrastructure and enhance energy performance in the construction sector underscore the urgency of deploying intelligent building technologies. However, the lack of robust fault detection mechanisms and the limited integration of automated diagnostic tools continue to result in significant energy losses, decreased occupant comfort, and rising operational costs.

The growing adoption of smart building technologies—driven by the proliferation of Internet of Things (IoT) devices and sensor networks—has enabled the continuous monitoring of indoor environments. These systems generate vast quantities of real-time data, creating opportunities for advanced data analytics to detect abnormal system behavior. Yet, the practical application of such analytics is often hindered by the scarcity of labeled datasets required for supervised machine learning approaches. In contrast, unsupervised learning techniques such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have shown promise in identifying hidden patterns in high-dimensional, noisy datasets without the need for manual labeling. As noted by Xu and Wunsch [6], DBSCAN is particularly effective in discovering anomalies in complex sensor environments.

This study builds on these insights by exploring the use of the DBSCAN algorithm for unsupervised anomaly detection in building microclimate systems. Utilizing sensor data collected from both residential and commercial facilities, the research aims to address the practical challenges of fault detection in unlabeled data environments. The overarching goal is to contribute to the development of intelligent diagnostic frameworks that enhance the energy efficiency, operational reliability, and sustainability of HVAC systems.

Specifically, this study seeks to answer the following research questions:

- How can hidden faults in building microclimate systems be detected in the absence of labeled data?
- What machine learning techniques are most effective for identifying anomalies in sensor data from diverse building types?
- How can data-driven diagnostics improve the energy performance and reliability of HVAC operations?
- To address these questions, the main objectives of the research are:
- To develop an experimental data collection framework for monitoring indoor environmental conditions;
- To implement a fault detection methodology based on unsupervised machine learning;
- To validate the use of the DBSCAN clustering algorithm for anomaly detection in HVAC-related sensor data;
- To evaluate the practical applicability of the proposed approach in the context of smart building automation.

The research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as a structured methodology to guide the end-to-end data analysis lifecycle, from problem understanding to deployment and evaluation. This approach ensures a systematic handling of the data mining tasks involved in fault detection within building microclimate systems.

Sensor data were collected from both residential and non-residential buildings using a custom-

designed environmental monitoring system, which included temperature, humidity, CO_2, and air pressure sensors, among others. The hardware was configured to capture high-resolution time-series data under real-world operational conditions, enabling the collection of a rich dataset for analysis.

The preprocessing phase consisted of multiple stages:

Data Cleaning: Outlier removal, handling of missing values using interpolation and imputation techniques, and sensor drift correction.

Normalization: Feature scaling was performed using min-max normalization to ensure uniform value ranges across sensor channels, improving algorithm performance.

Dimensionality Reduction: Principal Component Analysis (PCA) was employed to reduce the dimensionality of the dataset while preserving the variance structure. PCA helped mitigate the curse of dimensionality, accelerate computation, and improve the separation of normal and anomalous patterns in subsequent clustering steps.

For anomaly detection, the study utilized the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm—an unsupervised machine learning technique known for its ability to identify clusters of arbitrary shape and to distinguish noise points. DBSCAN does not require prior knowledge of the number of clusters and is particularly suitable for applications where labeled data are unavailable or infeasible to obtain. Key parameters, such as epsilon (ε) and minimum number of samples (minPts), were tuned experimentally to optimize the clustering output. The selection of ε was guided by k-distance plots to identify the point of maximum curvature, ensuring optimal density thresholding.

To assess the performance of the clustering, both qualitative and quantitative evaluation methods were employed:

Visual inspection: Clustering results were projected into two-dimensional space using the top principal components, allowing for interpretability and manual validation of anomaly separation.

Comparative analysis: The results from DBSCAN were compared to those obtained using the K-means clustering algorithm. Unlike DBSCAN, K-means assumes spherical clusters and requires the number of clusters to be predefined. This comparison highlighted DBSCAN's better ability to isolate noise and adapt to irregular cluster structures in the data.

Overall, the methodological approach ensures a robust, data-driven process for unsupervised anomaly detection, tailored to the unique challenges of fault identification in HVAC and indoor environmental systems.

2 State of the art

Recent advancements in building energy efficiency and automation have sparked significant interest in the development of data-driven methods for detecting faults in HVAC (Heating, Ventilation, and Air Conditioning) systems. Fault detection in HVAC systems plays a crucial role in maintaining optimal indoor comfort, reducing energy consumption, and ensuring system longevity. The literature on this topic categorizes fault detection techniques into two primary types: model-based and data-driven approaches.

Model-based methods rely on physical system models to simulate HVAC behavior, and deviations from these expected behaviors are flagged as potential faults. However, these methods are often system-specific, computationally expensive, and time-consuming to develop and calibrate. In contrast, data-driven methods leverage sensor data to detect anomalies, offering a more flexible and scalable alternative. Katipamula and Brambley [4] provided a comprehensive classification of these methods, highlighting the advantages of data-driven approaches in terms of adaptability and ease of integration into existing building systems.

Among data-driven techniques, supervised learning methods such as neural networks and support vector machines (SVMs) have been widely applied for HVAC fault diagnosis. For instance, Kusiak *et al.* [5] used these approaches to classify faults based on historical sensor data. While promising, these techniques have a significant limitation: they require large, labeled datasets for training, which are often unavailable in practice due to the difficulty of obtaining annotated fault data in real-world environments. The lack of labeled data

thus represents a key challenge in deploying supervised machine learning for HVAC fault detection.

To address this limitation, unsupervised learning methods, particularly clustering algorithms, have gained attention in recent years as an effective alternative. Unlike supervised methods, unsupervised techniques do not require labeled data and instead focus on discovering inherent patterns and anomalies in the data. Clustering, in particular, is useful for identifying regions of the data space that deviate from normal system behavior. Wang et al. [6] emphasized that clustering algorithms like k-means can provide a baseline for anomaly detection but are limited by their assumption of spherical clusters and the need to predefine the number of clusters, which can lead to poor performance in datasets with complex, irregular structures.

In contrast, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) offers a more flexible solution. As highlighted by several studies [7], DBSCAN excels at detecting dense regions in data and identifying outliers without requiring prior knowledge of the number of clusters. This feature makes DBSCAN particularly effective for handling real-world data, where cluster shapes may be non-spherical and the number of anomalies is not known *a priori*. The ability to detect noise, or outliers, is particularly valuable for fault detection, where faults often manifest as rare or abnormal events that do not conform to the general pattern of system behavior.

To further enhance the robustness of anomaly detection in high-dimensional sensor data, DBSCAN is often combined with Principal Component Analysis (PCA) for dimensionality reduction. PCA helps to reduce the data's dimensionality while preserving most of its variance, allowing clustering algorithms like DBSCAN to operate more effectively. PCA-based preprocessing has been shown to improve the detection of subtle faults by mitigating the "curse of dimensionality" and emphasizing the most significant features of the data [8]. This combined approach—DBSCAN with PCA—has proven to be a powerful tool for unsupervised fault detection in HVAC systems.

The work by Omarov et al. (2020) is thematically related to the present study, as both address building microclimate systems. However, their article primarily focuses on the development of an adaptive Fuzzy-PID controller for temperature regulation, whereas the current research concentrates on fault detection methods and improving the energy efficiency of microclimate systems through the use of machine learning algorithms. Furthermore, this study includes the development of an experimental model based on IoT devices for real-time data acquisition, alongside an analysis of clustering techniques for diagnosing hidden faults, thereby enhancing the reliability and performance of these systems. The effective utilization of the DBSCAN algorithm yielded visually well-separated clusters and clear identification of outliers, indicating that the main goal of accurate fault detection has been successfully achieved.

In modern integrated microclimate systems, intelligent methods such as multi-agent systems are widely applied to coordinate between various rooms in residential and non-residential buildings, optimizing the balance between comfort and energy consumption (Altayeva & Omarov, 2018). Thanks to distributed control and the ability to adapt to changing conditions, these systems enhance flexibility, adaptability, and scalability of microclimate management. To improve the reliability and robustness of such systems, the implementation of effective fault diagnosis methods is required, enabling timely detection of hidden equipment failures and reduction of excessive energy consumption - an essential factor for maintaining high comfort levels and minimizing operational costs.

3 Methodology

3.1 Theoretical Background

This research is firmly rooted in the principles of sustainable development and intelligent building management, with a focus on optimizing building operations through advanced data analytics and machine learning techniques. The study specifically applies unsupervised machine learning methods—including clustering and dimensionality reduction—to address the critical issue of fault detection in building microclimate systems, particularly HVAC (Heating, Ventilation, and Air Conditioning) systems. This approach aligns with the growing need to improve energy efficiency, reduce operational costs, and enhance

indoor comfort within modern buildings.

The core diagnostic challenge in this research is the identification of anomalies in the operational parameters of HVAC systems, which are continuously monitored through multivariate sensor data. These sensors collect high-dimensional data streams related to environmental factors such as temperature, humidity, airflow, and air quality, among others. The challenge lies in detecting subtle deviations from normal operational patterns that may indicate system malfunctions, such as sensor drift, clogged filters, or mechanical failures, which can often go unnoticed for extended periods.

To address this challenge, dimensionality reduction and clustering methods were employed. Dimensionality reduction was carried out using Principal Component Analysis (PCA), a widely used technique in multivariate data analysis. PCA projects the high-dimensional data into a lower-dimensional space by retaining the directions (principal components) that capture the maximum variance in the dataset. This transformation simplifies the data while retaining its essential structure, thus improving the performance of subsequent machine learning algorithms and mitigating the issues caused by high-dimensionality, such as the curse of dimensionality.

After reducing the data's dimensionality, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was applied for anomaly detection. DBSCAN is a powerful unsupervised clustering algorithm that identifies clusters of varying shapes and densities, making it particularly suitable for real-world applications where the data often exhibits complex structures. The algorithm requires two critical parameters:eps (epsilon), which is the maximum radius of the neighborhood around a point; and min_samples, which is the minimum number of points required within the eps radius to form a cluster.

A core point is defined as a point that has at least min_samples points within its eps neighborhood. Points that do not meet this condition are classified as border points or noise points (outliers). The ability of DBSCAN to detect these outliers is particularly valuable for identifying rare or abnormal conditions, such as system faults, which often manifest as isolated events not fitting into any defined cluster.

The overall diagnostic workflow is illustrated in the diagram of Figure 1, which outlines the step-by-step process for developing a fault detection system for building microclimate environments.

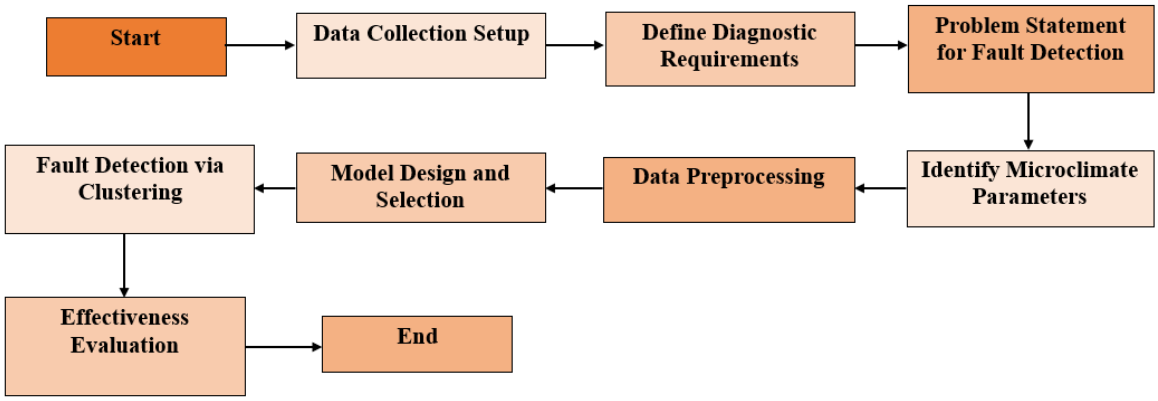


Figure 1: Workflow of Fault Detection System Development for Building Microclimate Environments

The procedure begins with data collection setup, where real-time sensor data from various environmental parameters is gathered from the building's HVAC systems. Following this, the diagnostic goals are defined, which include specifying which types of faults or anomalies need to be detected. The next step involves identifying the relevant operational parameters (e.g., temperature, humidity, airflow rates) and performing thorough data preprocessing, which may include cleaning, normalization, and handling missing values.

Once the data is preprocessed, the next phase involves model selection and clustering application. The choice of DBSCAN as the clustering algorithm is based on its capacity to handle noisy, high-dimensional

sensor data and detect anomalies without requiring pre-labeled datasets. After applying the clustering algorithm, the system's effectiveness is rigorously evaluated by comparing the clustering results with known fault patterns, if available, or by conducting visual and statistical assessments to ensure that anomalies are accurately detected.

By following this methodology, the research provides a robust framework for fault detection in building microclimate systems, with an emphasis on sustainability and energy efficiency. The system is designed to be adaptable, scalable, and capable of real-time monitoring, offering a significant improvement over traditional fault detection methods that rely on manual inspection or predefined thresholds.

4 Results

The study successfully developed an experimental microclimate monitoring model utilizing a custom-designed sensor system, designed in accordance with the recommendations of the International Energy Agency (IEA) and the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) [10,11]. A comprehensive fault detection methodology was implemented based on the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, incorporating advanced machine learning techniques.

The research demonstrated the effectiveness of unsupervised learning—particularly clustering algorithms—for anomaly detection in unlabeled sensor data. Among these, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, configured with optimized parameters, showed strong performance in identifying deviations indicative of hidden faults. To address the high dimensionality of multivariate sensor data, Principal Component Analysis (PCA) was applied for dimensionality reduction, preserving the underlying data structure while minimizing information loss.

The results confirm that the integration of PCA with density-based clustering enables robust detection of operational anomalies without the need for labeled datasets. These findings underscore the potential of unsupervised learning techniques to enhance fault diagnosis capabilities in building microclimate systems. By enabling timely detection of system faults, the proposed approach contributes to increased reliability, improved indoor environmental quality, and greater energy efficiency in modern building management.

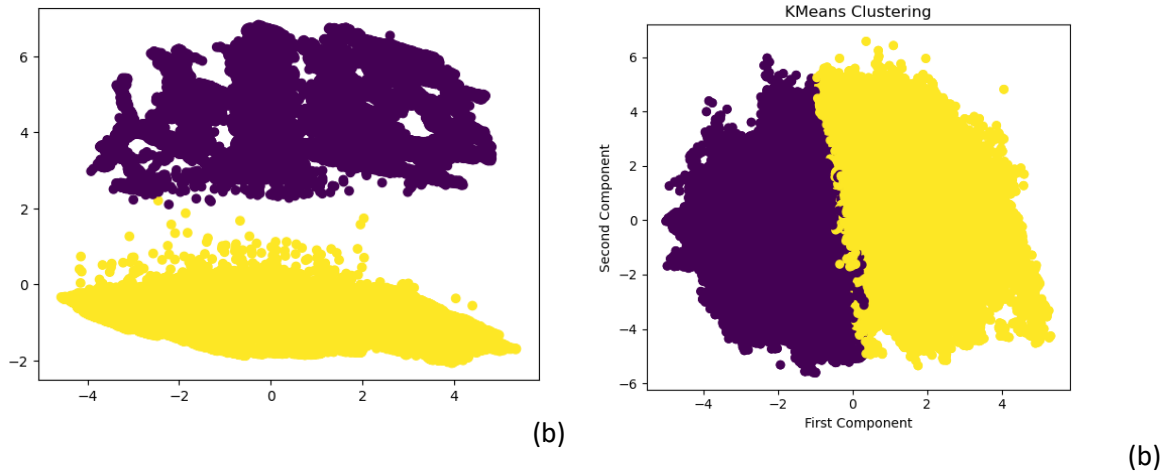


Figure 2: Using K-means for Residential building Cold season (a) and Hot season (b)

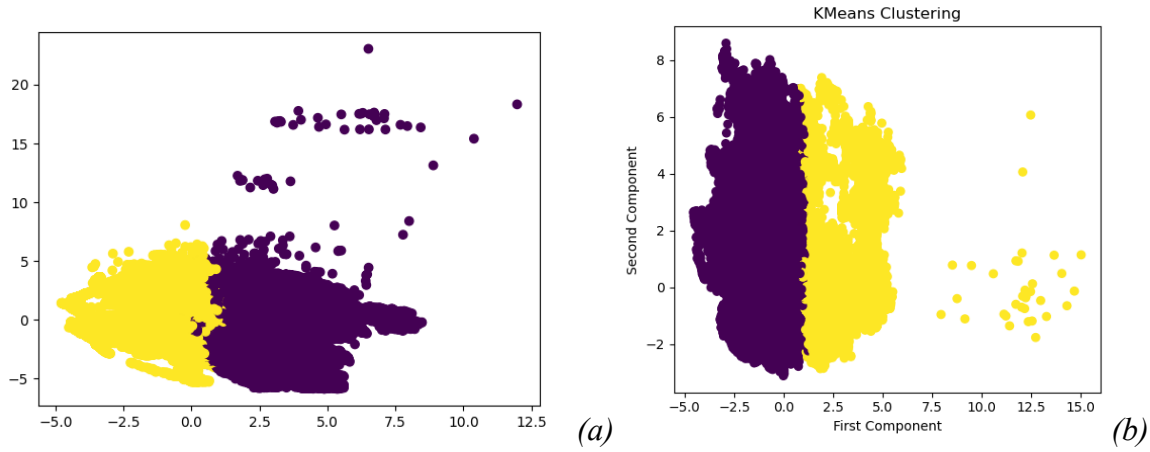


Figure 3: Using K-means for Non - residential building Cold season (a) and Hot season(b)

The K-Means algorithm provided only a simplified clustering of the data, failing to detect outliers and account for density variation, which limits its applicability for fault detection in building microclimate systems compared to the more adaptive DBSCAN algorithm. The K-Means clustering method was applied to microclimate data from a residential building during the cold season (Figure 2a) and the hot season (Figure 2b), as well as to data from a non-residential building during the cold season (Figure 3a) and the hot season (Figure 3b). Although K-Means successfully formed basic clusters, the analysis revealed certain limitations: the clusters were compact and uniform in size, but the algorithm failed to detect noise and outliers—key indicators of hidden faults. These visualizations demonstrate that while K-Means captures the general data structure, it lacks the sensitivity and flexibility required for precise fault detection in building microclimate systems.

Visual analysis of the clustering results confirmed that the DBSCAN algorithm, configured with parameters $\varepsilon = 1.1$ and min_samples = 5, produced well-defined, interpretable clusters and yielded outcomes that are both technically meaningful and practically relevant within the context of HVAC fault detection.

The proposed methodology demonstrates strong potential for integration into Building Management Systems (BMS), enabling real-time monitoring and early detection of anomalies. This would facilitate timely maintenance interventions and contribute to more energy-efficient and reliable building operations.

Figures 4 through 7 illustrate the clustering outcomes, clearly showing the separation between high-density clusters (representing normal operational patterns) and isolated noise points (indicative of potential faults or anomalies). Furthermore, the distribution of detected anomalies across both residential and non-residential buildings is summarized in Table 1, highlighting the method's applicability across different building types and operational contexts.

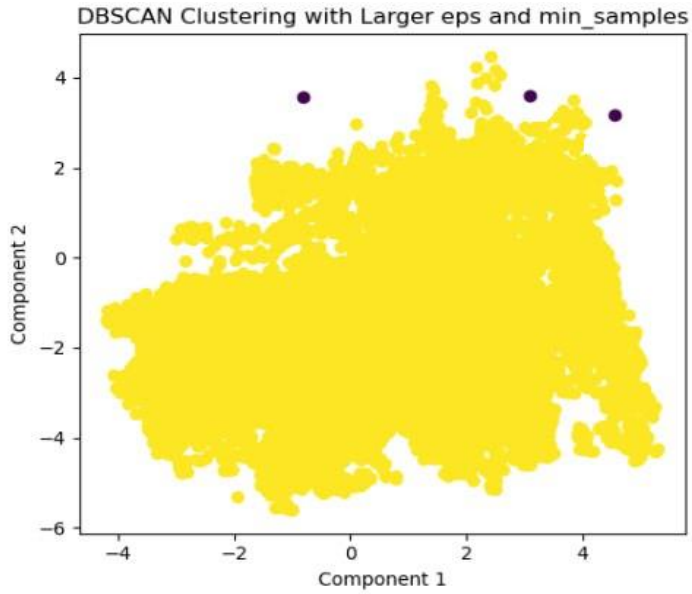


Figure 4: Using DBSCAN for Residential building (Hot season)

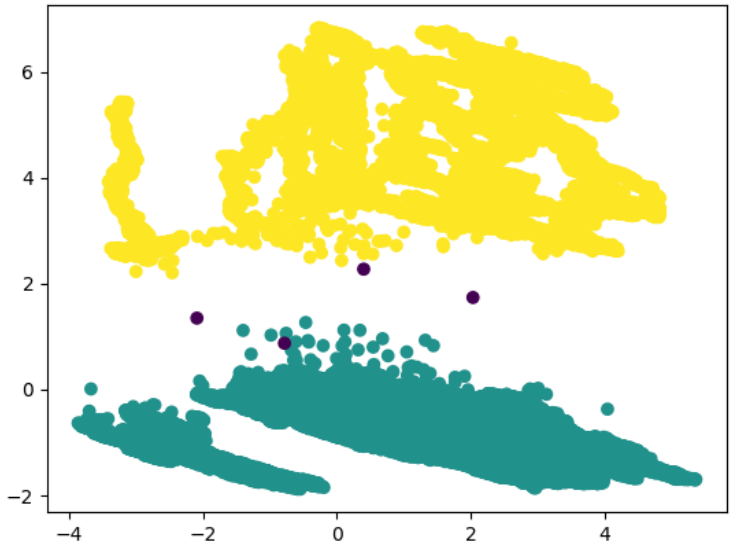


Figure 5: Using DBSCAN for Residential building (Cold season)

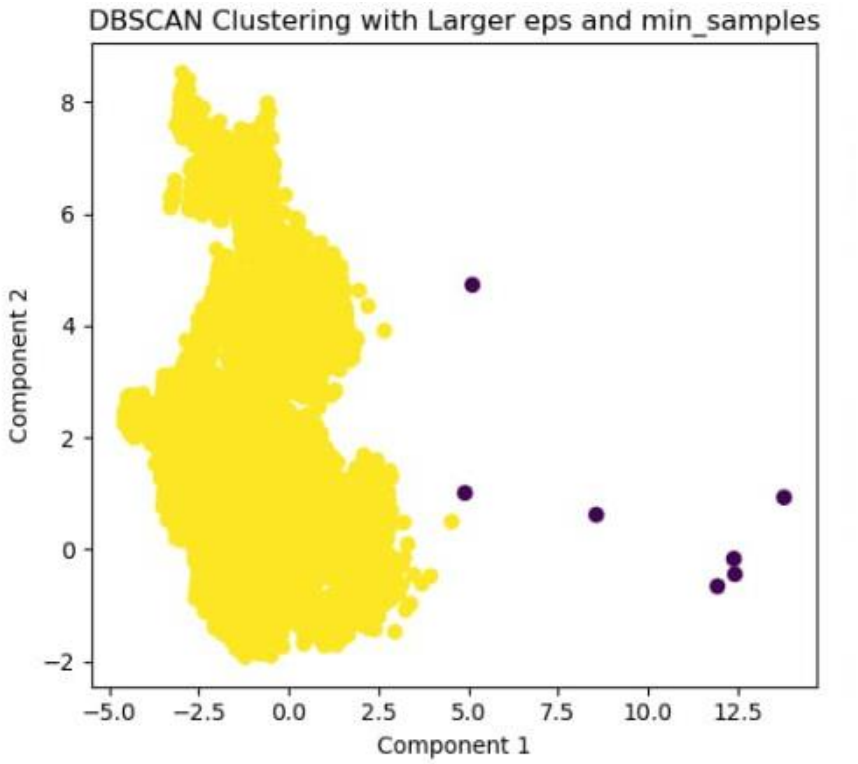


Figure 6 : Using DBSCAN for Non-Residential building (Hot season)

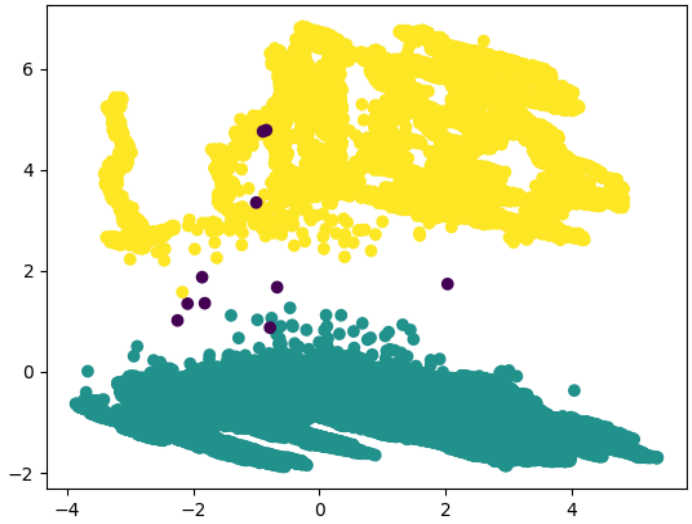


Figure 7: Using DBSCAN for Non-Residential building (Cold season)

Table 1: Residential and Non-Residential buildings result after using DBSCAN

Parameter	Residential (Mean / SD / Min / Max)	Non-residential (Mean / SD / Min / Max)	DBSCAN Result
T (°C)	27.64/164.81/25.49/ 45652.00	27.63/164.51/25.43/45652.00	▨ Outliers detected (extreme Max)
T_{out} (°C)	24.51 / 13.82 / 0.0 / 64.66	24.48 / 13.81 / 0.0 / 64.66	☑ No anomaly
H (%)	32.77 / 2.60 / 26.04 / 40.49	32.70 / 2.61 / 26.04 / 40.49	☑ Stable clusters
H_{out} (%)	35.62 / 20.39 / 0.0 / 82.69	35.60 / 20.37 / 0.0 / 82.69	☑ Stable clusters
DP (°C)	9.22 / 1.18 / 5.70 / 12.86	9.20 / 1.20 / 5.56 / 12.86	☑ No anomaly
P (hPa)	687.37 / 1.52 / 684.18 / 692.03	687.37 / 1.51 / 684.18 / 692.03	☑ Normal density
P_{out} (hPa)	666.42 / 119.52 / 0.0 / 693.12	666.48 / 119.36 / 0.0 / 693.12	▨ Some outliers (Min = 0.0)
CO_2	0.0 / 0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0 / 0.0	⚠ No data / sensor off
TVOC	0.0 / 0.0 / 0.0 / 0.0	0.0 / 0.0 / 0.0 / 0.0	⚠ No data / sensor off
V (V)	227.89 / 31.58 / 0.0 / 0.54	227.92 / 31.55 / 0.0 / 0.54	▨ Outliers (Min = 0.0)
C (A)	0.066 / 0.144 / 0.0 / 96.80	0.0657 / 0.1434 / 0.0 / 96.80	☑ Wide range, but dense
Pwr (W)	12.50 / 26.89 / 0.0 / 96.80	12.46 / 26.85 / 0.0 / 96.80	☑ Reflects usage pattern
UVr	9.19 / 1.65 / 0.0 / 9.60	9.19 / 1.65 / 0.0 / 9.60	☑ No anomaly
L	10548.45/17980.22 / 0.0 / 54612.0	10541.23 / 17948.36 / 0.0 / 54612.0	▨ Outliers (High variance)

The application of the DBSCAN algorithm successfully revealed multiple anomalies within the dataset, demonstrating its effectiveness for unsupervised fault detection in building microclimate systems. Among the anomalies identified were:

- An extremely high temperature reading exceeding 45,000 °C, strongly indicative of a sensor malfunction or data logging error.
- A zero value for outdoor pressure (Pout), which is physically implausible under normal atmospheric conditions and suggests a possible sensor fault or transmission failure.
- A zero voltage (V) reading, which deviates from expected operational behavior and may signal power disruption or measurement error.
- Significant variability and outlier values in illuminance (L), suggesting potential issues with lighting sensors or environmental disturbances during data capture.

Additionally, the dataset lacked readings for CO_2 and total volatile organic compounds (TVOC) across all records. Due to this complete absence of data, these parameters were excluded from the clustering analysis to maintain data integrity.

All anomalies detected by DBSCAN were classified as outliers, meaning they did not belong to any

high-density clusters representing typical system behavior. This aligns with the nature of DBSCAN, which explicitly identifies points lying in low-density regions as noise or anomalies.

Prior to clustering, the dataset underwent Z-score normalization, which standardized the features and helped mitigate the influence of extreme values. This preprocessing step also facilitated the identification and removal of statistical outliers, improving the quality of the input data. Principal Component Analysis (PCA) was then applied to reduce dimensionality, ensuring computational efficiency and improving cluster separability. The cumulative explained variance from the PCA confirmed that the primary structure and variability of the data were retained, indicating minimal information loss.

Subsequently, DBSCAN was employed for anomaly detection, using the normalized and PCA-transformed dataset. The results were evaluated through visual inspection, which provided an intuitive and informative representation of the clustering structure, especially in the absence of ground truth labels or conventional performance metrics—common challenges in unsupervised learning tasks [9, 10]. This visual evaluation proved particularly valuable in confirming the validity and interpretability of the identified anomalies.

5 Discussion

The results of this study support the initial hypothesis that the combination of DBSCAN and Principal Component Analysis (PCA) constitutes an effective unsupervised approach for fault detection in building microclimate systems. This methodology offers several advantages over conventional supervised learning techniques, particularly in domains where labeled data are scarce or unavailable. Key strengths of the proposed approach include:

- No reliance on labeled datasets, making it suitable for real-world environments where annotation is costly or infeasible;
- Capability to detect clusters of arbitrary shape, which is essential for capturing complex, nonlinear operational patterns;
- Explicit identification of noise and outliers, enabling the detection of subtle or isolated system anomalies that may otherwise go unnoticed.

These findings are consistent with prior work by Ester *et al.* [7] and Wang *et al.* [6], who demonstrated the effectiveness of DBSCAN in processing noisy, high-dimensional sensor data and identifying operational anomalies. The success of this study further validates DBSCAN's applicability in the context of intelligent building systems, particularly when paired with PCA to address challenges related to high-dimensional feature spaces.

This research makes several notable contributions to the field of intelligent fault detection in building environments:

- A novel fault detection methodology, structured within the CRISP-DM framework and integrating unsupervised machine learning techniques for robust anomaly identification;
- Empirical validation of DBSCAN as a reliable algorithm for real-time detection of faults in HVAC and microclimate systems, without the need for prior labeling or manual thresholding;
- Methodological justification for DBSCAN parameterization, including practical guidelines for selecting eps and min_samples in systems with similar operational characteristics and feature distributions.
- While the current approach demonstrates significant promise, several avenues for future work can enhance its applicability and scalability, namely:
- Integration into Building Management Systems (BMS): Embedding the developed fault detection model into operational BMS platforms will enable real-time monitoring and automated alerts, facilitating proactive maintenance and energy optimization.
- Hybrid models combining unsupervised and semi-supervised learning: Incorporating domain knowledge or leveraging limited labeled data could improve detection accuracy and reduce false

positives.

- Adaptive parameter tuning: Implementing dynamic adjustment mechanisms for DBSCAN parameters in response to changes in system behavior or data patterns could further increase robustness.
- Expansion to multi-building or district-level systems: Scaling the methodology for use across smart city infrastructures could support broader energy management strategies and cross-system anomaly detection.
- Development of evaluation benchmarks: Creating standardized datasets and performance metrics tailored to unsupervised HVAC fault detection would allow for more objective comparisons between methods.

6 Conclusion

This research addressed the critical issue of undetected faults in building microclimate systems, which can significantly compromise both energy efficiency and operational reliability. To tackle this problem, a comprehensive methodology was developed based on the CRISP-DM framework, incorporating unsupervised machine learning techniques for fault detection.

The core outcomes of the study include the design, implementation, and real-world deployment of an experimental sensor-based microclimate monitoring model. This system enabled the collection of high-resolution multivariate data from building environments. PCA was employed for dimensionality reduction, effectively preserving the data's variance structure while improving the performance of downstream analysis. DBSCAN was subsequently applied for anomaly detection, successfully identifying abnormal operational behaviors without the need for labeled datasets. The approach was validated using real-world data, confirming its practicality and robustness for fault diagnosis in HVAC systems.

The scientific contributions of this work are multifaceted. It presents a novel integration of unsupervised clustering methods into the diagnostic workflow for building automation systems. It offers an empirical validation of DBSCAN's parameter configuration (ε and min_samples) for practical application in noisy, real-world sensor environments. It demonstrates the feasibility of data-driven fault detection without reliance on predefined fault models or supervised learning.

Despite these strengths, the study also highlighted several limitations. DBSCAN's performance is highly sensitive to parameter tuning, and its clustering effectiveness decreases when applied directly to high-dimensional, unprocessed data. Furthermore, due to the lack of standardized evaluation metrics for unsupervised learning, the validation of clustering results relied primarily on visual assessment, which introduces subjectivity and limits reproducibility.

To address these challenges and further enhance the proposed methodology, future research will focus on the following directions:

Exploring advanced clustering techniques, such as HDBSCAN (Hierarchical DBSCAN), which can dynamically determine the number of clusters and is less sensitive to parameter selection;

Extending the system for real-time anomaly detection, incorporating temporal analysis to capture evolving fault patterns and time-dependent system behaviors;

Conducting comparative evaluations involving other machine learning methods (*e.g.*, Isolation Forests, Autoencoders, Gaussian Mixture Models) across diverse building types and climate zones;

Integrating evaluation frameworks for unsupervised models, including silhouette scores, Davies–Bouldin indices, and expert feedback loops, to establish more objective performance metrics.

This study demonstrates that unsupervised machine learning, when carefully integrated into building diagnostics, offers a scalable and effective pathway for enhancing energy efficiency and operational resilience in intelligent building environments.

References

[1] International Energy Agency. Energy Efficiency 2023. IEA Publications, 2023. Available at: https://www.iea.org/reports/energy-efficiency-2023 (last accessed on 2025-01-01).

[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers. ASHRAE Handbook—HVAC Applications, 2020 Edition, 2020.

[3] Katipamula S., Brambley M. R. Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I. HVAC&R Research, 11(1), 3–25, 2005.

[4] Katipamula S., Brambley M. R. Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part II. HVAC&R Research, 11(2), 169–187, 2005.

[5] Kusiak A., Li M., Zhang Z. Modeling and Optimization of HVAC Energy Consumption. Applied Energy, 106, 71–81, 2013.

[6] Wang S., Ma Z., Chen Q. Unsupervised Clustering-Based Fault Detection in Building Systems. Energy and Buildings, 143, 245–258, 2017.

[7] Ester M., Kriegel H.-P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), 226–231, 1996.

[8] Xu R., Wunsch D. Clustering Algorithms in Biomedical Research: A Review. IEEE Reviews in Biomedical Engineering, 3, 120–154, 2010.

[9] Daurenbayeva N., Nurlanuly A., Atymtayeva L., Mendes M. Survey of Applications of Machine Learning for Fault Detection, Diagnosis and Prediction in Microclimate Control Systems. Energies, 2023, 1–21.

[10] Nurlanuly A., Daurenbayeva N. The Study of Human Behavior in the House and Its Role in the Overall Life of the Building in the Field of Energy Consumption. World Sci. Eng. Sci., 1, 24–28, 2019. DOI: 10.31435/rsglobal_ws/31012019/6294.

[11] International Energy Agency. (2017). *Energy Efficiency 2017: Analysis and Outlooks to 2040*. Paris: IEA. (link:https://www.iea.org/reports/energy-efficiency-2017, Data access: 15.05.2025).

[12] ASHRAE. (2017). *ASHRAE Handbook—Fundamentals*. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc. ISBN: 978-1-939200-94-4.

[13] Aigerim Altayeva, Batyrkhan OMAROV (2018), Design of a multiagent-based smart microgrid system for building energy and comfort management// Turkish Journal of Electrical Engineering & Computer Sciences. 2018, 26(5), стр. 2714–2725.

[14] Omarov, B., Anarbayev, A., Turyskulov, U., Orazbayev, E., Erdenov, M., Ibrayev, A., & Kendzhaeva, B. (2020). Fuzzy-PID based self-adjusted indoor temperature control for ensuring thermal comfort in sport complexes. *J. Theor. Appl. Inf. Technol*, *98*(11), 1-12.

Chapter 34

Scalable Data Lake Ingestion using Apache Hudi

Contribution #59
Cite as:
Kafka and Docker, Vasco Oliveira, João Costa. Scalable Data Lake Ingestion using Apache Hudi. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-59,
  title ={Scalable Data Lake Ingestion using Apache
     Hudi},
  author={ Kafka and Docker and Vasco Oliveira and
     João Costa},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Scalable Data Lake Ingestion using Apache Hudi, Kafka and Docker

Vasco Oliveira[1], João Costa[2]
[1,2] Instituto Politécnico de Coimbra, Instituto Superior de Engenharia de Coimbra, Coimbra, Portugal
{a2019119011, jcosta}@isec.pt

Abstract

As organizations increasingly rely on real-time analytics and large-scale data processing, Data Lakes have become essential to ingest and store vast volumes of structured and unstructured data from diverse sources. While most Data Lakes are built on cloud platforms for their ability to scale with growing data volumes, the associated costs can be prohibitively high for smaller organizations. This study evaluates data ingestion performance, using an open-source Data Lake deployed on common off-the-shelf hardware. The setup integrates Apache Hudi for data management, Apache Kafka for real-time ingestion, and Docker for containerized and scalable deployment. Ingestion speed, latency, and scalability were evaluated through a series of controlled experiments varying insert modes, table types, batch sizes, and levels of consumer parallelism. Results indicate that ingestion performance varies according to the insert mode used and the level of consumer parallelism, demonstrating that Apache Hudi, when tuned appropriately, can support high-throughput workloads even on resource-constrained hardware. While most existing research on Data Lakes emphasizes architecture and general functionality, there is limited insight into implementing open-source, non-cloud-based alternatives. This paper bridges that gap by providing a practical, ingestion-focused evaluation of a low-cost, open-source Data Lake, highlighting its feasibility as an alternative to cloud provider solutions.

Keywords: Big Data; Data Lake; Data Ingestion; Open-source; Apache Hudi;

1 Introduction

In today's data-driven world, organizations are struggling with the rapid growth of data generated from a variety of sources, including social media, IoT devices, and business applications. This explosion of data has not only increased in volume but also in velocity and variety, making it increasingly challenging to manage, store, and process into data warehouses (DW). To address these challenges, modern enterprises have turned to Data Lakes, scalable and flexible repositories designed to accommodate large datasets of structured, semi-structured, and unstructured data [1].

Unlike DW, which rely on predefined schemas and structured data, Data Lakes offer a more flexible "schema-on-read" approach, allowing data to be stored in its raw form and structured only when queried or processed. This adaptability has made Data Lakes an appealing solution for organizations looking to manage big data analytics, machine learning, and real-time data processing. With the increasing adoption of Data Lakes, their architecture has become progressively more complex, as it now integrates multiple interconnected components, including storage, ingestion, processing, and governance layers.

Most Data Lake solutions make use of cloud providers tools such as Amazon Web Services, Google's BigQuery or Microsoft Azure's Data Lake [2], which can scale to current data requirements automatically, facilitating Data Lake development by having integrated data analytics and machine learning tools. However, these providers also accompany high costs derived from every transaction that occurs within the Data Lake. These costs may be prohibitive for smaller organizations with predictable data needs.

In comparison to cloud-based solutions, open-source Data Lake architectures present a compelling alternative for organizations aiming to reduce costs and maintain greater control over infrastructure. Frameworks such as Apache Hudi [3], Delta Lake [4] and Apache Iceberg [5] enable the construction of scalable Data Lakes without the need for proprietary services or vendor lock-in. When combined with containerization technologies like Docker and orchestration platforms such as Kubernetes, these solutions

allow for flexible, modular deployments on local or hybrid infrastructure. This approach can be advantageous for small and medium-sized enterprises with predictable data requirements and limited budgets, offering the potential to achieve many of the benefits of cloud-native Data Lakes such as real-time ingestion, schema evolution, and fast analytics at a fraction of the cost.

While most of Data Lake research focuses on the more conceptual side of data management, such as architecture and general functionality, there remains a gap in studies that evaluate the real-world performance of low-cost, open-source alternatives. Solutions that can be deployed on commercial off-the-shelf (COTS) hardware are rarely assessed in terms of their ingestion capabilities, scalability, and cost-efficiency. This lack of practical evaluation is especially relevant for organizations seeking affordable, on-premises Data Lake infrastructures. Addressing this gap, the present study offers a hands-on, performance-driven analysis of an open-source data ingestion pipeline using Apache Hudi, Apache Kafka [6], and Docker [7].

To this end, we explore the following research question: Can a Data Lake built entirely with open-source tools and deployed on local infrastructure provide adequate ingestion performance for periodic, high-volume data streams? Our objective is to measure ingestion throughput, latency, and scalability under various configurations to determine the feasibility and limits of this approach.

This paper is structured as follows: Section 2 reviews related work. Section 3 outlines the methodology and experimental setup. Section 4 presents the results and comparative performance analysis. Section 5 discusses key insights, practical implications, and future research. Section 6 concludes with the main findings.

2 State of the art

The concept of Data Lake emerged in response to the limitations in handling characteristics of big data, namely the need to manage structured data and unstructured data for storage and analysis [8]. To better understand the role of Data Lakes, it is essential to first grasp the core concept of DW and how their focus differs from that of Data Lakes. DW are repositories specifically designed for storing and analyzing structured data in a predefined schema, enforcing schema-on-write, requiring data to be structured before ingestion. DW focus on high-performance querying and reporting, making them ideal for business intelligence tasks [9]. While DW had long been the standard for structured data storage and analysis, they struggled to accommodate the increasing volume and variety of unstructured and semi-structured data sources such as IoT data, multimedia files and data streams from social media and web platforms.

A Data Lake is a centralized storage repository designed to hold massive amounts of data, encompassing structured, semi-structured, and unstructured formats [2]. Unlike DW, Data Lakes enforce schema-on-read, allowing raw data to be stored and structured later when querying or analyzing the data, offering flexibility to work with any type of data [2]. By postponing schema enforcement, Data Lakes enable rapid data ingestion and accommodate evolving or unpredictable data requirements, allowing for more pliability for data of high variety, reducing the workload required to alter schemas associated with DW. A key distinction between DW and Data Lakes lies in their data processing approaches, while DW usually make use of ETL (Extract, Transform, Load), Data Lakes use ELT (Extract, Load, Transform).

The ETL process is critical to ensure that only clean, structured, and pre-processed data is loaded into DW. During ETL, data is first extracted from source systems and then rigorously transformed to conform to the predefined schema of the warehouse. This transformation includes operations such as cleaning, deduplication, and formatting, which ensure data consistency and optimize it for high-performance querying and reporting. However, this comes at the cost of slower data ingestion due to computational complexity and time required for thorough data transformation before loading it.

In ELT, the data is first extracted from source systems and loaded directly into the Data Lake in its raw form, making use of its schema-on-read approach. By using ELT, Data Lakes prioritize rapid data ingestion without waiting for preprocessing or transformation, ensuring that no information is lost during ingestion making it much faster than ETL, as the data refining processes are done later.

Data ingestion is a foundational aspect of Data Lake architecture as it is designed to ingest raw data

rapidly, supporting high-volume, high-velocity workloads, enabling organizations to capture data from diverse sources. Ingestion can be performed in batch or in streaming, depending on the use case, and is facilitated by modern data frameworks such as Apache Kafka, Apache NiFi[10], or Flink [11]. The ability to ingest data as-is provides greater agility for data exploration and delayed transformation, aligning well with use cases like real-time analytics and machine learning where raw data may need to be revisited or reprocessed.

In addition to cloud-based offerings, there is an ecosystem of open-source solutions that support scalable and flexible Data Lake architectures. Tools like Apache Hadoop [12] and Apache Spark [13] provide foundational storage and processing capabilities, while Apache Kafka and Apache NiFi enable real-time and batch data ingestion from diverse sources. Apache Iceberg, Delta Lake and Apache Hudi offer open table formats that support ACID transactions, time travel, and schema evolution, enhancing reliability and consistency in large-scale data environments.

Apache Hudi (Hadoop Upserts Deletes Incrementals) is an open-source Data Lakehouse platform designed for high performance writes on incremental data pipelines and fast query performance [3]. It provides ACID transactions, enabling reliable data ingestion and updates while maintaining consistency and integrity. Hudi achieves this by using multi-version concurrency control (MVCC), maintaining a metadata structure that tracks which file versions constitute the current state of the table. Transactions read from consistent snapshots of this metadata [14] and commit atomically by updating the metadata structure. Coordination of concurrent access is managed via table-level locks implemented through external services such as ZooKeeper, Hive Metastore, or DynamoDB [15]. Furthermore, Hudi employs optimistic concurrency control to ensure transaction isolation by validating each transaction before committing, detecting conflicts with concurrently committed transactions. Its default isolation level is snapshot isolation, whereby transactions read from a consistent snapshot of committed data at their start time and are allowed to commit only if no other transaction has modified the same files they intend to update [14].

In Hudi, metadata for a lakehouse table is maintained in a dedicated metadata table, while a transaction log checkpoint tracks the current state of changes. For Merge-on-Read (MoR) tables, all record-level inserts, updates, and deletes are stored in row-based Avro files. During query execution, these incremental changes are merged with the base data stored in columnar Parquet files to present a consistent, up-to-date view. Importantly, Hudi performs deduplication and sorts ingested data by record keys by default, even in MoR mode, which, while intended to optimize incremental ingestion, can introduce additional write latency. Hudi exhibits significantly slower data loading times, approximately ten times slower than other open-source platforms (Apache Iceberg and Delta Lake) primarily because it is optimized for keyed upserts rather than bulk ingestion [14]. During data loading, Hudi performs computationally expensive pre-processing steps such as key uniqueness validation and key re-distribution, which contribute to the latency.

While numerous open-source and cloud-native solutions exist for building data lakes, such as Delta Lake and Apache Iceberg, this study focuses on Apache Hudi due to its distinctive capabilities in supporting record-level updates, incremental data processing, and versatile ingestion modes, including Copy-on-Write (CoW) and Merge-on-Read (MoR) table types. In addition to these table formats, Hudi offers multiple insert types, bulk insert for fast, scalable loading of large append-only datasets, and upsert for handling update-heavy workloads with fine-grained record-level mutations. These features make Hudi especially well-suited for near-real-time ingestion scenarios where low-latency updates and consistency are critical. These attributes position Apache Hudi as a prime candidate for research into scalable and consistent data ingestion strategies within Data Lake environments.

3 Methodology

This section describes the experiment methodology to evaluate the ingestion capability of the implemented Data Lake as well as the implementation itself and the many parts that make the Data Lake setup work.

3.1 Data Lake Architecture

The scenario chosen to build the Data Lake involved a use case of a ride sharing application such as Uber [16]. In this scenario we used multiple producers emulating data from ride requests and consumers ingesting these messages in batches to the Data Lake (Figure 1). Apache Hudi was chosen as the data ingestion and data management tool as it was originally developed by Uber to efficiently handle the ingestion of massive volumes of small log files, offering capabilities such as incremental processing and upserts [17].

Docker is acting as a containerization tool for its widespread adoption, developer-friendly environment, and ability to simplify scaling and deployment. Apache Kafka is being used as a distributed queue manager and was selected for its proven reliability and high throughput, making it ideal for processing massive volumes of messages in real time. Apache Hive Metastore is utilized to manage metadata and ensure concurrency control while Hudi writes data, with Hive metadata stored in a MySQL database for reliable access.

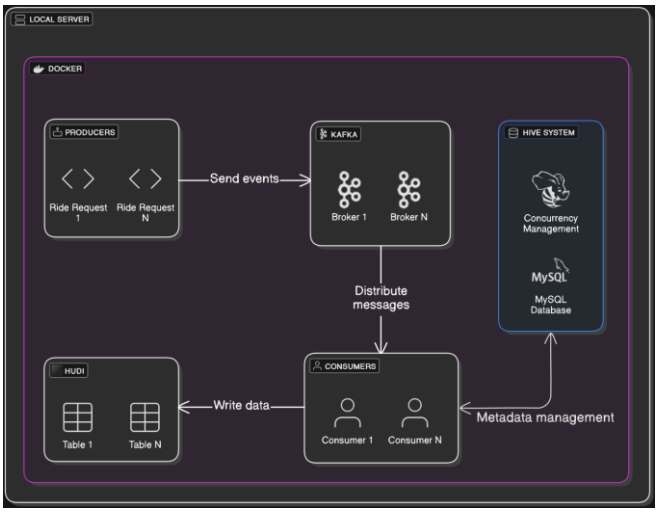


Figure 1: Data Lake Architecture

3.2 Experimental Approach

To simulate real-world Data Lake ingestion patterns, the producer was configured to emit a batch of ride requests ($P_{records}$) every T seconds, reflecting the periodic data loading behavior typical in such environments. The consumer has a limited time T_{limit} (in seconds) to process and ingest as many messages as possible, writing them to disk using Apache Hudi.

The overarching goal of these experiments is to identify the most effective Apache Hudi configuration for handling high-throughput data ingestion. Different settings were tested across multiple dimensions, including insert type, table type, and number of concurrent consumers.

To better emulate the resource constraints of a small organization, we used commercial off-the-shelf (COTS) hardware. The system was deployed on a Debian server equipped with an Intel Core i5 quad-core processor (2.8 GHz), 12 GB of RAM, and a 400 GB HDD.

For each experiment, we logged the following variables for performance analysis:

- **Average Ingestion Rate (*IR*)**: Average number of records ingested into the Data Lake per second.
- **Peak Ingestion Rate (*IRP*)**: Average Maximum ingestion rate achieved during runtime.

- **Average Batch Size (*B*)**: The average size in records of batches ingested to the Data Lake
- **Average Latency (*L*)**: Average time between each batch being inserted into the Kafka messaging queue and being fully written into the Data Lake.
- **Average Batch processing Time (*BP*)**: Average time spent ingesting a batch of records received.

The results in this paper were obtained from a minimum of 5 runs for each experiment to mitigate the influence of transient system fluctuations. An extended evaluation is ongoing to improve statistical reliability.

4 Analysis and Results

This section presents the experimental findings on data ingestion performance using Apache Hudi, focusing on key configuration aspects that influence throughput and scalability. The experiments evaluate how different ingestion strategies, specifically, insert modes (upsert vs. bulk insert), table types (Copy-on-Write vs. Merge-on-Read), and consumer parallelism (single vs. multiple consumers) affect average ingestion rate (*IR*) and overall system behavior.

Unless otherwise specified, upsert operations in the experiments generate unique IDs, thereby effectively functioning as pure inserts. Additionally, a single consumer configuration is employed, and Copy-on-Write tables are used as the baseline setup for comparative analysis across the dedicated experiments.

The experimental results aim to provide empirical insights into the trade-offs between ingestion speed, update handling, and resource utilization. Each subsection details the performance implications of individual settings, followed by a comparative analysis that highlights the most efficient configurations for common ingestion workloads.

4.1 Determining Maximum Data Ingestion Rates

Firstly, we determine the maximum *IR* achievable by a single consumer under varying batch sizes. The producer was configured to emit batches ($P_{records}$) of increasing sizes: 100, 1000, 10000, and finally 100000 messages, with $T = 5$ seconds. To this end, each experimental configuration was executed from a clean state.

As batch size $P_{records}$ increased, consumers were able to ingest more records (Table 1), enabling measurement of the upper bound of ingestion throughput in the test environment (Figure 2 a)). With a single consumer, the system reached a maximum *IR* of 2690 records/second, and maximum *IRP* of 3132 records/second at the highest batch load ($P_{records}$ =100000 messages).

To evaluate the scalability of Hudi, it is necessary to determine the ideal ingestion rate for multiple consumers. After determining the maximum ingestion rate achievable by a single consumer, the next step was to evaluate the system's scalability by increasing the number of consumers. Specifically, the number of consumers increased to three to assess how ingestion performance scales with parallelism. The choice of three consumers finds a balance between demonstrating horizontal scalability and avoiding excessive resource contention, which can obscure performance trends. After increasing the number of consumers to three, performance improved significantly, with maximum *IR* of 5159 records/second and a maximum *IRP* of 7990 records/second at the highest batch load ($P_{records}$ = 100000 messages).

Given the *IR* of 2690 records/second for a single consumer, we can estimate that the optimal batch size for every T=5 seconds would be approximately 13451 messages. This represents the theoretical upper limit of efficient data ingestion without causing backpressure or system overload for a single consumer setup.

Based on the estimated optimal batch size of approximately 13451 records per T= 5 seconds, we defined $P_{records}$ to 15000 for the subsequent experiments. This slightly exceeds the calculated threshold for a single consumer, providing a controlled workload that helps reveal how the system adapts under pressure. The

choice of $P_{records}$ =15000 records allow us to observe the consumer's ability to handle bursts of data beyond its optimal ingestion rate, shedding light on how ingestion performance degrades or stabilizes under sustained load. Additionally, this batch size facilitates more meaningful comparisons between single- and multi-consumer scenarios, especially in identifying where performance plateaus or bottlenecks begin to emerge.

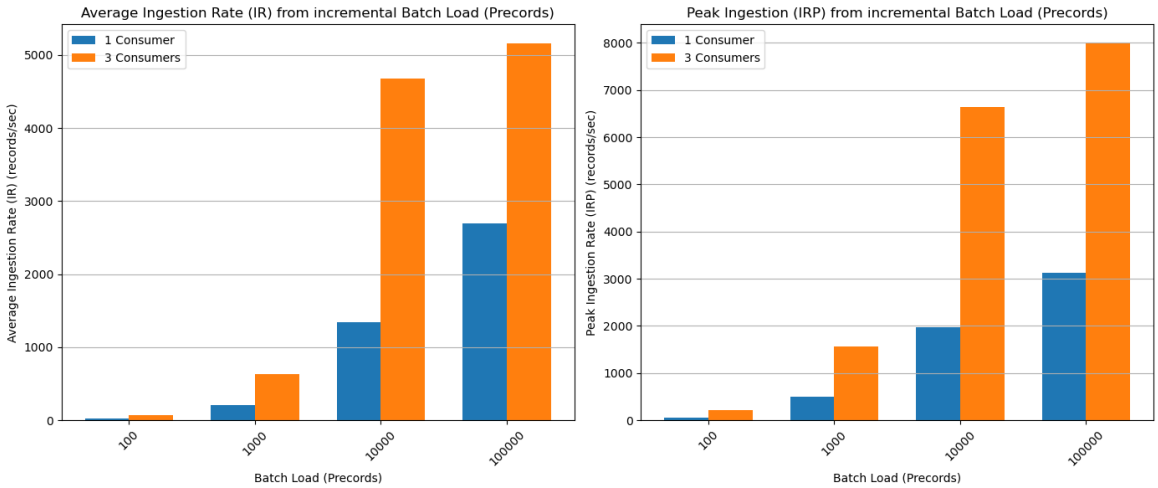


Figure 2: a) Average Ingestion Rate (IR) b) Peak Ingestion per Batch Load (IRP)

Table 1: Ingestion metrics from incremental batch loads

	1 consumer		3 consumers		Scale-Up(3C/ 1C)	
Batch Load	*IR*	*IRP*	*IR*	*IRP*	*IR*	*IRP*
100	20.82	52.83	65.21	217.72	**3.1**	**4.1**
1000	204.04	506.83	629.43	1563.21	**3.1**	**3.1**
10000	1346.36	1964.74	4682.1	6632.77	**3.5**	**3.4**
100000	2690.34	3132.47	5159.77	7990.41	**1.9**	**2.6**

4.2 Batch processing time impact by Insert Mode

This section evaluates how different insert modes (upsert, bulk insert) supported by Apache Hudi affect the batch processing time. The default operation in Hudi is upsert, where incoming records are first classified as inserts or updates by referencing an index. These records are then written to storage using heuristics designed to optimize file organization, such as efficient file sizing. This is ideal for tasks that possess plenty of updates, upserts ensure that the target table remains free of duplicate records [18].

Bulk insert is another ingestion mode optimized for the initial loading of large datasets. Unlike upserts, which retain input records in memory to apply storage heuristics, bulk insert employs a sort-based data writing algorithm that minimizes memory usage. While it offers similar functionality to standard insert operations, it prioritizes scalability over precision in file sizing, delivering best-effort results rather than the strict guarantees provided by bulk inserts or upserts [18].

As expected bulk insert operations are much faster, taking about 61% less time (Figure 3 a)) to ingest the same amount of data as upsert operations, as these include checking for the index of every ingested record for a possible update. Furthermore, bulk insert operations present a 61.9% lower *L*,

presenting a superior performance on real time ingestion scenarios. The lower standard deviation observed in bulk insert operations reflects a more predictable performance, attributable to the absence of index lookups and update checks that introduce variability in upsert operations.

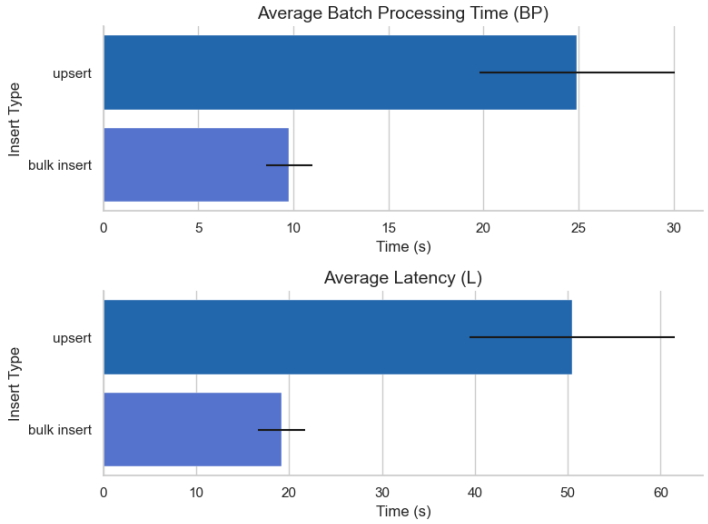


Figure 3: a) Average Batch Processing time (BP) b) Latency (L) by insert type

In a mixed scenario with upserts consisting of 50% inserts and 50% updates, we observed a notable reduction in *BP* approximately 16.4% faster (Figure 4), compared to the pure insert upsert scenario. This improvement can be attributed to reduced write amplification, when records are updated rather than inserted, fewer new files need to be written, allowing Hudi to more efficiently leverage its internal indexing and delta log structures.

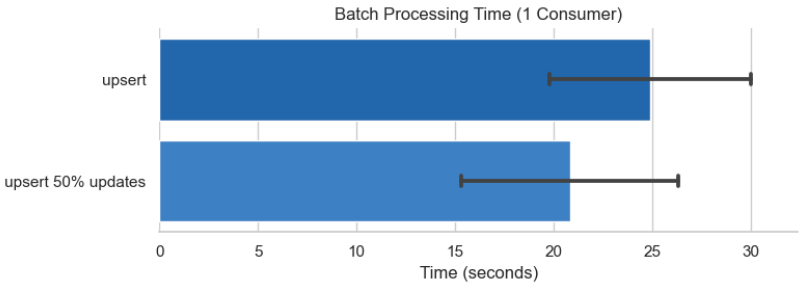


Figure 4: Batch Processing Time with 50% updates

Despite this improvement in mean processing time, the standard deviation remained relatively stable across both experiments, suggesting that while the average workload per batch was lower in the update-heavy scenario, the system still experienced fluctuations in performance likely due to the overhead of maintaining index consistency and handling merge logic during updates.

4.3 Consumer scaling impact on Ingestion Rate

In this section we evaluate the ingestion performance when we scale from 1 to 3 consumers. By introducing multiple consumers, the system can process data in parallel across different partitions, potentially increasing overall throughput and reducing end-to-end latency.

Results show, in Figure 5, that increasing the number of consumers significantly improves ingestion performance across both insert types.

As shown, using three consumers results in a 137.6% increase in *IR* for upserts (from *IR*=1614 to *IR*= 3837 records/sec) and a 241.9% increase for bulk inserts (from *IR*=1741 to *IR*=5954 records/sec), compared to using a single consumer. This improvement is primarily attributed to the parallel processing capabilities provided by Docker, Kafka, and the underlying distributed infrastructure, which enable concurrent ingestion with minimal coordination overhead.

Bulk insert consistently outperforms upsert in terms of *IR*, particularly when scaled to 3 consumers. These results are consistent with the previous experiments show similar improvements in *IR* depending on the insert mode. However, it should be noted that consumers using bulk insert scale *IR* near linearly when increasing consumer number, while consumers using upsert don't show the same improvement.

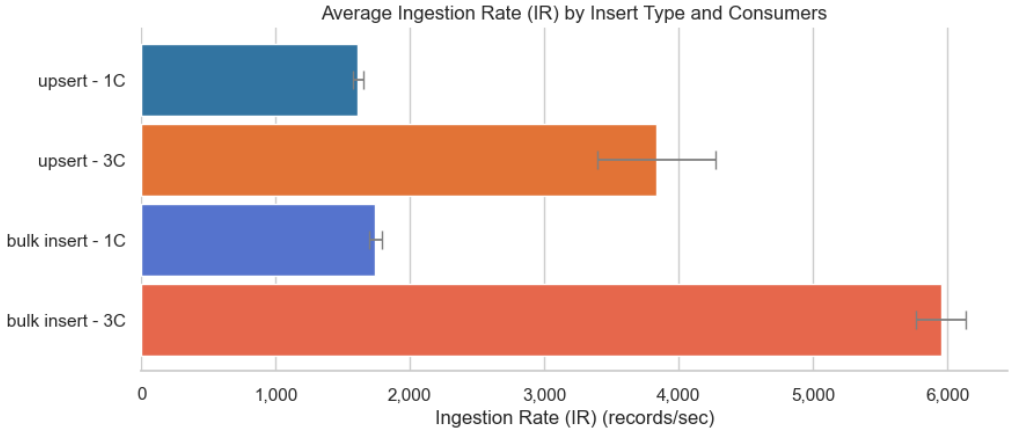


Figure 5: Average Ingestion Rate (IR) by Insert Type and Number of Consumers

The overhead generated by upsert operations becomes particularly impactful in a single-machine setup, as used in this experiment, where all three consumers share the same physical resources. Under such conditions, competition for CPU, memory, and disk further constrains the scalability of upserts. The higher standard deviation of *IR* observed in the three-consumer upsert case (stddev(IR)=439 records/sec) also indicates performance variability likely caused by this resource contention.

These findings highlight the trade-off between ingestion speed and update capabilities. While bulk insert is ideal for high-throughput, append-only workloads, upsert remains essential for scenarios requiring record-level updates and versioning. The choice between them should be guided by workload characteristics and the consistency requirements of downstream applications.

4.4 Batch processing time impact by Table Type

This experiment investigated how the choice of Apache Hudi table type, Copy-on-Write or Merge-on-Read affects data ingestion rates. The Copy-on-Write (CoW) table type is optimized for read-heavy workloads. In this mode, any update or delete operation results in the creation of a new base file within the corresponding file group, without generating log files. As a result, queries only read from base files,

delivering consistently high read performance without requiring on-the-fly log merging. However, CoW tables have the downside of higher write latency, as even minor updates require rewriting of entire base files, making them less efficient for write-intensive operations [19].

Merge-on-Read (MoR) table type offers a trade-off between write and read performance. Updates and deletes are initially recorded in lightweight log files, while base files remain unchanged until a scheduled compaction operation occurs. During query execution, changes from the log files are merged with the base files, either dynamically or after compaction. This results in lower write latency and supports near real-time data ingestion. However, query performance can vary depending on whether log compaction has occurred, as unmerged logs may slow down reads [19].

The difference in *BP* between CoW and MoR was modest, with MoR achieving approximately 2.6% lower *BP* and a 27% reduction in variance (Figure 6). These results indicate that while both table types offer comparable throughput under the tested conditions, MoR provides slightly faster and more consistent ingestion performance.

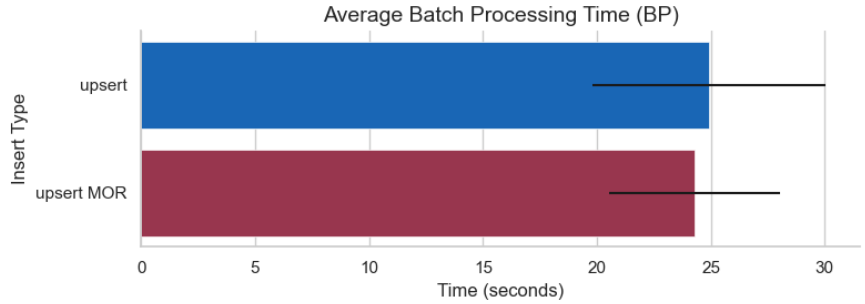


Figure 6: Average Batch Processing Time COW and MOR

5 Discussion and Future Work

Through a series of controlled experiments using Apache Hudi, Apache Kafka, and Docker, we assessed how ingestion configurations, particularly insert type, table format, batch size, and consumer parallelism affect ingestion throughput, latency, and scalability in a Data Lake built with open-source technologies and deployed on common hardware.

By tuning ingestion strategies and system parameters, it is possible to achieve *IR* that support near real-time workloads. From Figure 7 we observe that the most influential factor in ingestion throughput was the choice of insert mode. Bulk insert consistently outperformed upsert across all configurations, confirming that minimizing indexing and update logic allows for faster and more scalable ingestion. This makes bulk insert the clear choice for high-volume, append-only workloads. Upsert operations, however, introduce considerable overhead through indexing and merge processes, resulting in lower and more variable ingestion rates. The experiments revealed that Apache Hudi exhibits improved processing efficiency under update-heavy workloads, batches composed of 50% update records outperformed pure-insert scenarios in terms of *BP*, mostly attributed to reduced write amplification.

The choice of table type also impacted performance, though to a lesser extent. Merge-on-Read (MoR) tables provided slightly better batch processing efficiency and more consistent throughput compared to Copy-on-Write (CoW). This advantage stems from MoR's write-optimized design, which delays heavy file operations until compaction. For ingestion-centric use cases, particularly those prioritizing real-time data availability over immediate query performance, MoR offers a pragmatic balance.

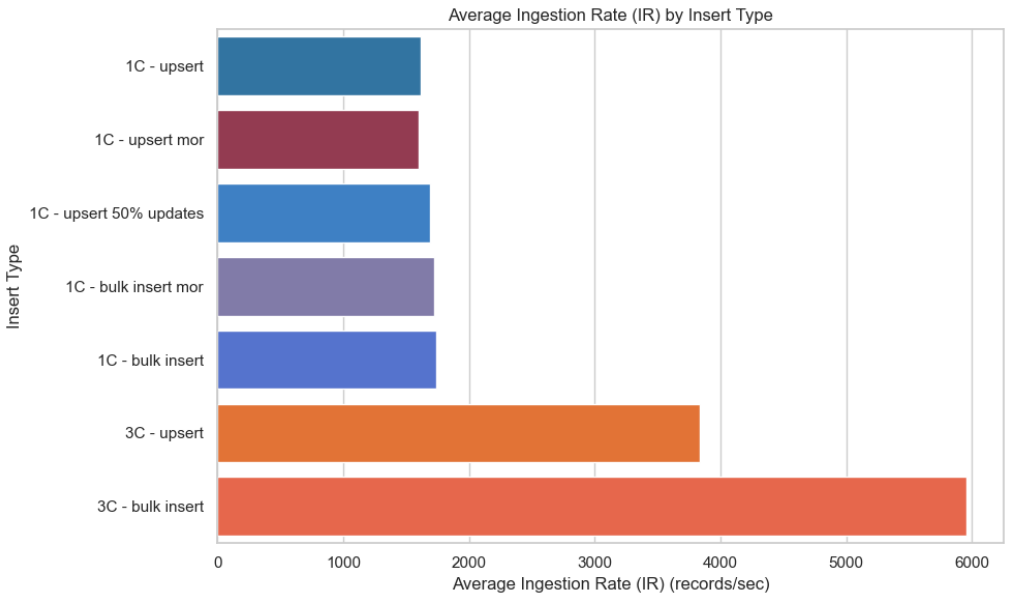


Figure 7: Average Ingestion Rate (IR) by Experiment.

Parallelism proved to be a critical factor in achieving ingestion scalability. Increasing the number of consumers from one to three resulted in significant throughput improvements: 137.6% for upserts and 241.9% for bulk inserts. Bulk insert operations demonstrated near-linear scalability with additional consumers, highlighting their low synchronization overhead and suitability for parallel ingestion pipelines in distributed architectures. Conversely, upsert operations exhibited diminishing performance returns under multi-consumer configurations, likely due to elevated contention for shared system, which is especially pronounced in single-node environments where all consumers compete for the same hardware stack.

The highest ingestion performance observed in this study was achieved using bulk insert mode with three parallel consumers, reaching $IR = 5954$ records per second. This demonstrates that, with minimal coordination overhead and effective resource utilization, the system can fully leverage parallelism through containerized consumers operating across independent Kafka partitions. The results suggest that for high-throughput, append-only workloads, the optimal ingestion pipeline used bulk insert mode to reduce write latency, and scale with multi-consumer parallelism. Such a configuration offers both performance and flexibility, even when deployed on cost-effective, resource-constrained infrastructure.

Building on the findings of this study the next steps involve scaling out consumers across multiple COTS machines, to evaluate the impact of distributed ingestion on performance, fault tolerance, and resource efficiency. This will provide a more realistic assessment of the system's behavior under production-like, multi-node environments, where factors such as network latency, inter-node coordination, and resource contention may become much more prominent. Another avenue is the integration of a dedicated Apache Spark cluster to manage ingestion tasks. Given Spark's capabilities for distributed computation and memory management, this setup is expected to significantly enhance throughput and parallelism.

Furthermore, expanding the scope of the study to include comparative performance evaluations with other open-source storage layers, such as Delta Lake and Apache Iceberg, would offer valuable insights into the many open-source Data Lake frameworks. These systems differ in metadata architecture, transaction models, and query optimization strategies, and a side-by-side analysis could

help identify the most suitable solution for specific ingestion workloads. Collectively, these future directions aim to deepen the understanding of how to build robust, scalable, and cost-effective Data Lake infrastructures using open-source technologies.

6 Conclusion

This study set out to evaluate the performance and feasibility of a low-cost, open-source Data Lake ingestion using Apache Hudi, Apache Kafka, and Docker, deployed on commodity hardware. With cloud-based data platforms often incurring prohibitive costs for small to mid-sized organizations, this work addresses a gap in the literature by offering a practical alternative capable of handling periodic, high-throughput data loads without relying on proprietary infrastructure.

Through a series of controlled experiments, we examined how ingestion performance varies with insert types, table formats, batch sizes, and the number of concurrent consumers. The best configuration achieved uses bulk insert and three parallel consumers (*IR*= 5954 records/sec), highlighting the scalability and efficiency possible without cloud infrastructure. These results validate the feasibility of open-source ingestion pipelines for near real-time, high-throughput workloads in resource-constrained environments.

References

[1] S. Azzabi, Z. Alfughi, and A. Ouda, "Data Lakes: A Survey of Concepts and Architectures," Jul. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/computers13070183.

[2] R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data Lakes: A Survey of Functions and Systems," IEEE Trans Knowl Data Eng, vol. 35, no. 12, pp. 12571–12590, Jun. 2021, doi: 10.1109/tkde.2023.3270101.

[3] "Overview| Apache Hudi." Accessed: 2025-05-28. [Online]. Available: https://hudi.apache.org/docs/overview/

[4] M. Armbrust et al., "Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores," Proceedings of the VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020, doi: 10.14778/3415478.3415560.

[5] "Apache Iceberg - Apache IcebergTM." Accessed: 2025-06-18. [Online]. Available: https://iceberg.apache.org/

[6] "Apache Kafka." Accessed: 2025-06-18. [Online]. Available: https://kafka.apache.org/

[7] "Docker: Accelerated Container Application Development." Accessed: 2025-03-21. [Online]. Available: https://www.docker.com/

[8] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," ITM Web of Conferences, vol. 17, p. 03025, 2018, doi: 10.1051/ITMCONF/20181703025.

[9] A. Nambiar and D. Mundra, "An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management," Big Data and Cognitive Computing 2022, Vol. 6, Page 132, vol. 6, no. 4, p. 132, Nov. 2022, doi: 10.3390/BDCC6040132.

[10] "Apache NiFi." Accessed: 2025-06-13. [Online]. Available: https://nifi.apache.org/

[11] "Apache Flink® — Stateful Computations over Data Streams | Apache Flink." Accessed: 2025-06-13. [Online]. Available: https://flink.apache.org/

[12] "Apache Hadoop." Accessed: 2025-06-13. [Online]. Available: https://hadoop.apache.org/

[13] "Apache SparkTM - Unified Engine for large-scale data analytics." Accessed: 2025-06-11. [Online]. Available: https://spark.apache.org/

[14] P. Jain, P. Kraft, C. Power, T. Das, I. Stoica, and M. Zaharia, "Analyzing and Comparing Lakehouse Storage Systems." [Online]. Available: https://github.com/lhbench/lhbench.

[15] "Concurrency Control | Apache Hudi." Accessed: 2025-05-28. [Online]. Available:

https://hudi.apache.org/docs/concurrency_control
[16] “Setting Uber’s Transactional Data Lake in Motion with Incremental ETL Using Apache Hudi | Uber Blog.” Accessed: 2025-03-03. [Online]. Available: https://www.uber.com/en-PT/blog/ubers-lakehouse-architecture/
[17] “Building a Large-scale Transactional Data Lake at Uber Using Apache Hudi | Uber Blog.” Accessed: 2025-03-19. [Online]. Available: https://www.uber.com/en-PT/blog/apache-hudi-graduation/
[18] “Write Operations | Apache Hudi.” Accessed: 2025-06-13. [Online]. Available: https://hudi.apache.org/docs/write_operations/
[19] “Table & Query Types | Apache Hudi.” Accessed: 2025-06-13. [Online]. Available: https://hudi.apache.org/docs/next/table_types/

Chapter 35

Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care

Contribution #63
Cite as:
Rui Castelo, Mateus Mendes. Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-63,
  title ={Adapting Large Language Models to Support
     Nutritional Decision-Making in Neonatal Intensive
     Care},
  author={Rui Castelo and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Adapting Large Language Models to Support Nutritional Decision-Making in Neonatal Intensive Care

Rui Castelo[1,2,*], Mateus Mendes[1,3]
[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal,
[2] Serviço de Neonatologia, Maternidade Daniel de Matos – ULS Coimbra,
Rua Miguel Torga, 1, 3030-165 Coimbra
[3] RCM2+ Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[*] Corresponding author
a2024161999@isec.pt, mmendes@isec.pt

Abstract

Nutritional adequacy is a key factor in improving the prognosis of very low birth weight infants (VLBW). However, the high workload and scattered clinical data hinder daily systematic analysis. This study investigates the use of locally deployed large language models (LLMs) as assistants for automated data extraction from semi-structured neonatal intensive care unit (NICU) clinical notes. Various text extraction tools were considered (RegEx, ClinicalBERT, MedSpaCy, LLaMA, and Gemma), and an automated analysis pipeline was planned. Then, it is possible that the extracted data can be compared against international neonatal nutritional guidelines (ESPGHAN, SPN) to generate individualized nutritional adjustment proposals. Preliminary results show that LLMs, particularly Gemma 3 12B, can extract clinical data with near-human accuracy and generate context-aware suggestions aligned with clinical practice. This approach offers promising benefits for personalized and efficient nutritional management in VLBW infants, potentially reducing morbidity, mortality, and resource waste.

Keywords: neonatal intensive care; personalized nutrition; LLM; automated data extraction; healthcare artificial intelligence

1 Introduction

Artificial Intelligence has been applied in Medicine with very favorable results, enabling the development and implementation of true precision medicine, with individualized therapeutic and intervention options. In Medicine, despite the large number of models applied in Imaging and in Intensive Care, there is a noticeable gap in the Pediatric and Neonatal areas.

Various Artificial Intelligence models, especially Large Language Models (LLMs), have gained increasing importance as assistants in different fields, as they are capable of quickly absorbing information and answering relatively complex questions in natural language. As an example, we can mention the use of Clinical Key AI©[1], already implemented in several institutions.

In clinical practice, LLMs can offer several advantages, notably assisting professionals in tasks such as identifying similar cases, proposing diagnoses, prognoses, or even therapeutic and nutritional adjustments.

In the present case, the goal is to analyze the feasibility, advantages and limitations of using LLM assistants as support in neonatal intensive care clinical practice, specifically evaluating their capacity to answer

1 https://www.clinicalkey.com/ (last checked on 2025-05-27)

questions in the field of nutrition, assess nutritional practices, and produce adjustment suggestions based on models, consensuses, and guidelines. We aim to improve the prognosis and morbidity/mortality of very low birth weight (VLBW) newborns through better nutritional adaptation to their needs, according to guidelines and consensuses. The model will prepare a nutritional adjustment proposal, based on a comparison between collected parameters and guidelines and consensuses, which the clinician will review and decide whether or not to use.

The population for the study is selected based on Birth Weight (BW) and Gestational Age. Newborns with BW < 1500 g and/or GA ≤ 32 weeks, admitted to the Intensive Care Unit, are selected for the study. The following data will be used: gestational age, day of life, birth weight, current weight, total fluid intake, total caloric intake, total protein intake, total lipid intake, caloric-protein ratio, z-score ratios for weight, length, and head circumference, growth velocity, nutritional supplements used (human milk fortifier, protein supplement), and type of milk used.

If the study results are positive, this model could become a useful tool for daily nutritional adjustment, leading to growth closer to ideal and improving the prognosis and morbidity/mortality of very low birth weight preterm infants. In addition to the health benefits for hospitalized VLBW infants, we also aim to reduce costs by minimizing waste resulting from inadequate and misaligned therapeutic charts.

Ultimately, this model could also be expanded to include other items related to Neonatal Intensive Care.

2 State of the art

Extracting clinical elements from medical texts is a cornerstone of digital health transformation and clinical decision support. Clinical narratives, whether highly structured or entirely free-form, contain the essential details of patient care. In neonatal intensive care units (NICUs), where timely and precise decisions are vital, automating data extraction from clinical notes is not only beneficial but increasingly necessary. This section reviews possible workflows developed to address clinical data extraction, offering a foundation to contextualize the use of large language models in neonatal settings.

2.1 Nature of Clinical Text and Extraction Challenges

Medical texts can be categorized into three levels of organization: structured (*e.g.*, lab results), semi-structured (*e.g.*, clinical progress notes with headings), and unstructured (*e.g.*, free-text consultations or discharge summaries). Structured data can be easily queried, but it often lacks nuance. Unstructured text contains rich clinical context but is difficult to mine systematically.

The extraction of clinical concepts from these texts must account for variability in terminology, abbreviations, negations, temporal references, and context. Medical notes are full of acronyms for which standardization is often lacking, even within a single institution.

2.2 Rule-Based and Pattern-Matching Approaches

Early work relied heavily on rule-based systems, including Regular Expressions (RegEx) and domain-specific scripts. Tools like cTAKES and MetaMap, built on rule-based architectures, map text to standardized medical ontologies like UMLS or SNOMED CT[1], [2]. These systems are designed to identify medical terms in free-text narratives—such as diagnoses, symptoms, or procedures—and align them with structured concepts from ontologies. For example, if a clinical note mentions 'neonatal respiratory distress syndrome', MetaMap can recognize it and link it to its corresponding concept identifier (e.g., C0035222 in UMLS or 69896004 in SNOMED CT). This standardization enables interoperability between systems and supports downstream tasks like analytics, decision support, or population health studies. While powerful in mapping well-known concepts, these tools may struggle

with ambiguous or institution-specific expressions that fall outside predefined vocabularies. These systems offer high precision in tasks such as medication or diagnosis identification but are fragile and require frequent updating to adapt to new text formats or clinical practices. Also, neonatal datasets are scarce.

Pattern-matching remains useful in specific scenarios, particularly when working with templated documentation, such as ICU flowsheets or neonatal daily logs. However, its limitations become apparent when the language used is inconsistent or when semantic understanding is necessary. Context awareness is also an issue.

2.3 Statistical and Machine Learning-Based Systems

The introduction of statistical machine learning improved upon rigid rule-based techniques. Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) became widely used for Named Entity Recognition (NER) in clinical text [3]. These models identify and classify text into predefined clinical categories such as diagnoses, procedures, or medications. For instance, in neonatal care, CRF-based models have been applied to identify 'respiratory support escalation', 'oxygen weaning', or 'necrotizing enterocolitis' within free-text daily progress notes. By learning contextual dependencies, CRFs can disambiguate whether a reference to "ventilation" refers to invasive mechanical ventilation or non-invasive methods like CPAP. Similarly, SVMs trained on annotated corpora can distinguish between therapeutic interventions and diagnostic observations. In neonatal contexts, SVMs have been used experimentally to classify whether phrases such as "NIV escalation" or "oxygen supplementation" represent active interventions versus monitoring status. They have also been tested to differentiate conditions like "intraventricular hemorrhage" as present diagnoses versus historical findings, based on phrase cues. Moreover, these models often struggle to incorporate context beyond sentence level, reducing their utility in extracting temporally and semantically related concepts. These models can be highly effective when trained on domain-specific corpora, but their performance often drops when applied to new datasets, particularly when the clinical language varies significantly across institutions or note types. However, these models often require domain-specific features and substantial amounts of manually labeled data, which are rarely available as neonatal datasets.

2.4 Deep Learning and Contextual Language Models

Deep learning techniques, particularly those based on recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), were a significant step forward. These models offered improved handling of temporal dependencies and non-linear relationships in text. Yet, they were eventually outperformed by transformer-based architectures.

In 2018, Google deployed Bidirectional Encoder Representations from Transformers (BERT). Still today, its clinical adaptations (e.g., ClinicalBERT, BioBERT) continue to offer robust solutions to NER and relation extraction[4], [5]. These models leverage attention mechanisms to consider context from both directions of the input text, providing nuanced understanding that is essential in clinical interpretation. In neonatal domains, these models can extract key information such as type and timing of respiratory support (e.g., 'mechanical ventilation started on D2') or nutritional status (e.g., 'enteral feeds suspended due to abdominal distension'). For example, ClinicalBERT has been applied to identify risk factors in premature infants by capturing subtle cues in documentation that correlate with outcomes like bronchopulmonary dysplasia. The ability to incorporate context allows the model to distinguish between past medical history and current conditions, which is critical in neonatal daily progress notes where temporal awareness is essential. Despite their strengths, such models still require cumbersome fine-tuning to the specific linguistic patterns found in neonatal intensive care notes, as general biomedical corpora may not fully represent neonatal terminology.

For instance, ClinicalBERT has been successfully applied to medication extraction, adverse drug event detection, and de-identification tasks in large datasets like MIMIC-III[6], [7]. MedSpaCy further enhances these capabilities with rule-based extensions, section detection, and negation handling tailored to the clinical domain [8]. It is particularly useful for recognizing the structure and context of clinical documentation. For example, MedSpaCy can identify whether 'apnea episodes noted' appears in the 'Assessment' section versus the 'History' section, changing its interpretation as active versus inactive/past diagnosis. Additionally, its ConText algorithm can determine whether a condition like 'necrotizing enterocolitis' is negated ("no signs of NEC") or hypothetical ("monitor for NEC"). MedSpaCy is especially powerful when used in combination with entity recognizers like spaCy or ClinicalBERT, forming a hybrid rule-based and statistical pipeline that is modular, explainable, and adaptable to local documentation styles.

However, a complete language support is restricted to english (present day), which constitutes a major drawback to effective deployment in this country. The necessary reverse translation is counterintuitive, adding complexity and error-prone.

2.5 Large Language Models in Clinical Practice

The introduction of LLMs like GPT, LLaMA, and PaLM marks a paradigm shift. Unlike earlier models, LLMs can be adapted to diverse tasks through prompting, without needing task-specific training. This flexibility is seen across different families of models: PaLM (Google©) offers multimodal, multitask capabilities integrated into tools like Bard; GPT (OpenAI)© powers applications like ChatGPT and Codex with strong language generation performance; and LLaMA (Meta©) provides efficient, open-source models for local deployment and research. Each has strengths in terms of accessibility, scalability, or domain adaptability, which makes LLMs particularly attractive in low-resource domains such as neonatology, where data are scarce and language diversity requires adaptable solutions [9], [10].

In recent studies, GPT-3 and similar models have shown promising results in summarizing clinical records, generating discharge instructions, and performing zero-shot classification of medical events [10], [11]. Importantly, they can handle temporal and contextual ambiguity more effectively than prior solutions.

However, clinical use of LLMs raises challenges. Hallucination—generation of incorrect but plausible information—is a well-documented issue. Moreover, models may reproduce biases present in training data, and their outputs are not inherently explainable[12]. Clinical deployment thus requires guardrails, such as restricted prompting, human-in-the-loop validation, and domain-specific fine-tuning [13].

2.6 Applications and Case Studies

Several case studies illustrate the successful implementation of clinical NLP tools:

- In oncology, NLP systems have been used to extract staging and treatment plans from pathology reports, enhancing registry completeness and care coordination [14].
- In adult ICU, NLP supports medical decision-making and identifies undiagnosed conditions [15], [16].
- In neonatology, research is more limited. SOFA Score can be predicted from patient notes (also nSOFA) [17]. Preterm birth risk prediction [18] and maternal data extraction are also possible [19].

2.8 Hybrid Pipelines and Integration with EHR Systems

Hybrid pipelines that combine deterministic rules, statistical models, and LLMs are emerging as best practice. For example, a pipeline may use MedSpaCy to detect sections and extract key entities, followed by an LLM to contextualize and summarize the findings. Integration with EHRs could be facilitated by standards such as HL7 FHIR and openEHR, which allow structured outputs to feed clinical dashboards or decision support systems. Presently, in this country, the lack of implementation of such protocols constitutes an irreversible loophole. Also, lack of Portuguese language support is a drawback. Still, theorethically speaking, the former would be the most accurate workflow.

3 Practical Exploration

I) Clinical Data Extraction Models

The most complex and probably least reliable step in the entire process is the extraction of clinical data from text, whether it is highly structured (e.g., medical certificate), semi-structured (e.g., ICU clinical notes), or unstructured/free text (e.g., consultation notes, addenda, etc.).

The available options vary according to the original text, but the analysis is based on the simplest option:

1 Structured Text

a Use of a RegEx (Regular Expression) tool only – as it uses a sequence of characters/search pattern, it can extract specific information. It relies on an exaction pattern. The text would have to be highly structured and rigid. No duplicate data formats are allowed in that text. It would require one RegEx per item intended for extraction.

b Possibly the use of a general LLM could also perform well, but it would require more complex computation.

2 Semi-Structured Text

a Extensive use of RegEx with multiple elements to constrain the extracted data, relying on the most basic structure of the text.

b LLM Models for Element Extraction and Methods for Named Entity Recognition

b.1) BERT / CLINICALBERT – an expanded version of the BERT model more adapted to find medical terms in free text; lighter in computational terms than LLMs; limited to the collection of medical terms; free; not cloud-based; requires the use of MedSpaCy for interaction.

b.2) Lamma x / Gemma x (other) – more generic, more demanding in terms of computing power, less dedicated or specific to medicine; free; not cloud-based; context-sensitive.

c. Custom commercial solutions (such as John Snow Labs©) – paid, dedicated and specific to medicine, customizable on demand (paid), cloud-based

d. Extensive use of RegEx only – it would imply a highly structured text as already stated; one “expression” per item; context-blind.

Table 1 summarizes the advantages and disadvantages of different solutions.

Table 1: Possible pathways – pitfalls and advantages

Solution	Description	Advantages	Pitfalls
RegEx	Uses a search string/pattern Can extract specific information. Implies an exact pattern.	Simple and fast Local; Free	Limited to highly structured texts 2 different data elements must never have similar text Does not contextualize Complex "programming" (?)
BERT/ClinicalBERT	NLP, NER-based Expansion of the BERT model Adapted for medical terms in free/semi-structured text Used for clinical classification, entity recognition extraction, relationship with ontologies (e.g. Snomed) Encoder Tokenization and Vectorization (1) Deterministic	Local; Free Computationally lighter than Lamma / Gemma	Limited to medical term collection only It doesn't work alone: - Complex Programming - Complemented with SpaCy - Complemented by RegEx
Lamma x Gemma x	"Complex", autoregressive NLP "Decoder" Tokenization and Vectorization (2) Probabilistic	Local; Free Can extract Clinical Entities Identifies clinical context and evolution Produces text, tables, etc Adjustment possible with prompt engineering The simplest to integrate into clinical practice (?) No programming required (?) Evolutive (?)	Computationally more demanding Not specific to Medicine Hallucination (can be minimized...) Can be combined with RegEx as well. Prompt Engineering If maximum accuracy is required, precede with ClinicalBert and / or Regex. (3)

Note: (1) fixed vectors per token/phrase for entity extraction (NER), encodes the whole phrase/expression at once. (2) dynamic, contextualized vectors to predict the next token, with inference. (3) extraction is done by ClinicalBERT and the extracted data is processed by Gemma / Lamma, which generates a report, potentially contextualized and interactive.

II) Comparison with Standards

The comparison with standards, aiming at frequent and individualized adjustment of nutritional intake, will be carried out automatically. These standards are available freely and consist, mainly, in the following:

1 *Recommendations for enteral nutrition in the preterm infant* – 2023 update from the Portuguese Neonatology Society:

Part I – Nutritional needs and enteral nutrition during hospitalization
Part II – Specificities of enteral nutrition in special clinical situations and post-discharge feeding

2 *Enteral Nutrition in Preterm Infants: ESPGHAN Position Paper 2022*

These will be summarized and relevant data presented in a table by day of life, gestational age, and birth weight, as well as other relevant somatometric data. A general LLM, possibly the same used previously, will compare the data gathered from the newborn and outline the deviations.

III) Nutritional Intervention Adjustment

The development of an individualized proposal will be carried out by the same LLM model used for data extraction. Eventually, it might be necessary to use another LLM model, more adapted to use the information gathered in steps I and II.

It involves, in a first approach, the preparation of a prompt to systematize the parameters to be compared, the parametrization based on birth weight, gestational age, and day of life, and finally a highly individualized report/proposal. This will be presented to the clinician.

Figure 1 illustrates the workflow proposed.

Workflow

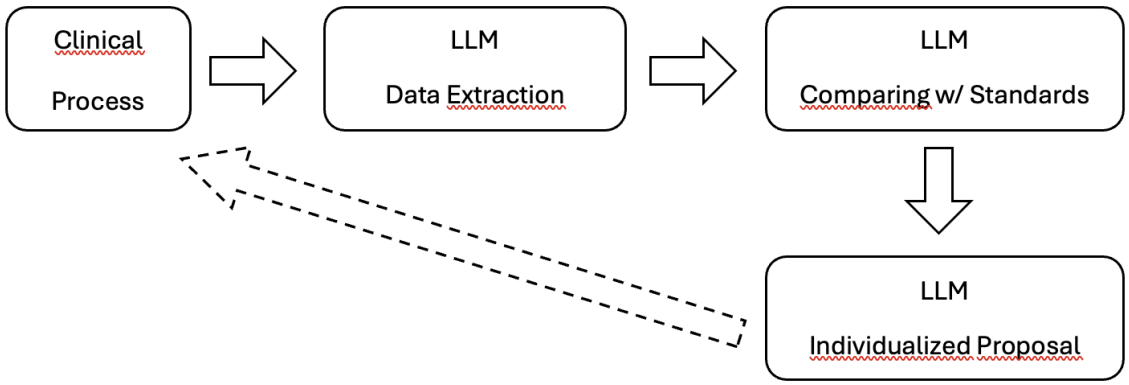


***Figure 1** – Workflow of the method proposed*

Features to Extract and Calculate

Gestational age, day of life, birth weight, current weight, total fluid intake, total caloric intake, total protein intake, total lipid intake, caloric-protein ratio, z-score ratios for weight, length, and head circumference, growth velocity, nutritional supplements in use (human milk fortifier, protein supplement), type of milk used.

4 Methods

A proof-of-concept approach was set up to ascertain the model capabilities to extract relevant data.

The project is based on archetypal clinical records from NICU settings, not real ones, according to patient privacy concerns. Daily progress notes were written in semi-structured format. These notes contain critical information on each infant's nutritional intake, growth parameters, and clinical course.

A pipeline was considered involving several steps:

1 Extract clinical variables using different natural language processing (NLP) strategies - general-purpose LLMs (LLaMA 3.1 4B, Gemma 3 12B) for context-aware data extraction

2 Compare extracted data to the real data in the clinical notes

3 Adjust the NLP parameters to minimize hallucination and grasp the real data

4 Elect the best setup

The models were implemented offline using LM Studio to ensure data security and compliance with hospital data policies. Accuracy and coherence of LLM-generated outputs were assessed against manual expert annotation in test cases. Decoy (non-medical texts) were also used in proofing the LLMs.

5 Results

Among the evaluated methods, Gemma 3 12B showed the best trade-off between performance and system requirements, achieving accurate extraction of key data with minimal hallucination when used with strict prompting. The use of self-consistency techniques (multiple output generation and majority selection) further improved reliability.

The LLM-based system was able to: Automatically identify and structure key clinical features; distinguish between real and decoy clinical notes; map the clinical features according to context; respect real data values.

Some minor improvements are still needed. The collected data need to be validated based on admissible ranges in order to ensure consistency and clinical plausibility (eg the newborn weight in grams must be between 250g and 6000g).

Compared to manual workflows, the system offered significant time savings and reduced variability, especially valuable in busy NICU settings with multiple critical patients.

```
prompt = f"""
Nota para o modelo: AET significa "Aporte Enteral Total". TGV significa "Transfusão de Glóbulos Vermelhos".

Abaixo está um relatório clínico de um recém-nascido.

Por favor, leia cuidadosamente o texto e extraia as seguintes informações clínicas, se estiverem disponíveis:

1. Idade gestacional (em semanas)
2. Peso ao nascimento (em gramas)
3. Dia em que foi iniciada a nutrição enteral
4. Dia em que o recém-nascido atingiu aporte enteral total (150 ml/kg/dia)
5. Tipo de leite utilizado (ex: leite materno, fórmula, leite de banco, fórmula extensamente hidrolisada)
6. Uso de nutrição parenteral (Sim/Não)
7. Se sim, dia de início e término da nutrição parenteral

Por favor, responda em formato estruturado como o exemplo abaixo:

Idade gestacional: ___ semanas
Peso ao nascimento: ___ g
Início da nutrição enteral: Dia ___
Aporte enteral total: Dia ___
Tipo de leite: ___
Nutrição parenteral: Sim/Não
Início da NP: Dia ___
Término da NP: Dia ___

Relatório:
{texto}
"""
```

Figure 1 – Prompt used (in Portuguese) – defining the context and the objective, and the desired structure of the collected data

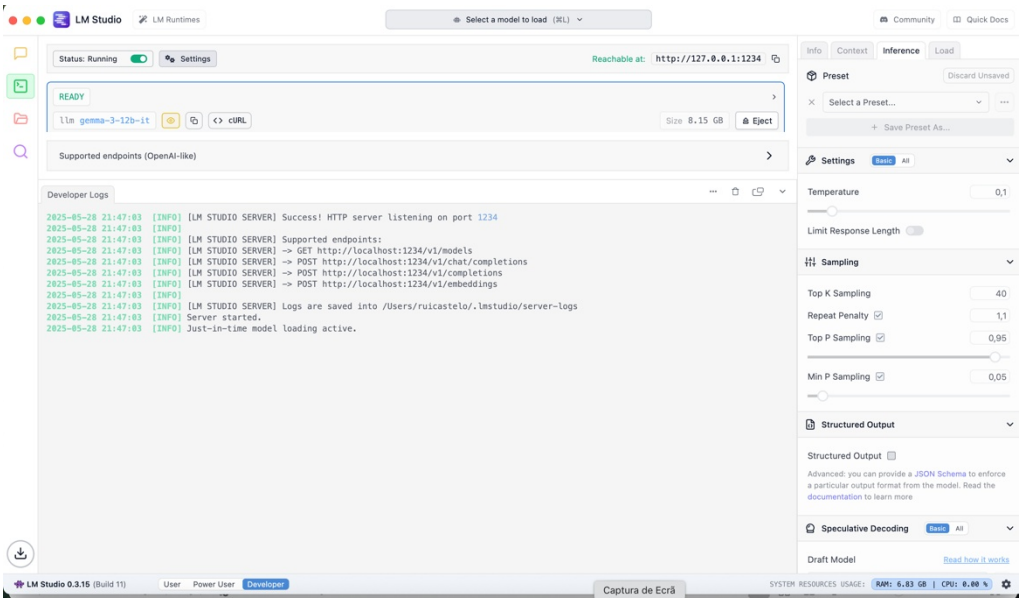


Figure 2 – Using LMStudio to adjust Gemma 3 – topK, topP and temperature

resultados_clinicos_20250527_182302_tabela

Arquivo	Idade gestacional	Peso ao nascimento (g)	Início nutrição enteral (Dia)	Aporte enteral total (Dia)	Tipo de leite	Nutrição parenteral	Início NP (Dia)	Término NP (Dia)
RN1.txt	26	932	7		Leite de banco	Não		
RN3.txt	26	944	9		Fórmula	Não		
RN4.txt	26	1079	8		Fórmula extensamente hidrolisada	Sim		
RN5.txt	26	1001	7		Fórmula extensamente hidrolisada	Sim		
RN6.txt	26	1056	7		Fórmula extensamente hidrolisada	Sim		
RN7.txt	26	1065	8		Fórmula extensamente hidrolisada	Sim		
RN8.txt	26	863	7		Fórmula	Não		
RN9.txt	26	887	8		Não informado	Não		
RN10.txt	26	917	8		Não informado	Não		
RN2.txt	26	1058	7		Fórmula	Não		
RN21.txt		9173469359			Não informado	Sim		
RN22.txt					Não informado	Não		
RN23.txt					Não informado	Não		

Figure 3 – csv table with extracted and structured data (in Portuguese)

6 Discussion

Despite known limitations of LLMs—such as occasional hallucination and non-deterministic behavior—this study shows that strict prompting strategies and offline deployment can mitigate risks and deliver clinically useful outputs. The automated workflow may match or even outperform manual data extraction in terms of speed and consistency, particularly in repetitive tasks. Importantly, the system remains under full clinical control, with LLMs acting as suggestive tools rather than autonomous agents.

This first approach sets the pace for further development of steps II and III

The proposed approach can evolve flexibly: the same pipeline may be adapted to extract additional clinical dimensions (e.g., somatometric data, ventilatory parameters, infection markers) by simply adjusting prompts, without reprogramming or retraining the model. This extensibility makes it suitable for broader integration into clinical decision support systems in neonatal care.

7 Conclusion

Local LLMs show promise as clinical assistants for neonatal nutritional care. Their capacity to extract and process information from semi-structured notes, and to contextualize outputs based on established standards, highlights their potential to support more personalized and efficient care. Although there are some limitations and specific concerns about reliability, this work highlights a non-inferiority level.

Future work will focus on scaling the solution, integrating real-time clinical data, and evaluating long-term outcomes in neonatal growth and morbidity.

References

[1] A. R. Aronson and F. M. Lang, "An overview of MetaMap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, May 2010, doi: 10.1136/jamia.2009.002733.

[2] G. K. Savova *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, Sep. 2010, doi: 10.1136/jamia.2009.001560.

[3] B. Settles, "Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets." [Online]. Available: http://us.expasy.org/sprot/

[4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[5] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/

[6] Y. Wang *et al.*, "Clinical information extraction applications: A literature review," Jan. 01, 2018, *Academic Press Inc.* doi: 10.1016/j.jbi.2017.11.011.

[7] J. Lee *et al.*, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[8] H. Eyre *et al.*, "Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python." [Online]. Available: https://spacy.io/

[9] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[10] H. Zhou *et al.*, "A Survey of Large Language Models in Medicine: Progress, Application, and Challenge." [Online]. Available: https://github.com/AI-in-Health/MedLLMsPracticalGuide.

[11] Z. Yang *et al.*, "Large Language Model–Based Critical Care Big Data Deployment and Extraction: Descriptive Analysis," *JMIR Med Inform*, vol. 13, 2025, doi: 10.2196/63216.

[12] L. Boonstra *et al.*, *Prompt Engineering*. Google Cloud, 2024.

[13] A. Gatt *et al.*, "From data to text in the neonatal intensive care Unit: Using NLG technology for decision support and information management," *AI Communications*, vol. 22, no. 3, pp. 153–186, 2009, doi: 10.3233/AIC-2009-0453.

[14] S. Abedian *et al.*, "Automated Extraction of Tumor Staging and Diagnosis Information From Surgical Pathology Reports," *JCO Clin Cancer Inform*, vol. 5, pp. 1054–1061, 2021, doi: 10.1200/CCI.21.

[15] E. Urquhart *et al.*, "A pilot feasibility study comparing large language models in extracting key

information from ICU patient text records from an Irish population," *Intensive Care Med Exp*, vol. 12, no. 1, Dec. 2024, doi: 10.1186/s40635-024-00656-1.
[16] W. Zhang *et al.*, "Risk prediction models for intensive care unitacquired weakness in intensive care unit patients: A systematic review," *PLoS One*, vol. 16, no. 9 September, Sep. 2021, doi: 10.1371/journal.pone.0257768.
[17] F. H. Saner, Y. M. Saner, E. Abufarhaneh, D. C. Broering, and D. A. Raptis, "Comparative Analysis of Artificial Intelligence (AI) Languages in Predicting Sequential Organ Failure Assessment (SOFA) Scores," *Cureus*, May 2024, doi: 10.7759/cureus.59662.
[18] L. Sterckx *et al.*, "Clinical information extraction for preterm birth risk prediction," *J Biomed Inform*, vol. 110, Oct. 2020, doi: 10.1016/j.jbi.2020.103544.
[19] S. Abhyankar and D. Demner-Fushman, "A simple method to extract key maternal data from neonatal clinical notes."

Chapter 36

ISO 9001 and ISO 14001 Recertification: A Case Study in a Thermoplastics Injection Factory

Contribution #64
Cite as:
João Simão, Hugo Raposo, Ana Vieira. ISO 9001 and ISO 14001 Recertification: A Case Study in a Thermoplastics Injection Factory. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-64,
  title ={ISO 9001 and ISO 14001 Recertification: A
     Case Study in a Thermoplastics Injection Factory},
  author={João Simão and  Hugo Raposo and  Ana Vieira},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

ISO 9001 and ISO 14001 Recertification: A Case Study in a Thermoplastics Injection Factory

João Simão[1*], Hugo Raposo[1,2], Ana V Vieira[1,2,3,4]

[1] Polytechnic University of Coimbra, Instituto Superior de Engenharia de Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.
[2] RCM[2+] Research Centre for Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal
[3] CISE - Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal
[4] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal
*Corresponding author
* a2019134181@isec.pt; hugo.raposo@isec.pt; avieira@isec.pt

Abstract

This paper presents a structured approach for the recertification of ISO 9001 and ISO 14001 in a Portuguese thermoplastic's injection factory. Through an integrated management system, the company ensured continued compliance with quality and environmental standards. Key activities included document management, data-driven decision-making using the Primavera ERP (Enterprise Resource Planning), internal audits, (KPI) (Key Performance Indicator) monitoring, lean methodology implementation (5S), and calibration plans for monitoring resources (RMMs). The project showcases how digitalization and standardized practices can optimize operations and contribute to strategic sustainability goals.

Keywords: ISO 9001; ISO 14001; Quality Management; Environmental Management; 5S; KPIs.

1 Introduction

Certification in international management standards, such as ISO 9001 and ISO 14001, has become a fundamental requirement for companies aiming to ensure the quality of their products and services while promoting the sustainability of their operations. These globally recognized standards provide a systematic framework that fosters continuous improvement in organizational processes and ensures compliance with applicable legal and regulatory requirements.

ISO 9001, focused on quality management, establishes guidelines that enable organizations to enhance internal efficiency, optimize processes, and meet the expectations of customers and stakeholders. Meanwhile, ISO 14001 is dedicated to environmental management, driving practices that reduce environmental impacts and contribute to the sustainable development of industrial activities. Together, these standards form an Integrated Management System (IMS) that strengthens competitive capacity and reinforces the credibility of companies in the global marketplace.

This paper presents a case study on the recertification process for ISO 9001 and ISO 14001 in a Portuguese thermoplastics injection factory. This sector is characterized by high competitiveness and a constant demand for innovation and continuous improvement, making the implementation of robust and effective management practices even more critical.

The research carried out involves a comprehensive approach encompassing the updating of

documentation, the analysis of performance indicators, the implementation of Lean methodologies - particularly the 5S approach - and the rigorous control of measurement and monitoring resources (RMMs). The use of the Primavera ERP management system enabled integrated, data-driven decision-making, enhancing operational efficiency and the overall effectiveness of the processes.

The main objective of this study is to describe the practices adopted during the recertification process, highlight the results achieved, and discuss the challenges faced. It aims to provide valuable insights into the strategic role of ISO certifications as tools for organizational excellence in the industrial sector.

2 Theoretical Framework

The ISO 9001 and ISO 14001 standards are globally recognized as fundamental references for the implementation of quality and environmental management systems in industrial contexts (Ahmed A. et al., 2021) (Cabecinhas M. et al., 2021). ISO 9001 focuses on quality management by establishing requirements and guidelines that ensure the systematic control of processes, the identification of risks and opportunities, and the pursuit of customer satisfaction and continuous improvement (Pallawala N. et al., 2018). This standard utilizes the Plan-Do-Check-Act (PDCA) cycle as a structural basis, emphasizing leadership, evidence-based decision-making, and the engagement of employees to achieve organizational goals (Pallawala N. et al., 2018).

ISO 14001, on the other hand, provides a framework for establishing an Environmental Management System (EMS) aimed at reducing environmental impacts and ensuring compliance with legal obligations (Ahmed A. et al., 2021) (Stankov I., 2020). This standard also adopts the PDCA cycle, reinforcing the commitment to continuous improvement and proactive environmental management (Ahmed et al., 2021). Key components include the identification and evaluation of environmental aspects, the definition of measurable environmental objectives, and the implementation of operational controls to mitigate negative impacts (Stankov I., 2020).

The integration of these standards within a single management system — an Integrated Management System (IMS) allows organizations to address quality and environmental challenges simultaneously, fostering synergy between processes and reinforcing a culture of sustainability and excellence (Cabecinhas M. et al., 2021).

In addition to these standards, Lean methodologies, such as the 5S approach, play a crucial role in enhancing the efficiency of industrial operations (Beno et al., 2021). The 5S methodology — Seiri (Sort), Seiton (Set in Order), Seiso (Shine), Seiketsu (Standardize), and Shitsuke (Sustain) promotes workplace organization, waste elimination, and the establishment of standardized practices that contribute to operational excellence (Hurtado C. et al., 2023) (Beno J. et al., 2021).

Key Performance Indicators (KPIs) are also essential tools for monitoring and evaluating organizational performance (Gu M. et al., 2023). In an industrial environment, KPIs provide objective data that support strategic decision-making, enabling managers to identify strengths, weaknesses, and opportunities for continuous improvement (Ahasan M. et al., 2022).

Finally, the effective control and calibration of Measurement and Monitoring Resources (RMMs) is critical to ensuring the accuracy and reliability of measurements, which are fundamental for quality assurance and regulatory compliance (Volodarsky Y. et al., 2020).

3 Methodology and Results

The recertification process of ISO 9001 and ISO 14001 at the Portuguese thermoplastics injection factory was conducted through a structured and systematic approach, integrating various activities aimed at ensuring full compliance with the standards requirements and fostering continuous improvement.

The first step involved a comprehensive review and update of the Integrated Management

System (IMS) documentation. This included the revision of procedures, work instructions, and records to ensure alignment with current operational practices and regulatory obligations.

Simultaneously, an extensive data collection and statistical analysis were carried out to assess the performance of key processes and identify areas requiring corrective actions. The use of the Primavera ERP system was essential in this phase, providing a centralized platform for monitoring data and supporting decision-making based on real-time performance indicators.

Internal audits were planned and executed to verify the effectiveness of the IMS and ensure compliance with the standards. These audits also provided valuable insights into best practices and identified potential areas for improvement.

The implementation of Lean methodologies, particularly the 5S philosophy, played a crucial role in optimizing workplace organization and eliminating waste. This initiative involved the active engagement of employees in organizing workspaces, standardizing processes, and promoting a culture of discipline and continuous improvement.

Another key aspect of the recertification process was the calibration and control of Measurement and Monitoring Resources (RMMs). Regular verifications and calibrations were performed to guarantee the accuracy and reliability of instruments, which are essential for maintaining product quality and regulatory compliance.

Throughout the recertification process, customer and employee satisfaction assessments were conducted to ensure that stakeholder needs were met and to strengthen the company's commitment to continuous improvement.

The results of the external audits confirmed the effectiveness of these initiatives. The company successfully maintained its certifications in ISO 9001:2015 and ISO 14001:2015, demonstrating not only compliance with the standards but also the existence of a strong culture of quality and environmental management. Furthermore, the audits identified areas of best practice and suggested opportunities for further enhancement of the management system.

4 Conclusion

This case study highlights the strategic importance of ISO 9001 and ISO 14001 recertification for industrial organizations seeking to ensure compliance, enhance operational efficiency, and reinforce their market competitiveness. The comprehensive approach adopted by the thermoplastics injection factory demonstrated how an Integrated Management System (IMS), supported by Lean methodologies, data-driven decision-making, and robust calibration practices, can deliver significant improvements in both quality and environmental performance.

The successful recertification process validated the company's commitment to continuous improvement and sustainability, while also identifying opportunities for further enhancement in the future. Moreover, the active involvement of employees and the use of tools such as the Primavera ERP system and the 5S methodology were critical in fostering a culture of excellence and discipline.

This study underscores the value of combining internationally recognized standards with practical, data-supported methodologies to create a resilient and forward-looking management system. It also provides insights that can be adapted and applied by other organizations aiming to achieve similar objectives within the competitive landscape of industrial production.

References

Ahasan, M. R., Haque, M. S., & Alam, M. G. (2022). Supervised learning based mobile network anomaly detection from key performance indicator (KPI) data. *2022 IEEE Region 10 Symposium (TENSYMP)*, 1-6. https://doi.org/10.1109/tensymp54529.2022.9864371

Ahmed, A., & Mathrani, S. (2021). An integrated lean and ISO 14001 implementation for sustaining environmental performance in the meat industry. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1-6. https://doi.org/10.1109/csde53843.2021.9718371

Beno, J., Rao, M. V., Beno, J., & Das, S. K. (2021). Process control & Inspection using 5s method and computation with Pareto analysis. *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, 1-5. https://doi.org/10.1109/icac353642.2021.9697131

Cabecinhas, M., Sampaio, P., & Casadesus, M. (2021). A model to assess the impacts of ISO management systems standards. *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1208-1215. https://doi.org/10.1109/ieem50564.2021.9672900

Gu, M., Luo, L., & Li, P. (2023). Study and development of intelligent KPI management system for discrete manufacturing enterprises in industry 4.0. *2023 4th International Conference on Mechatronics Technology and Intelligent Manufacturing (ICMTIM)*, 19-25. https://doi.org/10.1109/icmtim58873.2023.10246680

Hurtado, C. V., Castro, O. J., De la Garza Inzunza, E. E., & Saravia, R. G. (2023). Applying the lean manufacturing tools 5S and 7 wastes for the learning and development of undergraduate students. *2023 IEEE Global Engineering Education Conference (EDUCON)*, 1-5. https://doi.org/10.1109/educon54358.2023.10125273

Pallawala, N. M., Jayamaha, N. P., & Grigg, N. P. (2018). Testing the ISO 9001:2015 process model: An Australasian empirical study. *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 580-585. https://doi.org/10.1109/ieem.2018.8607683

Stankov, I. (2020). Environmental management information systems. *2020 12th Electrical Engineering Faculty Conference (BulEF)*, 1-7. https://doi.org/10.1109/bulef51036.2020.9326021

Volodarsky, Y., Kosheva, L., & Pototskiy, I. (2020). Methods for determining the recalibration intervals of measuring equipment during their operation. *2020 XXX International Scientific Symposium 'Metrology and Metrology Assurance (MMA)*, 1-6. https://doi.org/10.1109/mma49863.2020.9254259

Chapter 37

Sentiment analysis of patient complaints in healthcare systems using VADER: Can it contribute to a better service?

Contribution #67
Cite as:
Liliana Barbosa, João Vasco Coelho. Sentiment analysis of patient complaints in healthcare systems using VADER: Can it contribute to a better service?. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-67,
  title ={Sentiment analysis of patient complaints in
     healthcare systems using VADER: Can it contribute
     to a better service?},
  author={Liliana Barbosa and  João Vasco Coelho},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Sentiment analysis of patient complaints in healthcare systems using VADER: Can it contribute to a better service?

Liliana da Costa Barbosa [1,*], João Vasco Coelho [2]

[1] Polytechnic University of Coimbra, Coimbra - Portugal
[2] C I E S / ISCTE - IUL, Lisbon – Portugal
*Corresponding author:
liliana.barbosa@live.com.pt

Abstract

This study focuses on exploring the effectiveness of using VADER (*Valence Aware Dictionary and sEntiment Reasoner*), a rule-based sentiment analysis tool, in swiftly analyzing patient complaints within healthcare settings aiming to develop overall digital capabilities. Currently, the use of sentiment analysis in the healthcare space is considered to be under-explored. Free-text complaints from patients often contain valuable insights into their experiences. However, analyzing large volumes of textual feedback can be time-consuming and prone to subjective interpretation. Sentiment analysis tools such as VADER can provide meaningful solutions by automatically extracting and quantifying the emotional tone of textual data.

The dataset used in the study comprised the total number and written free-text of patient complaints collected during 2024, in a key surgery service of a 10,000 employee public hospital located in Portugal. In terms of methods, Orange text mining capabilities were intertwined with VADER sentiment analysis. Three key study findings demonstrate the benefits of using VADER to enhance healthcare patient experience management: (a) its ability to balance efficiency and accuracy in sentiment analysis without compromising precision, (b) its role in fostering a culture of *patient-centered* care decision-making, and (c) its support for process optimization and digitalization efforts, namely in terms of case triage and prioritization.

Keywords: Natural language processing (NLP); Sentiment analysis; Text classification; VADER sentiment analysis; BERT analysis; Healthcare; Patient experience; Patient complaints.

1 Introduction

Patient and family complaints are a valuable source of information to provide patient-oriented healthcare. Complaints made by patients and families typically express a request or dissatisfaction regarding health services, and are considered a valuable data source for hospitals to improve the quality of health services [1][2][3].

As healthcare progressively changes from *disease-centered* to *patient-centered*, it is expected that free-text data will be increasingly important as means to understand patient experiences, that can be somehow lost by relying in the more frequently used surveys based in closed-ended questions [2][3]. To support this intent and develop patient-oriented healthcare services, systems and processes need to be improved to collect, organize, analyze and quickly address patient and family complaints, and establish an accurate *patient's voice*, which is proposed to be one of the core building blocks for system building in contemporary complaint management in the healthcare sector [1][2][3].

To extract value from rich free-text data regarding patient experience, establishing trust by devising quicker case resolution loops, is considered to be of key relevance [4][5][6][7][8][9]. Regarding this intent, it

is appropriate to say that the scale and diversity of free-text feedback make it difficult for human agents to swiftly extract and digest actionable information. Sentiment analysis, a pivotal technique in natural language processing (NLP) that involves identifying and extracting subjective information from textual data, can help automate this process by classifying textual feedback into emotional categories such as positive, neutral, or negative [2]. Beyond basic polarity detection, advanced sentiment models can also detect intensity and emotional cues, offering a nuanced understanding of customer grievances [5][6][7][8][9].

The *Valence Aware Dictionary and sEntiment Reasoner* (VADER), a rule-based sentiment analysis tool, is considered to be a reliable solution in this domain [2]. It was initially designed to perform on social media text, but is increasingly being tested and valued as effective in other domains [10][11][12][13]. VADER combines a robust lexicon with heuristic rules to capture contextual nuances in sentiment, making it easy to use and highly accurate [13][14][15][16].

In a context (healthcare) where the use of sentiment analysis, namely in terms of patient experience management support, is considered to be still significantly under-explored [13], this paper aims to explore the extent to which rule-based sentiment analysis tools such as VADER can contribute to more efficient, swift and fair patient complaints resolution by quickly establishing an accurate *patient's voice* [1][2][3], in complex and critical sectors such as healthcare and large-sized organizations aiming to develop their technological capabilities such as public hospitals.

The study's dataset consists of the total number and content of patient complaints (n=63) collected throughout 2024, in one of the key surgery services of a 10,000 employee public hospital located in Portugal. In terms of research methods, the analytical apparatus was based in intertwining the usage of Orange text mining capabilities with its VADER sentiment analysis widget, and in comparing outputs, for accuracy control, with the ones derived a grouped sentiment analysis using BERT (*Bidirectional Encoder Representations from Transformers*), a transformer-based model. The paper is organized as follows: *Section 2* presents a review of the current state of the art, *section 3* presents the dataset and methods that were used, *sections 4* and *5* the research results and respective discussion, including observed challenges and limitations. A conclusion section presenting a vision for future work ends the paper.

2 State of the art

Over the last decade, sentiment analysis, conversational AI and opinion mining have been emerging as critical areas of research, with significant applications across industries such as e-commerce, healthcare, social media, and customer service [2][13][14][15][16]. During this timeframe, sentiment analysis methods, in concrete, have developed significantly, driven by advances in natural language processing (NLP) and machine learning techniques. Current primary methodologies in the sentiment analysis field include [10][11][12][13][14][15][16]:

1. *Rule-based, lexicon-based* approaches that rely on predefined linguistic rules to identify and extract sentiment from text; these approaches typically involve using predefined dictionaries of sentiment-bearing words to determine the sentiment orientation of text; each word in the lexicon is assigned a sentiment score, and the overall sentiment of a text is calculated based on the presence and scores of these words. While these methods are simple and easy to implement, they often struggle with context, sarcasm, and nuanced expressions. Despite the limitations, these methods are still widely used due to their interpretability and low computational requirements;
2. *Machine learning* models trained on labeled datasets, learning patterns in the data to predict sentiments in new, unseen text; these techniques offer greater flexibility and accuracy compared to rule-based approaches;
3. *Deep learning techniques*, such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Transformer-based models (e.g., BERT) have demonstrated superior performance in sentiment analysis, by processing complex representations of text that enhance sentiment classification accuracy.

Closely intertwined with the aforementioned developments, sentiment analysis methods have become increasingly explored and applied in industries and domains where understanding customer and user sentiments can drive more effective managerial decision-making and improve overall experiences [4] [5][6][7][8][9]. For instance, businesses have been leveraging sentiment analysis to gain insights into customer feedback, enhancing product offerings and service quality [13]. In healthcare, sentiment analysis has been explored to aid in monitoring patient opinions and emotions, contributing to better patient care and satisfaction [9][16].

As healthcare shifts from a *disease-centered* to a *patient-centered approach*, free-text data is expected to play an increasingly vital role in capturing and promptly handling patient experiences [2][3]. To derive meaningful insights from rich free-text data on patient experiences, build trust and establish faster resolution loops, it is essential to be able to systematize complaints and complaint data related to health services [4] [5][6][7][8][9]. To advance this goal and foster *patient-oriented* healthcare services, systems and processes must be enhanced to efficiently organize and address patient and family complaints. Establishing an accurate *patient voice* is proposed as a fundamental building block in modern complaint management within the healthcare sector. As illustrated by previous studies [3][4][9][13][16], it is frequent to find ambiguity and nuances in free text patient complaints, making it difficult to identify appropriate improvement actions. Complaints caused by hospital staff, as opposed to processes or environment (e.g., doctors, nurses, and other staff) are often more frequent and more reported as complaints triggered by a range of events or attributes (e.g., communication, attitude, nursing skill), with direct comparisons of complaint classification between hospitals and countries being difficult to do [3][4][9].

Despite the recent observed advancements in the healthcare space, sentiment analysis, particularly in patient experience management through free-text data mining, remains significantly under-explored [13]. However, it continues to gain traction as healthcare organizations strive to enhance and optimization the services provided to patients and their families, in particular through process digitalization and automation. The present study aims to contribute to this progress.

3 Methodology

Materials.

The dataset that was used in the study consisted of the total number and content of patient complaints (n=63) collected during 2024, in one of the key healthcare services of a 10,000 employee public hospital located in Portugal. Given its size, nature (general surgery) and exposure to be subject considered in the research (patient experience management in healthcare organizations with developing digital capabilities), this healthcare service was positioned as an *exemplary*, pregnant case to address, in empirical terms, the research proposed intents. The reliance of the service on free-text survey data to handle patient complaints was a key factor supporting its suitability to address the research objectives in empirical terms.

Methods.

Data was translated from Portuguese to English, labelled and processed to establish a base *corpus* and perform subsequent sentiment analysis. *Figure 1* presents the dataset labelling per feedback type. Analysis was made by intertwining the usage of Orange text mining capabilities with its VADER sentiment analysis widget. In Orange, VADER is used to analyze English-language text, providing positive, negative, neutral and compound sentiment scores. These scores help users understand the overall sentiment of the established *corpus*. For secondary comparative purposes (in terms of compound sentiment analysis scores accuracy), a grouped sentiment analysis using BERT (used for nuanced and complex sentence structures analysis) [17] was made. The text mining model that was used is presented in *Figure 2*.

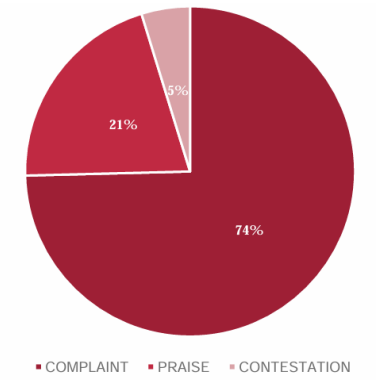


Figure 1. Dataset labelling (per feedback type).

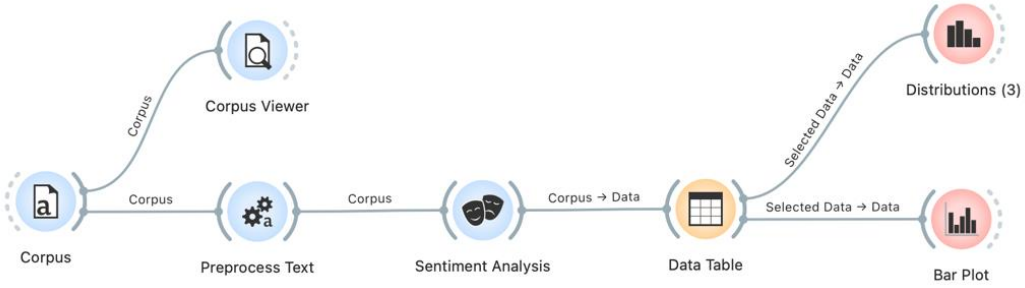


Figure 2. Text mining model (using Orange).

Composite scores were obtained by summing the valence scores of each word in the lexicon, including adjustments based on predefined rules, and normalizing the score to a range between -1 (extremely negative) and +1 (extremely positive). In the classification process, the following criteria was used to determine the sentiment (positive, negative, or neutral) of each comment [14]: *positive sentiment*: compound score >= 0.5; *neutral sentiment*: (compound score > -0.5) and (compound score < 0.5); *negative sentiment*: compound score <= -0.5.

In terms of analytical results presentation, given the research goal of exploring the potential usefulness or appropriateness of using a sentiment analysis technique and tool as means to digest patient complaints in healthcare settings and support quick resolution actions and/or decisions by key stakeholders, it was decided to split analytical outputs per emotional tone - *negative* sentiments vs. *neutral* or *positive* -, regardless the nature of the feedback being presented by the patients (e.g., a complaint, a compliment or a dispute).

4 Results

Overall, three main research results should be highlighted: *1)* the existence of a significant number of neutral and positive scores in complaints free-text suggesting that, although labelled as complaints, the tone isn't necessarily (strongly) emotional; *2)* where an overall positive emotional tone was observed (compound score), text labelling was mixed (complaint, praise and/or disputes); *3)* to be able to identify and differentiate these nuances in a timely manner can be particularly useful for users and decision-makers to quickly trigger (or not) follow-up case resolution actions.

In a more qualitative perspective, complementing VADER's sentiment analysis output, it was possible to summarize the most frequent content feedback and link it with negative toned feedback (e.g., misdiagnosis and incorrect surgery scheduling, post-surgery pain and neglect, poor communication/ updates), neutral feedback (e.g., constructive suggestions of service improvements) and positive feedback (e.g., gratitude for the professionalism and care).

Figure 3 presents a sample of the output that was obtained with VADER that had a negative tone, *Figure 4* a output sample where a neutral or a positive tone was observed. In *Figure 3*, that aggregates negative tone cases, most texts were labelled as complaint, with one case of dispute. Compound scores range from -0.9984 (extremely negative) to -0.0323 (slightly negative), indicating VADER's ability to differentiate the intensity of the sentiment expressed in a complaint. Most complaints have negative compound scores (check *Figure 5*), reflecting predominantly negative sentiments. The coexistence of neutral and positive scores in complaint texts suggests that, although they are classified as complaints, the tone isn't necessarily (strongly) emotional.

	Text	Label	positive	negative	neutral	compound
15	**COMPLAI...	Complaint	0.038	0.125	0.838	-0.9984
23	**Subject:** ...	Complaint	0.058	0.134	0.808	-0.996
33	**Subject:** ...	Dispute	0.076	0.167	0.757	-0.9948
7	Good Mornin...	Complaint	0.075	0.129	0.796	-0.9921
49	Formal Com...	Complaint	0.056	0.162	0.783	-0.9912
16	To Whom It ...	Complaint	0.05	0.189	0.761	-0.9904
59	**Outeiro Ma...	Complaint	0.064	0.117	0.819	-0.9863
10	**NCC. 998...	Complaint	0.049	0.16	0.791	-0.9861
4	Suspected M...	Complaint	0.063	0.102	0.835	-0.9857
12	**To Whom I...	Complaint	0.098	0.129	0.773	-0.9842
17	Urgent Com...	Complaint	0.061	0.11	0.83	-0.9784
31	**Subject:** ...	Complaint	0.071	0.082	0.847	-0.97
40	Dear Sir/Mad...	Complaint	0.122	0.137	0.741	-0.9692
45	Hello, good a...	Complaint	0.033	0.067	0.901	-0.9682
32	**Subject:** ...	Complaint	0.079	0.119	0.802	-0.9584
2	I underwent ...	Complaint	0.042	0.158	0.8	-0.9555
36	Good aftern...	Complaint	0.062	0.093	0.845	-0.9555
1	I, Maria, a pa...	Complaint	0.055	0.116	0.829	-0.9544
39	**Case Num...	Complaint	0.016	0.049	0.935	-0.9523
26	**Subject:** ...	Complaint	0.049	0.087	0.864	-0.9458
43	Dear Sir/Mad...	Complaint	0.051	0.071	0.879	-0.8373
57	Dr. José Guil...	Complaint	0.084	0.097	0.819	-0.8305
37	I've been wai...	Complaint	0	0.169	0.831	-0.7579
51	Dear Sirs,I a...	Complaint	0.045	0.06	0.895	-0.7556
3	My complain...	Complaint	0	0.067	0.933	-0.7227
8	Good aftern...	Complaint	0.113	0.127	0.759	-0.712
25	The complai...	Complaint	0	0.113	0.887	-0.6705
38	I went to the ...	Complaint	0.034	0.092	0.874	-0.6483
44	I've been wai...	Complaint	0	0.065	0.935	-0.3506
22	**Subject:** ...	Complaint	0.075	0.076	0.849	-0.0451
9	Dear HUC, ...	Complaint	0.069	0.086	0.846	-0.0323

Figure 3. VADER output data table with negative emotional tone (compound score).

In *Figure 4*, despite overall positively toned answers, text labelling is mixed - complaint, praise or disputes were observed. Compound scores range from 0 (neutral) to 0.9994 (extremely positive), indicating the intensity of the sentiment. The negative scores are generally low, with most values close to 0, indicating minimal negative sentiment. Neutral scores are high (ranging from 0.577 to 1), suggesting that many texts have a neutral tone. Some of these cases likely do not require the immediate direct intervention of a healthcare service member.

	Text	Label	positive	negative	neutral	compound
63	**Good morn...	Praise	0.264	0.046	0.69	0.9994
19	**Subject:** ...	Praise	0.173	0.027	0.8	0.9981
54	Good evenin...	Praise	0.289	0	0.711	0.9961
35	Dear Sirs,I a...	Praise	0.282	0.012	0.705	0.9955
56	**Dear Sirs,*...	Complaint	0.115	0.074	0.811	0.9938
6	Subject: Con...	Praise	0.237	0	0.763	0.9924
30	Luís Joaquim...	Praise	0.311	0	0.689	0.9869
53	Dear Sir,Hea...	Praise	0.423	0	0.577	0.9863
34	**Subject:** ...	Dispute	0.106	0.04	0.854	0.9858
55	Dear Profess...	Praise	0.28	0	0.72	0.9773
42	I would like t...	Praise	0.416	0	0.584	0.976
58	**To the Citiz...	Complaint	0.092	0.014	0.893	0.9698
52	Dear Dr. Alex...	Praise	0.217	0.022	0.761	0.9595
61	To the entire ...	Praise	0.231	0.04	0.728	0.9504
62	Dear Profess...	Praise	0.394	0	0.606	0.9423
13	Good aftern...	Complaint	0.125	0	0.875	0.936
5	Good mornin...	Complaint	0.229	0.037	0.734	0.9235
20	**Subject:** ...	Complaint	0.106	0.013	0.881	0.9197
21	**Subject:** ...	Complaint	0.161	0	0.839	0.9125
28	Request for ...	Complaint	0.141	0	0.859	0.8916
48	Exmo. Sirs,I ...	Dispute	0.068	0.054	0.878	0.8799
60	I would like t...	Praise	0.234	0	0.766	0.8658
27	Complaint R...	Complaint	0.138	0.085	0.777	0.7707
18	Urgent Com...	Complaint	0.126	0.1	0.775	0.7523
46	Dear Sir/Mad...	Complaint	0.061	0.044	0.895	0.743
11	Whom It May...	Complaint	0.103	0.081	0.816	0.7261
50	Good mornin...	Complaint	0.095	0.064	0.841	0.6871
14	**To the Dire...	Complaint	0.121	0.099	0.781	0.5759
24	**Subject:** ...	Complaint	0.046	0.028	0.926	0.5565
41	To the Estee...	Complaint	0.052	0.04	0.908	0.4776
47	Good mornin...	Complaint	0.064	0.051	0.885	0.2732
29	They called ...	Complaint	0	0	1	0

Figure 4. VADER output data table with a neutral or positive emotional tone (compound score).

Figure 5 presents the compound scores obtained with the use of VADER: as mentioned, the existence of a significant number of neutral or positive scores for texts labelled as complaints, suggests that, despite the classification, the tone isn't necessarily (strongly) emotional.

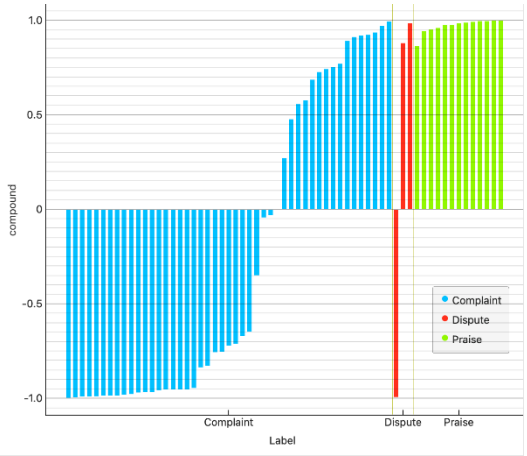


Figure 5. VADER output data table – compound scores.

For analytical control and output comparison purposes, BERT was used to double-check the accuracy of the results achieved with VADER. BERT grouped results are presented in *Figure 6*.

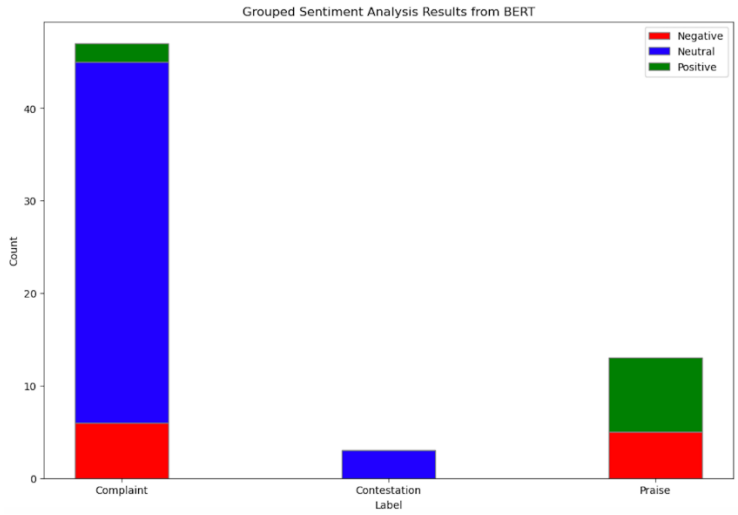


Figure 6. BERT output data table – grouped sentiment analysis.

Overall, as illustrated in the literature [2], the overall orientation of VADER and BERT results can be considered to be similar, namely the existence of a significant number of neutral and positive scores in complaints free-text. VADER analysis, given its rule-based nature, is more straightforward in terms of legibility and call-to-action, but the chance of missing subtle sentiments or contextual-dependent details exists (e.g., a sarcastic comment can classified as positive; overestimation of neutral or positive sentiments in ambiguous cases can be observed). BERT results are likely to be more accurate, especially for nuanced or sarcastic text. A sarcastic comment classified as neutral by VADER is likely to be classified as negative by BERT, thus providing approach to analytical accuracy [18].

5 Discussion

As illustrated by the present research, the use of sentiment analysis techniques and tools such as VADER to support patient experience management in healthcare settings with developing digital capabilities seems to present significant potential to contribute to transform free-text patient feedback into actionable, prompt insights. As illustrated, when analyzing informal free-text and requiring quick results, VADER is able to provide overall meaningful output, namely if broader, helicopter view or snapshot perspectives are required.

In terms of research results, the research provides an illustration of how VADER sentiment analysis can support the extraction of meaningful insights from rich free-text data on patient experiences, offering a view of customer grievances. As illustrated by past studies [4] [5][6][7][8][9], VADER human-centric design enhances sentiment understanding, making it a potentially valuable tool in healthcare-related analytics, going beyond basic polarity detection, by detecting the intensity of emotional cues. This makes it particularly useful for applications and domains where interpretability and speed are crucial, as is the case of patience experience management and the growing need to craft an accurate *patient's voice* in healthcare settings [1][2][3].

Beyond reinforcing previous studies findings [2][3][13][15][16], it is worthwhile highlighting, in particular, three concrete aspects provided by the present study that suggest the appropriateness of deepening and expanding the usage of VADER ito the patience experience management space:

1. *Efficiency – effectiveness (accuracy) balance*: VADER's rule-based approach offers a quick and computationally efficient method for sentiment analysis without sacrificing accuracy. This is particularly beneficial in healthcare settings and patient experience handling where timely, easy-to-understand insights are of paramount importance;
2. *Support to patient-centric healthcare services*: by analyzing patient complaints, healthcare providers

can gain a deeper understanding of patient experiences and identify areas needing improvement. This aligns with the broader trend of *patient-centered* care, which emphasizes the importance of considering patient feedback in healthcare service delivery;

3. *Process optimization support (e.g., case prioritization)*: the sentiment analysis made in the study revealed that most complaints had a negative tone, pinpointing concrete areas where patients frequently encounter problems; this can help healthcare providers prioritize these issues and allocate resources more effectively.

While VADER-based sentiment analysis may offer valuable, actionable insights to patient feedback handling, three challenges and limitations need to be considered:

1. *Dealing with nuances and ambiguity*: Sentiment analysis models tend to struggle with ambiguous expressions where the sentiment is not clear-cut. This can lead to misclassification and reduced accuracy. In the present study, the existence of a significant number of neutral and positive scores in complaints free-text signals possible advantages in using a deeper integration of different models in future work, hybridizing traditional techniques with deep learning models. Developing model capabilities to understand nuanced opinions and subtle sentiments remains a significant challenge;
2. *Unstructured data and domain adaptation*: Models trained on specific datasets often perform poorly when applied to different domains. To this regard, in the present study, a significant challenge was dealing with unstructured data, often containing noise, slang, abbreviations, and inconsistent use of language. Effective preprocessing methods, such as tokenization, stemming, lemmatization, and noise removal, are essential to enhance the quality of the data and improve analysis outcomes;
3. *Contextual dependency (for interpretation)*: Understanding the context in which a sentiment is expressed is crucial for accurate analysis. In the present study, following existing literature [17][18], the use of BERT, with comparative purposes, highlighted the relevance of considering the surrounding context to correctly interpret sentiments, especially in complex sentences.

Continuous advancements in sentiment analysis require maintaining a laser focus on developing innovative approaches to address existing gaps and challenges. As mentioned, future work will include a deeper integration of different models, hybridizing traditional techniques with deep learning models, a way to balance the limitations of each approach. For example, rule or lexicon-based methods can be used to provide initial sentiment scores, subsequently refined using deep learning models. This hybrid approach will help improve accuracy by incorporating both rule-based knowledge and pattern recognition capabilities of deep learning models. Additionally, integrating syntactic and semantic features extracted from traditional methods with embeddings learned by deep learning models can enhance the overall performance of sentiment analysis systems, namely to better understand and detect textual subtle and context-dependent nuances.

6 Conclusion

The present study illustrates the effectiveness and potential usefulness of using VADER sentiment analysis to analyze patient complaints in healthcare settings. Three key study findings suggest the advantages of using VADER to support healthcare patient experience management: a) the efficiency-effectiveness (accuracy) balance offered by VADER, by delivering quick and computationally efficient sentiment analysis without sacrificing accuracy; b) support given to patient-centered care; and c) support given to process optimization intents (via, for example, sentiment data-infused case prioritization). Future work will focus on integrating advanced machine learning models to enhance the accuracy and depth of sentiment analysis, as demonstrated by the comparison with the BERT sentiment analysis model. Additionally, expanding the dataset to include cross-sectional feedback from multiple departments and hospitals will provide a broader and more comprehensive perspective on patient sentiments across healthcare systems, helping to identify recurring issues and prioritize areas for improvement.

Acknowledgments

The authors would like to thank the healthcare organization included in the study for providing access to the data used in this research.

References

[1] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," Expert Syst Appl, vol. 162, 2020, doi: 10.1016/j.eswa.2020.113746.

[2] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014, 2014. doi: 10.1609/icwsm.v8i1.14550.

[3] R. Harrison, M. Walton, J. Healy, J. Smith-Merry, and C. Hobbs, "Patient complaints about hospital services: Applying a complaint taxonomy to analyse and respond to complaints," International Journal for Quality in Health Care, vol. 28, no. 2, 2016, doi: 10.1093/intqhc/mzw003.

[4] C. D. Mann, J. A. Howes, A. Buchanan, and D. J. Bowrey, "One-year audit of complaints made against a University Hospital Surgical Department," ANZ J Surg, vol. 82, no. 10, 2012, doi: 10.1111/j.1445-2197.2012.06240.x.

[5] L. Donaldson, "An organisation with memory," Clin Med (Lond), vol. 2, no. 5, 2002, doi: 10.7861/clinmedicine.2-5-452.

[6] S. Lloyd-Bostock and L. Mulcahy, "The Social Psychology of Making and Responding to Hospital Complaints: An Account Model of Complaint Processes," Law Policy, vol. 16, no. 2, 1994, doi: 10.1111/j.1467-9930.1994.tb00120.x.

[7] L. Mulcahy and J. Q. Tritter, "Pathways, pyramids and icebergs? Mapping the links between dissatisfaction and complaints," Sociol Health Illn, vol. 20, no. 6, pp. 825–847, 1998, doi: 10.1111/1467-9566.00131.

[8] S. Carmel, "Patient complaint strategies in a general hospital," Hosp Health Serv Adm, vol. 35, no. 2, 1990.

[9] T. W. Reader, A. Gillespie, and J. Roberts, "Patient complaints in healthcare systems: A systematic review and coding taxonomy," 2014. doi: 10.1136/bmjqs-2013-002437.

[10] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in Conference on Empirical Methods in Natural Language Processing, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:990233

[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," Proceedings of the Workshop on Languages in Social Media, Jan. 2011.

[12] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, A System for Real-time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle. 2012.

[13] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," Journal of King Saud University - Computer and Information Sciences, vol. 36, no.

4, p. 102048, 2024, doi: https://doi.org/10.1016/j.jksuci.2024.102048.

[14] M. Hoti and J. Ajdari, “Sentiment Analysis Using the Vader Model for Assessing Company Services Based on Posts on Social Media,” SEEU Review, vol. 18, pp. 19–33, Dec. 2023, doi: 10.2478/seeur-2023-0043.

[15] D. Youvan, Understanding Sentiment Analysis with VADER: A Comprehensive Overview and Application. 2024. doi: 10.13140/RG.2.2.33567.98726.

[16] G. Al Qarana, L. Rianto, and L. Purnomo,"Analyzing Perceptions of Maternal and Pediatric Care in Jakarta: An Integrated VADER and GloVe Analysis of Google Reviews in Mother and Child Hospitals.” [Online]. Available: www.joiv.org/index.php/joiv

[17] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 2019.

[18] A. de L. Languré and M. Zareei, “Improving Text Emotion Detection Through Comprehensive Dataset Quality Analysis,” IEEE Access, vol. 12, pp. 166512–166536, 2024, doi: 10.1109/ACCESS.2024.3491856.

Chapter 38

Case Study - Failure analysis of a raw mill in the cement industry: FMECA with fault tree and Ishikawa diagram

Contribution #71
Cite as:
Paulo Duarte, Hugo Raposo, José Farinha. Case Study - Failure analysis of a raw mill in the cement industry: FMECA with fault tree and Ishikawa diagram. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-71,
  title ={Case Study - Failure analysis of a raw mill
     in the cement industry: FMECA with fault tree and
     Ishikawa diagram},
  author={Paulo Duarte and  Hugo Raposo and  José
     Farinha},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Case Study - Failure analysis of a raw mill in the cement industry: FMECA with fault tree and Ishikawa diagram

Paulo Duarte[,1], Hugo Raposo[1,2], Torres Farinha[1,2]

[1] Polytechnic University of Coimbra, Instituto Superior de Engenharia de Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.
[2] RCM[2+] Research Centre in Asset Management and System Engineering, Polytechnic University of Coimbra, Coimbra, 3030-199 Coimbra, Portugal

Corresponding authors
paulo.j.duarte98@gmail.com; hugo.raposo@isec.pt; tfarinha@isec.pt

Abstract

The cement industry faces challenges related to the need to increase efficiency, ensure high equipment availability, reduce costs and reduce environmental impact. In this context, asset management plays a key role in promoting continuous improvement, ensuring the continuity of the production process and the sustainability of operations

In the cement sector, physical asset management is particularly demanding due to the size, complexity and criticality of equipment such as kilns and mills. The raw mill, responsible for grinding the raw materials that feed the clinker kiln, stands out as one of the most critical assets. Failures in this equipment can result in long periods of downtime, poor product quality and increased operating costs.

In this study presents an FMECA (Failure Modes, Effects and Criticality Analysis) analysis of the raw mill (vertical roller mill) in a cement production center. The methodology made it possible to identify the main failure modes, assess their effects and determine criticality based on occurrence, severity and detectability, resulting in the calculation of the Risk Priority Number (RPN). Complementary tools such as Fault Tree Analysis and the Ishikawa Diagram were also used for a more detailed analysis, with the aim of improving the analysis of the causes and effects of the failure modes found.

This analysis proved essential in identifying the most critical failure modes, serving as a basis for future reformulation of maintenance strategies. It is expected that, in the medium term, the implementation of these measures will contribute to greater equipment reliability and availability. On the other hand, the study showed that the current failure classification system is too simplistic, and it is recommended that it be revised to improve the accuracy of the analysis and the effectiveness of the actions implemented

Keywords: Cement Industry, Raw Mill, Physical Asset Management, FMECA, Fault Tree Analysis, Ishikawa Diagram

1 Introduction

The increasing complexity of industrial processes, combined with the pressure for efficiency, sustainability, and operational continuity, has highlighted the importance of physical asset management, which plays a fundamental role in modern organizations. The goal of asset management is to maximize value throughout the asset's lifecycle, balancing performance, cost and risk. In the context of digital transformation and Industry 4.0, physical asset management has been adapting to a rapidly evolving

technological landscape. The introduction of tools such as the Internet of Things (IoT), Artificial Intelligence (AI), and real-time data analytics is transforming the way assets are monitored and managed.

In the cement sector, due to the complexity and scale of the equipment, the reliability and availability of assets are essential to ensure production continuity and reduce the costs associated with unplanned downtime. One of the most critical pieces of equipment in this context is the raw mill, responsible for grinding the raw materials required for clinker production. Failures in this system can compromise operations and generate significant financial and productivity-related impacts.

This article presents a case study conducted at a cement plant, focusing on the analysis of the raw mill. The study began with the identification of recurrent failure modes in this stage of the process and aimed to apply reliability analysis and improvement methodologies. A Reliability Block Diagram (RBD) was developed, followed by a Failure Modes, Effects and Criticality Analysis (FMECA), and complemented by the use of an Ishikawa Diagram and a Fault Tree Analysis (FTA).

This article is structured as follows: Section 2 presents the state of the art; Section 3 describes the methodologies applied and the case study; Section 4 presents the results obtained; Section 5 discusses the findings in comparison with the literature; and, finally, Section 6 outlines the conclusions drawn from the study.

2 State of the art

The evolution of industrial maintenance reflects the growing need for increased reliability, availability, safety, and operational efficiency. Historically, maintenance was performed solely in a corrective manner, meaning after the occurrence of a failure. However, with economic and technological development, new approaches have emerged, giving rise to preventive and predictive maintenance strategies. [1]

According to the standard NP EN 13306:2021, maintenance is divided into three categories: preventive maintenance, corrective maintenance, and improvement maintenance. Corrective maintenance is carried out after a failure has occurred, while preventive maintenance involves scheduled interventions to prevent future failures. Within the scope of preventive maintenance, predictive maintenance stands out; this approach uses sensors and real-time data analysis to detect changes in operating patterns, thereby enabling the prevention of failures. Finally, improvement maintenance aims to increase the reliability of equipment or systems through actions designed to optimize their performance. [2]

In this article, four methodologies were used: Reliability Block Diagram (RBD), Failure Modes, Effects and Criticality Analysis (FMECA), Ishikawa Diagram, and Fault Tree Analysis (FTA).

The Reliability Block Diagram (RBD) is a methodology used to logically structure the operation of equipment or a system. The IEC 61078:2016 standard establishes the requirements for graphical representation, application, and interpretation of reliability diagrams. The standard also specifies methods for estimating reliability, availability, and the impact of possible redundancies or sequences within the system.

This methodology is widely used for identifying the absence of redundancies and understanding the effects of failures in series systems. The reliability of a system in series can be calculated using Equation 1:

$$R(t) = \prod R_i(t) \quad (1)$$

Where Ri(t), R_i(t) and Ri(t) represents the reliability of component *i* over time.

In parallel systems, reliability is calculated using Equation 2:

$$R(t) = \prod[1 - R_{i(t)})] \quad (2)$$

FMEA (Failure Mode and Effect Analysis) is an inductive and systematic methodology used in risk management, as well as in the improvement of products, processes and systems. Over time, this methodology has been adopted by the industrial sector across various fields. FMEA enables the identification of potential failure modes and the analysis of their respective causes and effects, so that preventive actions can be taken before failures occur. This method is widely recognized as an engineering technique aimed at anticipating, preventing, and eliminating known or potential failures before they reach the end user. [3], [4].

FMEA analysis can take two forms: a qualitative analysis, typically known as FMEA (Failure Mode and Effect Analysis), or a more detailed and quantitative approach called FMECA (Failure Modes, Effects and Criticality Analysis), in which the criticality of the identified failure modes is assessed. [5]

Criticality is assessed through the application of the Risk Priority Number (RPN), which is calculated using the Equation 3:

$$RPN = S \times O \times D \quad (3)$$

Where S is the Severity of the failure effect, O is the Occurrence of the failure mode and D is the Detectability of the failure. [5]

Despite its usefulness, FMEA/FMECA has limitations. It depends on the quality of the available information, the experience of the analysts, and can become subjective in the assignment of ratings. Moreover, the RPN may lead to incorrect interpretations if the factors S, O, and D are poorly weighted. Therefore, in more rigorous contexts, it can be complemented with other methodologies, such as Fault Tree Analysis (FTA) or Root Cause Analysis (RCA). [3], [5], [6]

FMEA or FMECA analyses are conducted according to the IEC 60812:2018 standard. This analysis is one of the fundamental tools for early failure identification, enabling a proactive response from engineering and maintenance teams, and significantly contributing to the reliability, safety, and efficiency of industrial assets. [5]

The Ishikawa diagram, also known as the fishbone diagram, is a cause-and-effect analysis tool used to identify, organize, and visually represent the possible causes of a problem, facilitating analysis in more complex issues. By clearly structuring the potential causes of the problem, it enables organizations to develop maintenance strategies—whether corrective, preventive, or even predictive—thereby improving the reliability and efficiency of assets. [7]

The development of the analysis using the Ishikawa diagram involves the correct identification of the problem, which is positioned on the right side of the diagram. From this point, the main lines represent the primary causes (Method, Manpower, Material, Measurement, Environment, Machine). Smaller lines branching from these are used to record the possible secondary causes grouped under each primary cause. [7]

Fault Tree Analysis (FTA) is a deductive methodology used for identifying, structuring, and evaluating the causes that may lead to a specific undesired event. This technique involves representing in a hierarchical and logical manner the sequence of failures or events that contribute to a critical equipment failure. Due to its ability to break down complex problems into basic causes, fault tree analysis is a fundamental tool in asset management, safety, and industrial maintenance, being particularly effective in analyzing systems with multiple interdependencies. [8]

According to the IEC 61025:2006 standard, fault tree analysis consists of the logical decomposition of the top event into simpler events until reaching the root causes or basic events. This international standard defines the principles, symbols, and procedures to be followed for constructing and analyzing fault trees, promoting standardization and reliability of the results obtained. For this purpose, the standard recommends the use of real data whenever possible (e.g., failure histories) to ensure greater reliability in the quantification of events. [9]

The causes are then organized into a tree diagram where, through logical gates—AND and OR gates—they express the relationships between the different events. [9]

OR logic gates represent situations where the output event occurs if at least one of the input events occurs; whereas the AND logic gate represents a situation where the output event only occurs if all the input events occur simultaneously. [9]

This logical structure also allows for the evaluation of failure probability based on component reliability data, failure histories, or expert estimates. Its practical application is highly valued in sectors where reliability analysis and risk minimization are critical, such as aerospace, energy, and industrial processes. [10]

For probability calculation Equations 4 and 5 are used, as stipulated in the IEC 61025:2006 standard:

Equation 4- OR gate:

$$F(t) = 1 - \prod_{i=1}^{n}[1 - R(t)] \qquad (4)$$

Equation 5- AND gate:

$$F(t) = \prod_{i=1}^{n}[1 - R(t)] \qquad (5)$$

Where *F(t)* is the probability of occurrence of the event, and *R(t)* is the reliability of the element.

However, despite being a powerful and widely used tool, Fault Tree Analysis (FTA) is not without limitations. In practice, it is not always easy to gather reliable and sufficient data to perform the analysis, especially when dealing with complex systems or in environments dominated by uncertainty and empirical knowledge.

To overcome these challenges, there has been an evolution in the methodology through the integration of fault tree analysis with other approaches, such as fuzzy logic, Bayesian networks, or even Reliability, Availability, Maintainability & Safety (RAMS) analysis, among others. As demonstrated by Soltanali *et al*., "The integrated methodology improves the predictive capability of failure models and simultaneously optimizes maintenance decisions based on real data and the tacit knowledge of operators." [11]

3 Methodology

The raw mill is one of the fundamental pieces of equipment in the cement manufacturing process. Its function is to grind the raw materials used for clinker production. Raw material preparation is an essential step in cement production, aiming to obtain a homogeneous mixture of raw materials with a chemical composition suitable for clinker formation. This mixture mainly consists of limestone and clay but may also include corrective materials or industrial by-products to adjust the final composition.

The operation of vertical mills is based on the principle of compression grinding through a horizontal rotating table and a set of vertical rollers subjected to constant pressure via hydraulic cylinders. The raw material is fed into the center of the table, where, due to the centrifugal force generated by the table rotation, it is conveyed towards the edges of the table. As the material passes under the rollers, it is ground

into fine particles. Simultaneously, hot air is blown into the bottom of the mill and travels through the interior. This hot air serves two functions: drying the material if moisture is present and transporting the ground particles to the upper part of the mill where the separator is located. The separator's function is to classify the material—the ground particles with the correct size pass through the separator following the gas flow to the next stage of the process, while the larger particles are returned via the rejection cone to be ground again.

The analysis conducted in this study was based on the failure history recorded between January 2022 and April 2025. The main objective was to identify, understand, and characterize the failure modes that impact the performance and availability of this equipment.

The study was structured in four stages, each with a distinct but interconnected approach, allowing for a detailed analysis of this equipment.

In the first stage, a block diagram was developed representing the main components that constitute the raw mill. This representation allowed for the identification of system elements and the assessment of redundancy solutions. In the second stage, a FMECA analysis was performed, focusing on the mill body to identify the most relevant failure modes, their effects, and criticality. In the third stage, as a complement to the FMECA analysis, an Ishikawa diagram was created, focusing on the failure mode "low oil level in the gauges," enabling an analysis of possible causes. Finally, in the fourth stage, a FTA diagram was developed, representing the failure modes identified in the FMECA analysis, as well as their interconnections and probabilities of occurrence, based on the equipment's historical data.

The study began with an analysis through a reliability block diagram (series-parallel), representing the main equipment that makes up the raw mill, such as, turner, coupling, main motor, gearbox, mill body, recirculation elevator and separator.

The construction of this diagram allowed for the visualization of the interdependencies between the different pieces of equipment and the assessment of possible redundancy solutions, that is, the existence of equipment capable of ensuring operation in the event of a failure of any component (Figure 3.1).



Figure 3.1 Raw Mill Reliability Block Diagram

After analyzing the diagram, it was confirmed that there are no redundancy systems— all equipment operates sequentially and interdependently, with every piece necessary for the continuous operation of the process. The only equipment that does not operate continuously with the others is the Reclaimer, which is only used when a specific maneuver is required on the mill, for example during maintenance or when it is necessary to adjust the thickness of the material layer. In such cases, the Reclaimer is used to perform the first pass before starting the main motor.

Consequently, whenever any of the constituent equipment fails, it inevitably results in the shutdown of the raw mill, classifying this system as critical to the process.

It is important to recall that in situations where the raw mill stops and there is insufficient stored raw meal in the silos, this failure can also cause the shutdown of the kiln stage, as the raw material feed to the kiln is interrupted, leading to significant consequences in terms of efficiency, operational costs, and production losses. This analysis highlights the importance of continuous monitoring of this equipment to minimize the risk of unexpected failures.

After preparing the block diagram and analysing the operation of the equipment and its components, the FMECA analysis follows, with the aim of analysing one of the most critical components of this equipment, in which case the mill body was chosen. This choice was based on the nature of the failures in this equipment and the size of the consequences of the failures for the operation of the mill.

The first phase of the analysis consisted of collecting incidental downtime data for the raw mill from January 2022 to April 2025. After data collection, a data screening was performed, which revealed some difficulties. Approximately 30% of the downtimes attributed to the mill body were classified under the generic category "Others." Specifically, out of 27 recorded downtimes attributed to the mill body, 8 were classified as "Others." Overall, of the 151 incidental downtimes recorded during this period for the raw mill, 69 were classified as "Others," representing about 45% of the total incidental downtimes, which complicated the analysis and understanding of the failure modes.

Given this situation, a more detailed analysis of the descriptive comments associated with these downtimes was necessary. This analysis allowed the identification of the failure modes related to each occurrence, enabling the reclassification of the data and ensuring a more reliable basis for the development of the FMECA analysis. After data screening and reclassification of some downtimes, the FMECA Analysis was performed. For this purpose, a form was prepared, as shown in Figure 3.2, following the IEC 60812:80 standard with some modifications.

FMECA Analysis											
System / Equipment	Component / Subpart	Function	Failure Mode	Cause(s)	Local Effect	Global / Final Effect	Severity (S)	Occurrence (O)	Detectability (D)	RPN	Mitigation Actions

Figure 3.2 Form for FMECA analysis

The form was completed systematically after the information was collected and organized. Initially, the data were structured by equipment subpart and identified failure modes, with this identification complemented by a description of the function performed by each subpart, allowing for the contextualization of the failure's impact on the overall system. For each failure mode, the possible causes were identified, as well as the effects caused both locally and on the operational performance of the raw mill (Figure 3.3).

FMECA Analysis						
System / Equipment	Component / Subpart	Function	Failure Mode	Cause(s)	Local Effect	Global / Final Effect
Raw Mill (SE12)	Hydraulic Cylindre	Ensure the pressure on the rollers required for raw material grinding	Tie rod fracture	Component fatigue	Pressure imbalance in the roller grindings	Mill stop
			Cylinder position detector failure	Pressure imbalance in the cylinders	roller grinding out of level	Risk of shutdown to restore normal operating conditions of the roller grindings
			Cylinder base bolts broken	Component fatigue	Increase in vibrations	Risk of mill shutdown
	Roller Grinding	Grind the material coming from the quarry	Low oil level	Damaged seals	Loss of lubrication	Risk of mill shutdown caused by low oil level
				Probe installed without measurement capability	Incorrect sensor reading	Incorrect oil level information
				Loose plug	Loss of lubrication	Risk of mill shutdown caused by low oil level
	Dam ring and wear plates	Dam ring that retains the material on the grinding table during milling	Broken fixing of the gauge liner	Wear	Wear plates drop caused damage to the dam ring and the lower plates	Mill shutdown due to structural damage in the mill
	Sealing air fan for grinding rollers	Supply constant air pressure for roller sealing	Loss of air pressure in the sealing line	Damaged pipeline (ruptured)	Air pressure loss in the sealing line	Air sealing failure in the roller grinding
				Breakage of the bolt fixing the piping support	Drop in roller grinding sealing pipe and loss pressure	Air sealing failure in the roller grinding
	viabration sensor	Vibration measurement	Vibration warning	Possible lack of material on the mill plate	Increased vibrations detected in the system	Risk of mill shutdown due to high vibration levels
				Sensor wiring loose	High vibrations	Shutdown due to high vibrations
	Wear plates	Protect the base plate against wear to reduce abrasion	Lower liner dropped (bottom part of the wear plates)	Fracture of the weld supporting the wear plates	Abnormal noise at the frequency of a rotation in the elevator SE14 zone	Mill stop
	Interbus card	Layer control card	Communication failure	Interbus card damaged	Loss of communication	Mill stop
	Low-pressure oil Pump 1	Lubrication	Low lubrication level	Clogged oil filters of the low-pressure pump	Lack of oil	Risk of mechanical failure
	Hydraulic cylinder position detectors	Measuring the position of the cylinders to ensure the correct position of the gauges	Broken position detector support	Component fatigue	roller grinding out of level	Risk of shutdown to restore normal operating conditions of the roller grindings
	Hydraulic Unit for Raw Mill 3	Control of the hydraulic cylinders	Isolated cylinder	Oil intake valve closed	Lack of cylinder pressure	Cylinder failure
	Water pipework	Water supply to the mill	Water leakage in the pipeline	Defective accessory	Water leakage	

Figure 3.3 First phase of the FMECA analysis

After completing the identification of the failure modes found, the next phase involved the criticality analysis through the classification of the three fundamental FMECA parameters: Severity (S),

Occurrence (O), and Detectability (D).

Although the IEC 60812:2018 standard typically recommends using a scale from 1 to 10 to assign values to each parameter, this analysis adopted a scale from 1 to 4. This decision aimed to simplify the analysis and ensure greater consistency in classification, given the nature of the available data and the complexity of the system analyzed (Figure 3.4).

Caption		
Level		Description
Severity (S)		
1	Very low	Insignificant functional impact
2	Low	Minimal reduction in perfotmance or quality with tolerable effects
3	Moderato	Significant reduction in performance, potential impact on functionality
4	High	Critical failure may compromise safety or cause system shutdown
Occurrence (O)		
1	Rare	Less than one occurrence per year
2	Occasional	Approximately one occurrence per year
3	Frequent	Several times per year
4	Highly frequent	Monthly or more frequently
Detectability (D)		
1	Very high	Very higth probability of detecting the failure
2	High	High probability of detecting the failure
3	Low	Some probability of detecting the failure
4	Very low	Low probability of detecting the failure

Figure 3.4 Caption FMECA Analysis

The classification of the parameters for each identified failure mode was carried out with the support of a maintenance technician from the Inspection and Preparation Department of the Souselas Production Center. This collaboration was crucial to ensure the consistency of the assigned ratings, as the technician possesses in-depth knowledge of the equipment and its operation, as well as past failures. This process helped guarantee that the analysis accurately reflected the operational reality of the equipment, reinforcing the results obtained and the subsequent prioritization of failure modes based on the Risk Priority Number (RPN) – Figure 3.5.

RPN	Color code	Description
0-16		Low risk
17-32		Moderate risk
33-48		High risk
49-64		Critical risk

Figure 3.5 Caption with values and color code for RPN interpretation

Finally, after completing the FMECA analysis, it was possible to identify the most critical failure modes related to the mill body, based on the calculated RPN values. Using these values along with a color-coded scheme, the criticality of the various failure modes was effectively demonstrated.

The analysis revealed that certain failure modes, although of low occurrence, exhibit high severity and low detectability. This situation is mainly due to the size of the components involved and the difficulty in performing corrective interventions, as these failures tend to occur unexpectedly and require prolonged downtime for repair, significantly impacting the operational availability of the equipment.

It was concluded that the most critical failure mode corresponds to the breaking of the tie rod, caused by progressive wear, which implies an immediate shutdown and, consequently, the stoppage of the kiln.

This result highlights the importance of studying monitoring strategies, such as ultrasonic testing, and intensifying routine inspections to anticipate component degradation. In this way, tie rod replacements are intended to no longer occur reactively and unexpectedly but to become part of the preventive maintenance plan based on objective data. This approach allows minimizing the impact of failures, reducing unplanned downtime, and increasing equipment availability and reliability.

The second most relevant failure mode, considering its occurrence, is low oil level in the gauges, resulting

from seal leakage. Although this failure mode has higher detectability, its high occurrence makes it a priority for maintenance management. Whenever an oil level below the recommended threshold is detected, replenishment is carried out, allowing the mill to continue operating until the next scheduled shutdown. However, this temporary solution does not represent a definitive correction, incurring costs related both to the labor required for refilling and to the oil consumption.

This procedure is justified by the fact that replacing the leaking pair of gauges with a spare requires a prolonged shutdown, which leads to attempts to postpone this intervention as much as possible until it coincides with a planned maintenance stop.

Given its relevance, this failure mode was subject to a more in-depth analysis through the construction of an Ishikawa diagram, aiming to understand its possible causes and identify effective mitigation measures.

Finally, after completing the FMECA analysis, it was possible to identify which failure modes are most critical regarding the mill body, based on the calculated RPN values and, through its value and a colour scheme, it became possible to demonstrate the criticality of the various failure modes.

The FMECA analysis thus proved to be a strategic tool for operational decision-making and continuous improvement, contributing to the enhancement of equipment availability and reliability (Figure 3.6).

FMECA Analysis

System / Equipment	Component / Subpart	Function	Failure Mode	Cause(s)	Local Effect	Global / Final Effect	Severity (S)	Occurrence (O)	Detectability (D)	RPN	Mitigation Actions
Raw Mill (SE12)	Hydraulic Cylindre	Ensure the pressure on the rollers required for raw material grinding	Tie rod fracture	Component fatigue	Pressure imbalance in the roller grindings	Mill stop	4	3	4	48	Replacement of the tie rod
			Cylinder position detector failure	Pressure imbalance in the cylinders	roller grinding out of level	Risk of shutdown to restore normal operating conditions of the roller grindings	3	2	3	18	Pressure regulation of the cylinders
			Cylinder base bolts broken	Component fatigue	Increase in vibrations	Risk of mill shutdown	2	2	4	16	Replacement of broken bolts
	Roller Grinding	Grind the material coming from the quarry	Low oil level	Damaged seals	Loss of lubrication	Risk of mill shutdown caused by low oil level	4	3	3	36	Restoring the oil level followed by roller grinding repair
				Probe installed without measurement capability	Incorrect sensor reading	Incorrect oil level information	4	1	3	12	Replacement of the sensor with one of higher measurement capacity
				Loose plug	Loss of lubrication	Risk of mill shutdown caused by low oil level	2	1	2	4	Retightening of the plug
	Dam ring and wear plates	Dam ring that retains the material on the grinding table during milling	Broken fixing of the gauge liner	Wear	Wear plates drop caused damage to the dam ring and the lower plates	Mill shutdown due to structural damage in the mill	4	2	4	32	Replacement of the north roller grinding, full replacement of the dam ring, mill skirt, drag plates, and repair of the mill body
	Sealing air fan for grinding rollers	Supply constant air pressure for roller sealing	Loss of air pressure in the sealing line	Damaged pipeline (ruptured)	Air pressure loss in the sealing line	Air sealing failure in the roller grinding	3	2	4	24	Repair of piping
				Breakage of the bolt fixing the piping support	Drop in roller grinding sealing pipe and loss pressure	Air sealing failure in the roller grinding	2	1	4	8	Replacement of broken bolts and lubrication of the spherical joints
	viabration sensor	Vibration measurement	Vibration warning	Possible lack of material on the mill plate	Increased vibrations detected in the system	Risk of mill shutdown due to high vibration levels	3	4	1	12	Restart of the system
				Sensor wiring loose	High vibrations	Shutdown due to high vibrations	1	1	3	3	Retightening of the wire
	Wear plates	Protect the base plate against wear to reduce abrasion	Lower liner dropped (bottom part of the wear plates)	Fracture of the weld supporting the wear plates	Abnormal noise at the frequency of a rotation in the elevator SE14 zone	Mill stop	2	2	3	12	Weld repair of the wear plates
	Interbus card	Layer control card	Communication failure	Interbus card damaged	Loss of communication	Mill stop	2	1	4	8	Replacing the control card
	Low-pressure oil Pump 1	Lubrication	Low lubrication level	Clogged oil filters of the low-pressure pump	Lack of oil	Risk of mechanical failure	2	1	4	8	Replacement of filters
	Hydraulic cylinder position detectors	Measuring the position of the cylinders to ensure the correct position of the gauges	Broken position detector support	Component fatigue	roller grinding out of level	Risk of shutdown to restore normal operating conditions of the roller grindings	1	2	3	6	Repair of the support structure
	Hydraulic Unit for Raw Mill 3	Control of the hydraulic cylinders	Isolated cylinder	Oil intake valve closed	Lack of cylinder pressure	Cylinder failure	2	1	2	4	Restoration of operating conditions (valve opening)
	Water pipework	Water supply to the mill	Water leakage in the pipeline	Defective accessory	Water leakage		1	1	3	3	Disassembly of accessory, condition analysis, reassembly and retightening of the union to eliminate water leakage

Figure 3.6 FMECA Analysis

After identifying the failure modes related to the mill body, the failure mode "low oil level in the gauges" stood out due to its recurrence. This recurrence justified a more in-depth analysis through the creation of an Ishikawa diagram, also known as a fishbone diagram.

This method allows for a structured organization of factors that may cause the failure, grouping them into categories such as: method, labor, machine, maintenance, measurement, and materials; the objective is to identify possible causes contributing to this failure mode.

According to historical data, the low oil level failure mode in the gauges is mainly attributed to seal leaks, directly affecting the oil retention in the gauges. However, a thorough analysis revealed other possible causes for this failure mode, such as:

- Method: Failure in the verification of plug tightening;
- Machine: Loose plug, damaged seals;

- Material: Poor seals quality;
- Measurement: Low-capacity sensor, incorrect reading;
- Maintenance: Failure to tighten the plug;
- Manpower: Incorrect plug tightening.

The diagram resulting from this analysis (Figure 3.7) clearly shows the possible root causes associated with the failure mode, serving as support for defining more effective corrective and preventive actions.

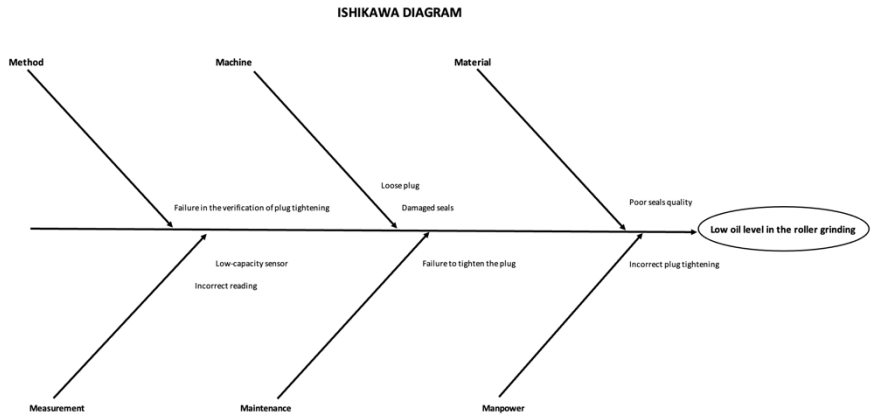


Figure 3.7 Ishikawa diagram: Low oil level in roller grinding

Just like the Ishikawa diagram, an FTA was developed as a complement to the FMECA analysis carried out with a specific focus on the mill body.

This tool allowed a deeper investigation by quantifying the probability of occurrence of the failure modes previously identified in the FMECA, through a logical tree structure. This approach proved particularly useful to understand which events carry the greatest statistical weight in causing the mill to stop.

It is important to highlight that this analysis was constructed based on historical shutdown data of the mill, recorded between January 2022 and April 2025. Unlike theoretical or estimation-based approaches, this analysis relied on concrete facts and records, allowing for more precise quantification of the causes that contributed most to system failures.

Through the fault tree analysis, it was clearly identified that the failure chain with the highest probability of occurrence is associated with the low oil level in the bearings, because of oil leaks caused by damaged seals. This failure mode, already highlighted as critical in the FMECA analysis and included in the Ishikawa diagram as one of the main causes, is now supported by concrete data obtained from the shutdown history.

The convergence of results obtained through qualitative (Ishikawa diagram), semi-quantitative (FMECA analysis), and quantitative (FTA analysis, Figure 3.8) approaches clearly validates the consistency of the conclusions of this study. The gathered evidence points to the need to strengthen monitoring and prevention of failures caused by oil leaks in the bearings, whose contribution to mill shutdowns is

significant.

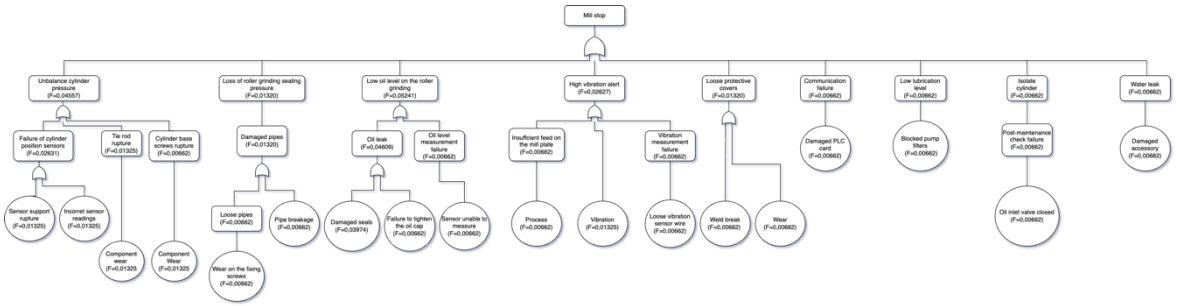


Figure 3.8 Fault tree analysis

4 Results

The reliability block diagram analysis identified the absence of redundancies in the raw mill. All equipment within the raw mill system is arranged in series, implying that the failure of any single component results in the shutdown of the entire mill.

The results obtained from the FMECA analysis revealed the failure modes with the greatest impact on operation: the tie rod breakage mode, which presents a high RPN due to its high severity and low detectability, and the oil leakage mode in the roller grindings, which exhibits a high RPN primarily due to its high occurrence rate.

Complementing the FMECA analysis, an Ishikawa diagram was used to further detail the causes of the low oil level failure mode in the roller grinding system.

Finally, the Fault Tree Analysis (FTA) confirmed that the failure mode with the highest probability of occurrence in the system is the oil leakage in the roller grindings, resulting from multiple combined failures. The application of OR logic demonstrated that the occurrence of any of the identified causes leads to the top event.

5 Discussion

The results of this study demonstrate that the combined application of methodologies such as the RBD, FMECA, Ishikawa Diagram, and FTA enables a comprehensive understanding of potential failure modes in the raw mill. The absence of redundancy, identified through the RBD, amplifies the impact of isolated failures, reinforcing the importance of a preventive and structured maintenance approach.

The FMECA proved effective in identifying and prioritizing failure modes, while the Ishikawa Diagram clearly represented the possible root causes of the problem, particularly regarding the low oil level failure in the roller grinding system. The FTA complemented the process by providing logical and quantitative validation of failure mode probabilities. As described by several authors [5], [11], [12], this approach aligns with international best practices in reliability management and the principles of Reliability-Centered Maintenance (RCM).

The study also highlights the importance of reliable and properly categorized historical data, whose continuous improvement will contribute to more rigorous and well-founded analyses in the future.

6 Conclusion

This study aimed to demonstrate the application of integrated methodologies in the management of industrial physical assets, specifically focusing on the analysis of a raw mill in a cement plant.

As previously mentioned in earlier chapters, this study was based on the use of four integrated methodologies with the goal of understanding and quantifying the existing failure modes in the selected equipment. Through this study, it was possible to confirm the absence of redundancies in this equipment, which contributes to an increased likelihood of equipment failure.

The low oil level in the roller grinding system was identified as the most recurrent failure mode, while the failure mode of tie rod breakage stood out as the most critical event in terms of severity, given its potential to cause severe damage and prolonged downtime.

As future actions, it is recommended to improve the digitization and categorization of failure modes to enable more precise and historically comparable analyses. Additionally, a review of preventive maintenance plans for the components associated with the most critical failure modes identified in this study is advised. Finally, the integration of predictive models through machine learning is recommended to detect possible deviations in operational patterns, thereby supporting decision-making based on real-time condition data of the assets.

These initiatives align with the principles of maintenance digitalization, the paradigms of Industry 4.0, and the natural evolution of maintenance towards a more intelligent, automated, and data-driven model.

References

[1] C. V. Pinto, Organização e gestão da manutenção. IAPMEI, 2002. Monitor. ISBN: 9789729413391.

[2] IPQ - Instituto Português da Qualidade, "NP EN 13306:2021 - Manutenção: Terminologia de Manutenção," Lisboa, 2021.

[3] Stamatis, DH. (2003). *Failure Mode and Effect Analysis* (2nd ed.). ASQ Quality Press. Retrieved from https://www.perlego.com/book/3772269/failure-mode-and-effect-analysis-fmea-from-theory-to-execution-pdf (Original work published 2003)

[4] P. C. Teoh and K. Case, "Failure modes and effects analysis through knowledge modelling," Journal of Materials Processing Technology, vol. 153–154, pp. 253–260, Oct. 2004, doi: 10.1016/j.jmatprotec.2004.04.298.

[5] International Electrotechnical Commission, *IEC 60812:2018 - Failure modes and effects analysis (FMEA and FMECA)*, 3rd ed. 2018. Accessed: May 05, 2025. [Online]. Available: www.iec.ch

[6] Mikulak, R. J., McDermott, R., Beauregard, M. (2017). *The basics of FMEA*. New York: Productivity Press. ISBN: 9780429244773

[7] A. S. Varde, J. Liang, Z. Yang, and R. D. Sisson, "Ishikawa, JESS, and Visual Analytics for Engineering," 2021 IEEE International Conference on Big Data (Big Data), pp. 6824–6826, Dec. 2022, doi: 10.1109/bigdata55660.2022.10020637.

[8] E. Zio, An introduction to the basics of reliability and risk analysis. 2007. Series on Quality, Reliability and Engineering Statistics: Volume 13. https://doi.org/10.1142/6442

[9] International Electrotechnical Commission, *IEC 61025: Fault tree analysis (FTA)*. Geneva, Switzerland: International Electrotechnical Commission, 2006.

[10] C. A. Ericson, Fault Tree Analysis primer. CreateSpace Independent Publishing Platform. 2011. ISBN-13: 978-1466446106 2011.

[11] H. Soltanali, M. Khojastehpour, J. T. Farinha, and J. E. De Almeida E Pais, "An Integrated Fuzzy Fault Tree Model with Bayesian Network-Based Maintenance Optimization of Complex Equipment in Automotive Manufacturing," Energies, vol. 14, no. 22, p. 7758, Nov. 2021, doi: 10.3390/en14227758.

[12] H. Raposo, J. T. Farinha, I. Fonseca, and D. Galar, "Predicting condition based on oil analysis – A case

study," Tribology International, vol. 135, pp. 65–74, Feb. 2019, doi: 10.1016/j.triboint.2019.01.041.

Chapter 39

Challenges in Preparing a Procurement Specification for Hospital Equipment

Contribution #72
Cite as:
Ana Martins, Fernanda Coutinho, Hélio Rodrigues, Alberto Adrêgo. Challenges in Preparing a Procurement Specification for Hospital Equipment. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-72,
  title ={Challenges in Preparing a Procurement
     Specification for Hospital Equipment},
  author={Ana Martins and  Fernanda Coutinho and  Hélio
     Rodrigues and  Alberto Adrêgo},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Challenges in Preparing a Procurement Specification for Hospital Equipment

Ana Martins*[1], Fernanda Coutinho[1,2], Hélio Rodrigues[3], Alberto Adrêgo[3]

[1]Polytechnic University of Coimbra, Coimbra Institute of Engineering, Portugal
[2] RCM2+ Research Center, Coimbra, Portugal
[3] Unidade Local de Saúde de Coimbra, Coimbra, Portugal
*Corresponding author
a2022138435@isec.pt

Abstract

A hospital engineer is a multidisciplinary professional who must possess competencies in a range of areas, including electrical engineering, mechanical engineering, healthcare equipment management, and hospital services management. In addition, cross-cutting skills are required to enable effective interaction with various healthcare professionals. Among the typical responsibilities of a hospital engineer is the preparation of technical specifications (tender documents) for the procurement of hospital equipment, as well as the oversight of the corresponding public procurement procedures. This paper focuses on this fundamental dimension of the hospital engineer's role, by systematizing the main steps to be adopted and identifying the key challenges inherent in the process. These include the precise definition of technical specifications for the equipment to be procured, a thorough understanding of the applicable legal and regulatory frameworks for drafting procurement documents (e.g., relevant legislation, guidelines issued by the national health authority), and the establishment of objective criteria for the evaluation of bids, among others. Furthermore, the paper discusses strategies to ensure proper coordination among the various stakeholders involved and to safeguard transparency throughout the procurement process. Finally, a practical example of a technical specification document is presented (with appropriate anonymization of sensitive data).

Keywords: Hospital engineering; Medical equipment procurement; Technical specifications.

1 Introduction

The contemporary healthcare landscape is marked by an unprecedented reliance on sophisticated medical technology, with hospitals in developed countries investing approximately €93 billion annually in the life cycle costs of medical equipment [1]. This technological evolution has fundamentally transformed the delivery of healthcare, creating opportunities for improved diagnostic and treatment outcomes while simultaneously introducing complex challenges in the procurement and management of technology [2]. At the heart of this technological ecosystem, biomedical engineers serve as critical intermediaries between clinical needs and technological solutions, bearing the responsibility of ensuring that equipment procurement decisions optimise both clinical effectiveness and economic efficiency.

The role of biomedical engineers in hospital equipment procurement has evolved significantly beyond traditional technical support functions. Modern biomedical engineers are required to possess multidisciplinary competencies encompassing principles of electrical and mechanical engineering, regulatory compliance knowledge, project management skills, and advanced interpersonal abilities for effective stakeholder coordination [3].

Hospital equipment procurement operates within a complex regulatory environment that varies significantly across jurisdictions but consistently emphasises transparency, competition, and value optimisation. In the European context, procurement processes must comply with both national public contracting laws and European Union directives governing the safety and performance of medical devices [4]. This regulatory complexity presents substantial challenges for biomedical engineers, who must be well-versed in legal requirements while ensuring that technical specifications adequately address clinical needs and institutional constraints.

Recent systematic reviews of hospital equipment procurement practices reveal significant inefficiencies and quality concerns within current approaches [5]. Studies indicate that procurement decisions often prioritise cost containment over quality and clinical outcomes, leading to the selection of undervalued equipment that may compromise patient care and increase long-term operational costs. Moreover, the involvement of multiple stakeholders with differing perspectives and priorities frequently results in conflicts and decision-making delays, underscoring the need for structured approaches to stakeholder management and consensus building.

Translating clinical needs into precise technical specifications represents one of the most complex aspects of hospital equipment procurement. Healthcare professionals articulate requirements in terms of clinical outcomes and patient care goals, whereas procurement processes demand detailed technical parameters and measurable performance criteria [6]. This translation process requires a deep understanding of both clinical workflows and technological capabilities, as well as the foresight to anticipate future needs and technological evolution over equipment life cycles that may span 10 to 15 years.

This article addresses these challenges by presenting a comprehensive framework for biomedical engineers involved in hospital equipment procurement. The framework integrates current best practices in the development of technical specifications, stakeholder management, and multi-criteria evaluation, while ensuring compliance with essential regulatory requirements. Through a systematic review of the existing literature and the application of a practical case study, we demonstrate how structured procurement approaches can significantly improve outcomes for healthcare institutions, equipment suppliers, and ultimately, patients.

2 Challenges in the Procurement of Medical Equipment

Hinrichs-Krapels et al. [5] conducted a comprehensive systematic review on the procurement of high-cost medical equipment in hospitals located in high-income countries, analysing academic literature from the year 2000 to 2020. Their findings reveal significant systemic challenges in current procurement practices, including an excessive focus on cost containment at the expense of quality and health outcomes. The study identifies numerous stakeholders involved in equipment selection decisions, including clinicians, nurses, biomedical engineers, financial officers, and managers, but notes that their roles and protocols are often poorly defined, leading to fragmented decision-making processes. The systematic review further highlights that these processes go beyond basic procurement to encompass user needs assessment, technical maintenance planning, training requirements, consumables management, and disposal considerations. Additionally, the study underscores the lack of stakeholder analysis in these planning processes, which frequently results in conflicts and delays in decision-making.

3 Impact and Competencies of Clinical Engineering

The authors of [2] present compelling evidence for the critical role of clinical and biomedical engineers in the management of healthcare technologies through an international study encompassing 400 case studies from 125 countries. This research demonstrates that the involvement of biomedical engineers in guiding health technology yields measurable positive impacts on patient outcomes, challenging the perception that evidence for such contributions is limited.

The study highlights the unique "horizontal expertise" provided by biomedical engineers, which complements the "vertical expertise" of traditional medical professions. This horizontal expertise encompasses the ability to assess and manage the integration of complex systems across diverse service areas, technology life cycles, hardware and software platforms, and middleware requirements for integrated medical and surgical services.
Furthermore, the research identifies biomedical engineers as essential professionals in providing guidance throughout the full equipment life cycle, including the stages of innovation and development with regulatory compliance, evaluation, installation, training, integration, performance assurance, and risk mitigation. The international scope of the study, covering both developed and developing healthcare systems, offers valuable insights into the universal challenges of health technology management and the consistent value delivered by biomedical engineering expertise across diverse settings.

4 Life Cycle Assessment and Patient Safety Considerations

Harkin et al. [7] this study provides critical insights into the complexity of life cycle assessment (LCA) of medical equipment through a comparative analysis of 52 evaluation models used by different stakeholder groups. The investigation reveals significant heterogeneity in how the concepts of "life cycle" and "life cycle assessment" are defined and applied, with important implications for patient safety and evidence generation. The study identifies four distinct perspectives on life cycle assessment: (i) Healthcare (focusing on clinical adoption decisions); (ii) Policymaking (emphasizing resource allocation); (iii) Commerce and Industry (prioritizing investment and marketing strategies); and (iv) Regulation (centered on market oversight). This multi-perspective analysis underscores the importance of clearly defining evaluation objectives and the roles of stakeholders in procurement processes. The study raises serious concerns about the risk that life cycle approaches may delay evidence generation until the post-market phase. Such delays could compromise patient safety, particularly if the information used for procurement decisions is deemed insufficient.

5 Methodology for the Preparation of a Statement of Requirements

This section provides a comprehensive framework for biomedical engineers to develop technical specifications and evaluation criteria for the procurement of hospital equipment. The methodology integrates established best practices in engineering, stakeholder management, and multi-criteria decision-making, while ensuring compliance with regulatory frameworks. The main phases of the process are as follows (Fig. 1).



Figure 1 - Methodology for the Preparation of a Statement of Requirements (step-by-step).

1. **Stakeholder Analysis and Needs Assessment:** The initial phase focuses on identifying and analyzing all stakeholders involved in the procurement process, followed by an assessment of clinical and operational needs. This phase is critical to ensure that the development of technical specifications addresses all relevant perspectives and requirements:
    - **Stakeholder Identification:** Stakeholder identification employs a systematic

approach to identify all individuals and groups with an interest in the procurement outcome. Primary stakeholders include clinical users (e.g., physicians, nurses, technicians), technical support staff (e.g., biomedical engineers, IT specialists), administrative personnel (e.g., procurement officers, financial managers), and external parties (e.g., suppliers, regulatory bodies, patients).

- **Clinical Needs Assessment:** Clinical needs assessment utilises structured interview techniques and observational studies to understand current workflows, identify performance gaps, and define desired outcomes. The process focuses on translating clinical objectives into measurable parameters that can be incorporated into the technical specifications.
- **Operational Requirements Analysis:** Operational requirements analysis examines the broader institutional context in which the equipment will operate. This includes physical infrastructure constraints, maintenance and support capabilities, financial limitations, and strategic alignment with institutional goals.

2. **Regulatory Compliance:** The second phase ensures a comprehensive understanding and integration of all applicable regulatory requirements into the procurement specifications. The integration of standards encompasses both mandatory requirements and voluntary best practices from relevant international organisations. Key standards include ISO 13485 (quality management for medical devices), the IEC 60601 series (safety of medical electrical equipment), and device-specific standards relevant to the category of equipment being procured. The integration process involves a detailed analysis of each applicable standard to identify specific requirements that must be reflected in the technical specifications. This includes performance parameters, safety requirements, testing protocols, documentation standards, and quality assurance procedures.

3. **Development of Technical Specifications:** The third phase translates stakeholder needs and regulatory requirements into precise technical specifications that can be effectively communicated to potential suppliers and objectively assessed during procurement. Performance criteria compassing aspects such as accuracy, precision, speed, reliability, and availability are objective, measurable parameters essential for evaluating device capabilities. Their definition results from a careful analysis of clinical workflows, comparison with existing equipment, consultation of manufacturer specifications, and anticipation of technological trends. It is crucial that these criteria include target values, acceptable ranges, and measurement methodologies, while also accounting for performance degradation over time, thereby requiring the specification of maintenance and calibration needs. In parallel, interface and integration requirements are vital to ensure seamless interaction of the equipment with hospital systems, other medical devices, infrastructure, and, naturally, end-users.

4. **Designation of Evaluation Criteria:** It is essential to clearly define how supplier proposals will be assessed. This phase focuses on establishing a fair and transparent evaluation system through a set of essential criteria. Main criteria, such as technical quality, cost, supplier capability, and strategic factors, are prioritised and weighted according to the specific importance for the equipment in question. Once scores are assigned, methods combining quantitative values and qualitative assessments are applied. Finally, scores are aggregated using weighting and consolidation techniques to generate the final ranking of proposals.

5. **Implementation and Validation:** The implementation and validation of procurement outcomes rely on integrating the previously defined specifications and evaluation criteria into the institution's existing procurement procedures. This involves not only technical aspects, such as document formatting, but also procedural coordination, including team management and deadline adherence. Following implementation, it is crucial to monitor and evaluate outcomes. This entails tracking both immediate results, such as process efficiency and stakeholder satisfaction, and long-term impacts, including equipment performance and clinical outcomes. Continuous monitoring forms the foundation for an ongoing process of improvement.

6 Example of a Statement of Requirements for a Surgical Robot

This section presents a case study involving the development of a procurement specification document by the Coimbra Hospital and University Centre (CHUC) for the acquisition of a surgical robot for the Central Operating Theatre at CHUC [8]. It offers the reader a detailed perspective on the technical and operational requirements involved in integrating advanced robotic technology within healthcare settings. The specification document served as the foundation for an international public tender and established the guidelines for the acquisition of a surgical robotic system comprising four robotic arms, a visualisation tower, a surgical console, and the imaging system (Figure 2). Its implementation aims to enhance surgical capabilities in specialties such as General Surgery, Gynaecology, and Urology, with the potential for future expansion into Thoracic Surgery, Paediatric Surgery, Otorhinolaryngology, and Maxillofacial Surgery. Beyond the technical specifications of the equipment, the document emphasises the importance of a comprehensive set of complementary services to ensure the system's operability. These include installation with the necessary adaptations to the operating theatre, training for healthcare and IT professionals, preventive and corrective maintenance during the warranty period, and integration with existing applications at CHUC. The inclusion of consumables for the first year of operation reflects a holistic approach to the procurement, ensuring that the initial investment is accompanied by the necessary resources for immediate and effective use. The base value of the tender, €2,455,000.00, reflects the magnitude of the investment in this cutting-edge technology and underlines its strategic relevance in modernising healthcare services and improving patient care outcomes.
In accordance with the methodological framework proposed in the previous section, the objective is to demonstrate, concisely and step-by-step, how each phase contributes to the elaboration of the Statement of Requirements for this equipment, thereby ensuring a proper procurement process for this complex medical technology.



Figure 2 – Surgical Robot (Da Vinci Robotic Surgical System) [10].

1. **Stakeholder Analysis and Needs Assessment:** The process is initiated by identifying relevant stakeholders and assessing their needs with respect to the system to be acquired. Clinical and operational requirements are then identified, to be included explicitly in the

document. Article 1 (Subject Matter) indicates that the robot is intended for multiple surgical specialties (General Surgery, Gynaecology, Urology, etc.), identifying surgeons as primary stakeholders with requirements for precision and versatility. The provision for training in Article 1, paragraph 3, subparagraph ii), highlights the importance of operating theatre nurses and technicians. The Facilities and Equipment Services (SIE) and the Information Technologies and Systems Services (STSI) are critical for the system's installation and integration, as stated in Article 1, paragraph 3, subparagraphs i) and iv), and Article 17. Finally, patients are the ultimate stakeholders, benefiting from improved surgical outcomes and safety, as evidenced by the requirement for consumables to support a significant surgical volume (Article 1, paragraph 1, and Annex II). In essence, the needs of all stakeholders are addressed in the document's specifications, with the aim of enhancing patient care. [9]

2. **Regulatory Compliance:** An analysis of relevant regulatory requirements is made which leads to several provisions mentioned throughout the document aiming to ensure regulatory compliance. Article 15 (Applicable Legislation) establishes the Public Contracts Code (CCP) as the legal framework. Article 13 (Data Protection) mandates compliance with the General Data Protection Regulation (GDPR) and Law no. 58/2019, both critical for safeguarding patient data. The Technical Standards (Article 18), though not detailed, require the equipment to meet recognised safety and performance benchmarks (e.g., CE marking). Additionally, Article 5 (General Obligations of the Parties), paragraph 1, subparagraph v), refers to the payment of fees to the Court of Auditors and licensing authorities, further reinforcing the legal dimension. [9]

3. **Development of Technical Specifications:** Technical specifications are derived from intended use and known best pracices and workflows, ensuring suitability for purpose. The technical specifications are detailed in Articles 16 and 17. Article 16, paragraph 1, requires functionalities to be presented in Portuguese. Specifications for the Surgical Console focus on high-definition 3D imaging, ergonomic controllers, real-time ultrasound display, and safety systems (e.g., positioning sensors). The Robotic Arms must feature a minimum of four arms, instrument exchange capabilities, remote centre technology to ensure patient safety, and seven degrees of freedom. The Visualisation Tower must include a 21-inch Full HD monitor, audio/video outputs, and safety alerts. Article 17 addresses the integration of the robot with CHUC's information systems, emphasising interoperability and data security. These specifications cover hardware, software, ergonomics, and safety, ensuring a technologically advanced and integrated system. [9]

4. **Designation of Evaluation Criteria**: The evaluation criteria encompass both financial and technical aspects. Article 4 (Base Price and Contractual Price) sets a base price of €2,455,000.00 and includes an exclusion criterion: proposals exceeding this amount are automatically disqualified (Article 4, paragraph 7). The contract is to be awarded as a single lot (Article 4, paragraph 8). The Technical Specifications (Articles 16 and 17) serve as eliminatory compliance criteria. Implicit qualitative criteria include Training (Article 1, paragraph 3, subparagraph ii)), Maintenance and Technical Support (Article 1, paragraph 3, subparagraph iii)), and the Management of Consumables (Article 1, paragraph 3, subparagraph v)), all aimed at ensuring sustainability and controlling long-term operational costs. This approach ensures selection based on a comprehensive and sustainable solution. [9]

5. **Implementation and Validation:** This phase is supported by clauses that define responsibilities and verification procedures. Article 1, paragraph 3, subparagraph i), requires the installation of the equipment and necessary adaptations in the operating

room. Training (Article 1, paragraph 3, subparagraph ii) and Article 21) is essential for user empowerment. Technical reception and acceptance testing (Article 20) serve as the formal procedure to verify equipment compliance. Maintenance and Technical Support (Article 1, paragraph 3, subparagraph iii) and Article 22) ensure ongoing operability. Lastly, Article 6 (Compliance Verification) allows for the replacement of non-conforming goods, reinforcing post-delivery quality assurance. [9]

7 Conclusion

An in-depth analysis of the technical specifications document for the acquisition of a surgical robot, in conjunction with the principles and challenges outlined in the supporting content, enables a robust understanding of the complexity and criticality involved in procuring advanced medical technology within hospital settings. This particular case study, is not merely an example of a procurement process, but rather a practical illustration of the intersection between clinical needs, technological innovation, and stringent regulatory and operational requirements.

The technical specifications document for the surgical robot, while normative in nature, goes beyond a simple listing of technical features. It reflects a holistic approach encompassing installation, the training of multidisciplinary teams (including physicians, nurses, technicians, and biomedical engineers), preventive and corrective maintenance, as well as integration with existing hospital information systems.

The role of the hospital engineer emerges as central in this process. As detailed in the Section 1, this professional, equipped with "horizontal expertise", acts as the essential link between clinical needs and technological solutions. Their ability to translate the requirements of healthcare professionals into measurable technical specifications, to navigate the complex regulatory frameworks (both national and European), and to manage the various stakeholders involved, is decisive to the success of the procurement process.

In summary, the acquisition of a surgical robot, such as the da Vinci system, represents a multifaceted challenge that demands a thorough understanding of technical, clinical, regulatory, and managerial dimensions. The articulation between theory and best practices, as presented in the analysed documents, and their practical application to real-world scenarios, are essential to ensure that investments in medical technology yield tangible improvements in healthcare delivery and patient safety. This process underscores the importance of interdisciplinary collaboration and a long-term strategic vision for the successful integration of technological innovations within the healthcare sector.

References

[1] Philips Healthcare, “Why lifecycle technology asset management?,” *Philips Healthcare Technology Services,* 3 March 2025.

[2] Y. David and T. Judd, “Evidence-based impact by clinical engineers on global patients,” *Health Technology,* pp. 19, July, 2019.

[3] Ordem dos Engenheiros, “Especialidade de Biomédica e Bioengenharia,” Ordem dos Engenheiros, June 26th 2025. [Online]. Available: https://www.ordemdosengenheiros.pt/pt/especialidades/biomedica-e-bioengenharia/.

[4] União Europeia, “Regulamento (UE) 2017/745 do Parlamento Europeu e do Conselho,” EUR-Lex, April 5th 2017. [Online]. Available: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32017R0745.

[5] S. Hinrichs-Krapels, B. Ditewig, H. Boulding, A. Chalkidou, J. Erskine and F. Shokraneh3, “Purchasing high- cost medical devices and equipment in hospitals: a systematic review,” *BMJ Open,* pp. 20, 12

February 2022.

[6] " Biomedical engineers are crucial for effective health-care systems," in *The Lancet Global Health*, vol. 13, 2025, pp. 3.

[7] K. Harkin, J. Sorensen and S. Thomas, "Lifecycle evaluation of medical devices: supporting or jeopardizing patient outcomes? A comparative analysis of evaluation models," *International Journal of Technology Assessment in Health Care,* pp. 12, 2023.

[8] Vortal, "Anúncio de Concurso: PT1.NTC.2466143," Vortal, 2025. [Online]. Available: https://community.vortal.biz/Public/contract-notice-view/PT1.NTC.2466143/?SkinName=VortalSkin1&currentLanguage=pt&currentLanguage=pt.

[9] Centro Hospitalar e Universitário de Coimbra, E.P.E., "Caderno de Encargos – Aquisição de Robot Cirúrgico para o Bloco Operatório Central do CHUC," 2023. [Online] https://community.vortal.biz/public/?SkinName=VortalSkin1&Page=login&currentLanguage=pt.

[10] P. Romanelli, "PEDROROMANELLI urulogia e cirurgia robótica," July 2nd 2020. [Online]. Available: https://urocirurgia.com.br/robo-cirurgiao-decifrando-o-sistema-cirurgico-da-vinci/.

Chapter 40

Exploration of vibration signatures and patterns for bearings failure modes

Contribution #73
Cite as:
Abdu Shaalan, Osarumen O. Ogbomo, David Baglee. Exploration of vibration signatures and patterns for bearings failure modes. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-73,
  title ={Exploration of vibration signatures and
     patterns for bearings failure modes},
  author={Abdu Shaalan and  Osarumen O. Ogbomo and
     David Baglee},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Exploration of vibration signatures and patterns for bearings failure modes

Abdu Shaalan[1*], Osarumen O. Ogbomo[1], David Baglee[1]

[1]University of Sunderland, School of Computer Science & Engineering, Faculty of Business & Technology, Sunderland, UK
*Corresponding author {Abdu.Shaalan@sunderland.ac.uk}

Abstract

Vibration analysis plays a pivotal role in machinery health assessment, especially within industrial contexts. It can support early fault detection for critical machines' components and prevent sudden and catastrophic failures. This article utilizes advanced tools like the Machinery Fault Simulator (MFS) with a Multi-channel DAQ and Vibra-Quest (VQ) to scrutinize vibration patterns of bearing failure modes including eccentric sheave, loose belt, defected shaft, defected bearing and worn gearbox. Distinct vibration signatures, such as Vibration velocity (Vrms) and Crest Factor (CF), which indicate machinery health, were experimentally determined for various mechanical faults at different frequencies and speeds. A clear link was established between fault severity and vibration magnitude, particularly at high speeds. The defected bearing had the highest Vrms and CF of 0.3048 mm/s and 7.0276 respectively. Also, Vibration Increment was observed on different sensor channels and frequency speeds of 10 - 20Hz demonstrated more prominence for most fault types. The ideal placement of sensors was dependent on fault type; vertical for bent shaft and worn gearbox but horizontal for cracked shaft and loose belt. Key findings contribute to shedding light on ideal sensor placements and frequency ranges for fault detection. These insights are crucial for shaping predictive maintenance strategies and real-time monitoring recommendations for research and industry.

Keywords: Bearing; Failure modes classification; Maintenance practices; Predictive analysis; Vibration analysis

1 Introduction

Predictive maintenance is the cornerstone of ensuring the reliability and longevity of machinery across various industries. One of the most significant techniques in this domain is vibration analysis. Vibration analysis is not only crucial for industries and plants that utilize rotational equipment but also plays a pivotal role in machinery health assessment, especially within industrial contexts [1].

This article underscores the importance of vibration analysis by utilizing advanced tools like the Machinery Fault Simulator (MFS) and Vibra-Quest (VQ) to meticulously scrutinize vibration patterns in mechanical parts. Through rigorous experiments, distinct vibration signatures for various mechanical faults were discerned. A clear correlation was established between the severity of a fault and the magnitude of vibration, especially at high operational speeds. Metrics such as the Crest Factor and Vibration Velocity have been identified as reliable indicators of machinery health.

In the broader context, the packaging materials industry has recognized the significance of Total Productive Maintenance, with vibration signal analysis being a primary method for early fault detection in rolling-element bearings [2]. Similarly, the structural integrity of machinery, especially those with bolted joints, is of paramount importance. Such joints, compared to welded structures, have been found to significantly influence the dynamic behavior of assembled structures [3].

The current research delves deeper into the significance of maintenance and vibration analysis, exploring its applications in various industries, the methodologies employed, and the benefits derived from its implementation. Emphasis will be placed on real-time monitoring using time and frequency domain analyses, and insights into ideal sensor placements and frequency ranges for fault detection.

The remainder of this paper will be presented as follows: Section 2 and 3 present the literature review and methodology respectively, section 4 discusses results and analysis, section 5 examines findings and section 6 concludes the paper.

2 Literature Review

The research indicates that maintenance costs in industrial plants can account for a significant portion, ranging from 15% to 60% of overall operating expenses [4]. However, companies often overlook the proper measurement of these costs. This oversight has led to the emergence of advanced maintenance strategies, such as Predictive Maintenance (PdM) and Condition-Based Maintenance (CBM), to address the challenges of effective maintenance management. Corrective Maintenance is a basic maintenance approach where repair actions take place only after a breakdown occurs. It can be cost-intensive, especially for large-scale applications, due to potential catastrophic failures and production capacity impact due to extended unplanned downtime [5]. Preventive Maintenance (PM) is considered a proactive approach based on scheduled maintenance tasks to prevent failures. PM takes place based on a periodic schedule carrying out pre-planned tasks. Tasks are mainly basic checks, lubrications, and the change of degrading parts. Even though the strategy prevents sudden unexpected failures, catastrophic failures might still occur. PM drawbacks can be expressed in the fact that it does not have a direct link to assets. As a result components and, the actual condition is not determined [6].

Total Productive Maintenance (TPM) is a maintenance strategy that originated in Japan in the 1970s, TPM is a holistic strategy to optimize equipment effectiveness and eliminate breakdowns. It involves all organizational levels to improve manufacturing equipment efficiency. TPM focuses on multiple areas that would impact the assets directly and indirectly including the surrounding environment, training, and staff engagement from all levels [7]. A different approach for maintenance is Condition-Based Maintenance (CBM), which originated in the 1940s, CBM has expanded across various industries, from railways to manufacturing, due to advancements in information technology [8]. CBM assesses real-time machine data to detect functional parameter fluctuations and potential abnormalities. It combines diagnostics, which handles fault detection, with prognostics, which manages fault prediction [9]. Vibration levels are pivotal in CBM, with the "bathtub curve" depicting the life of a machine [10]. Machine failure probability and time as a result of Condition-Based Maintenance (CBM) are depicted in Figure 2.3. Modern technological approaches over the past five decades have enabled condition monitoring, facilitating informed maintenance decisions [11]. CM involves measuring machine characteristics, such as motion and temperature, to identify potential defects. Comprehensive CM strategies consider various criteria, including machinery type and operating conditions [12]. Vibration measurement is a cornerstone of CM, with advanced systems relying on vibration sensors. The ISO establishes standards for equipment vibration, and thorough analysis of vibration data aids in assessing machine health [10].

Predictive Maintenance (PdM) has evolved significantly, becoming integral to Industry 4.0. This shift from reactive to predictive maintenance is driven by advancements in analytical, machine learning and data analysis applications [13]. PdM's primary objective is to foresee faults in deteriorating systems, leading to optimized maintenance interventions. Through PdM, early signs of potential problems are detected, resulting in timely maintenance actions. The data from these actions offers insights into the nature, location, and probable timing of system failures, ensuring proactive maintenance tasks [14].

Machine learning algorithms are central to PdM, automating fault detection and diagnosis, thereby reducing downtime, and maximizing industrial component use [13]. These algorithms are particularly adopted in analysing complex vibration signatures linked to specific faults. Vibration analysis, recognized for condition monitoring and fault diagnosis, has applications ranging from general machinery to specialized equipment like wind turbine drivetrains as an example [15]. Emerging techniques in PdM, such as convolutional neural networks and isolation forests, offer scalable solutions for fault diagnosis, capable of autonomously learning unique vibration signatures and enhancing predictive maintenance precision [16].

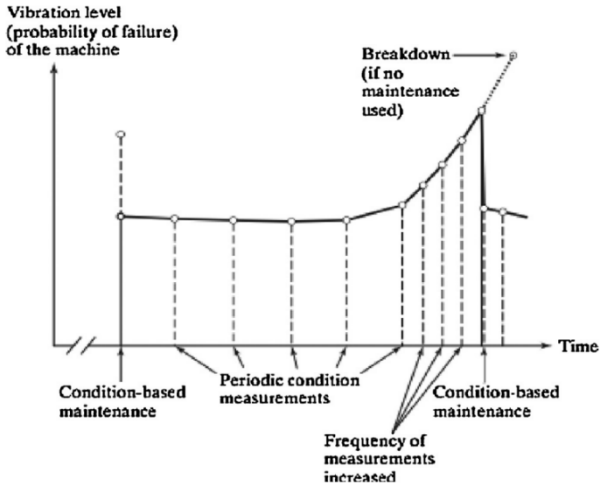


Figure 1: Condition-Based Maintenance [10]

2.1 Vibration Diagnostics and Analysis

Vibration is a fundamental phenomenon that occurs in both natural and man-made systems. It is the repeated movement of an object or system about an equilibrium point [17]. Vibration signatures, signals from machines under normal operations, are crucial in revealing the state of a mechanical system [18]. When a machine faces issues, these signatures change, showing variations in frequency, amplitude, or phase connections, crucial for detecting faults [11]. Vibration analysis emerged in the mid-20th century as a pivotal tool for machine condition monitoring. With advancements in sensor technology and signal analysis, it became essential for assessing rotating machinery health, and detecting potential defects [19][20]. Vibrations, oscillatory motions and their accompanying forces can strain machine components, affecting performance [17][21][22].

In condition monitoring (CM), vibration analysis is paramount. It aids in maintenance decisions by tracking real-time signals from machinery [23]. Vibration-based bearing CM has become integral to systematic maintenance over the years [24]. The varied vibration patterns under different fault conditions allow professionals to analyse a machine's internal workings without disassembly, making vibration analysis a top choice for machine CM [25].

2.2 The Pivotal Role of Vibration Analysis in Industrial Machinery Health and Safety

Vibration analysis is a premier technique for machinery assessment [26]. It offers benefits such as early fault detection and predictive maintenance [27]. This method not only reduces costs and downtime but also enhances equipment longevity and safety [14][28]. Its accuracy in fault identification is over 82% [29], but its effectiveness requires expert analysis, quality equipment, and a holistic maintenance strategy [12][26]. A Swedish case study on Vibration-Based Maintenance (VBM) highlighted its efficiency, revealing an annual maintenance profit of 3.58 million SEK (USD 0.358 million) and potential savings of SEK 30 million (USD 3 million), emphasizing its pivotal role in boosting operational efficiency [26].

Vibration analysis is pivotal in various industrial contexts. In pulp mills, it's essential for detecting bearing faults, with machinery signatures offering insights into system health [30]. Vibration analysis ensures the safety and performance of mechanical systems, such as the elevator chassis design [31].

2.3 Vibration Signal Analysis Techniques

Vibration analysis techniques such as Fourier analysis, time waveform analysis, and frequency spectrum analysis have been widely utilised to characterise the dynamic behaviour of machinery [10]. However, vibration analysis alone may not always provide an accurate picture of machinery health, particularly for defects that present subtle or non-vibrational symptoms [32].

2.4 Time Domain Analysis

Vibration signals from machinery are collected using systems with transducers like accelerometers [33]. These time domain signals, though rich in data, can be challenging to interpret due to noise. To aid in fault diagnosis, specific features are generated from these signals. Techniques such as time domain and frequency domain analyses are employed, with manual inspections being either visual or feature-based as shown in Figure 2 [5].

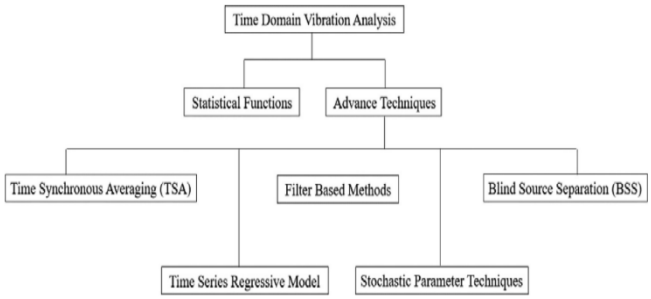


Figure 2: Time domain vibration analysis techniques [33]

Visual Inspection involves comparing the vibration signal of a machine to that of a healthy one, using tools like oscilloscopes [5]. However, this method can be imprecise for monitoring rotating machines due to various challenges like background noise and ambiguous signals [34]. Features-Based Inspection evaluates machine condition by examining key vibration signal features, often employing machine learning classifiers for accurate monitoring [35].

Statistical functions, also known as statistical techniques, are mathematical methods used to analyse vibration signals in the time domain. Some of which are presented in Fig. 3. They focus on signal amplitude and offer insights into the random behaviour of machine components [5].

Feature Name	Description	
	Brief Definition	Formula
Peak Amplitude (xp)	The maximum positive value of a vibration signal is referred to as the peak amplitude. It can also be defined as half of the difference between the maximum positive and maximum negative peak amplitudes of the vibration signal	$x_p = \frac{1}{2}[x_{max}(t) - x_{min}(t)]$
Mean Amplitude ($\bar{x}$)	Represents the average value of the vibration signal within a specific interval of sampling	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
RMS amplitude (RMS)	The RMS value rises progressively as a fault evolves. Yet, RMS doesn't offer insights into the initial stages of a fault, even though it escalates with fault progression.	$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert^2}$
Peak-to-Peak Amplitude (xp−p)	Refers to the difference between the highest point of positive amplitude and the lowest point of negative amplitude	$x_{p_p} = x_{max}(t) - x_{min}(t)$
Crest factor (CF)	calculates how much impact occur during the rolling element and raceway contact. CF is appropriate for "spiky signals	$x_{CF} = \frac{x_p}{x_{RMS}}$

Figure 3: Statistical functions with their mathematical expressions

2.5 Advanced techniques

Time Synchronous Averaging (TSA) was developed to extract periodic waveforms from noisy vibration signals, especially useful for monitoring gearboxes and bearings. Time series regression models, including AR, MA, ARMA, and ARIMA, utilize the Yule-Walker equation for parameter computation. Filter-based methods, such as the Prony model and adaptive noise cancellation, aim to minimize noise in vibration signals. Stochastic parameter techniques focus on the stochastic features of vibration signals for defect diagnosis, encompassing methods like correlation dimension and chaos theory. Lastly, the Blind Source Separation technique is employed to discern previously undetected signals from observed ones, crucial for identifying defects influenced by external factors like equipment noise [33].

2.6 Time Waveform Analysis in Vibration Situations

A time waveform displays a signal's variation over time, capturing its amplitude at each moment. Useful in low-speed operations and diagnosing bearing issues, gears, and machines with X-Y probes, it detects faults like looseness and rubs [17]. While versatile, time waveform analysis might not fully capture issues like unbalance in normal-speed machinery, necessitating combined spectrum and phase data for a holistic understanding [2].

2.7 Crest Factor (CF)

Crest Factor (CF) rises with peaks in a vibration signal, aiding in early failure detection and real-time monitoring [5]. A signal with a CF value exceeding 1.414 indicates an abnormality. Crest factor can indicate early machinery faults, but its detection can be overshadowed by other vibrations [36]. However, it is argued that the crest factor isn't effective for early detection in rolling element bearings [37][38].

Different research indicates that x_{CF} values above 4.5 suggest mechanical degradation, with gears in good condition having values between 3.5 and 4.0 [20]. It has been found that defective bearings have higher x_{CF} values than healthy ones [39]. An emphasis has been made that the vulnerability of mechanical equipment's rolling bearings to failure attributes to the rotation of bearing components [40][41]. Lastly, it has been argued that the crest factor of vibrations varies with the rotational speed of gearboxes [42].

2.8 Vibration Velocity (Vrms)

Rresearch suggests employing statistical characteristics of vibration signals, such as peak value, peak-to-peak value, root-mean-square (RMS), and crest factor, to assess machinery condition [43] and peak levels can fail to serve as a reliable indicator of damage in continuous-operation systems, owing to the absence of a statistical foundation [20]. RMS (Root Mean Square) values, on the other hand, thought to be more appropriate for long-term applications. The research suggest that the signals recorded in this context are typically the vibration velocities from the machine's bearing cover [4]. The subsequent comparisons correspond with recognised vibration standards such as ISO 10186 to determine the machine's RMS vibrating severity level, which aids in the identification of potential malfunctions or unsafe operating conditions [44].

The findings presented previously support previous research in the field of Root Mean Square (RMS) vibration monitoring. It has been noted that while RMS-based methods are effective in detecting gradual mechanical issues such as bearing pitting and shaft cracks, they have certain limitations when used to monitor gearbox conditions [44]. One of the major flaws is that RMS values are not sensitive enough to detect early signs of gear tooth problems because they do not respond quickly to isolated signal peaks [44]. Furthermore, brief, weak bursts of vibration have little effect on RMS readings, making early detection of bearing failures difficult. Because of these limitations, the authors recommend using peak vibration values as a more sensitive alternative for condition monitoring [44].

2.9 Frequency Domain Analysis

Frequency Domain Analysis is a prevalent vibration analysis technique used for machine monitoring. It reveals details based on frequency features not easily detected in the time domain. Rotating components produce time domain vibration signals, and their frequency spectrum helps identify vibration sources. Fourier analysis techniques, introduced by J. Fourier in the 19th century, are essential for these frequency representations. Variations in amplitudes can indicate machinery defects and thorough data analysis provides insights into equipment health, guiding timely maintenance [21][45]. As shown in Figure 4, these frequencies are shown as a sum of the signals.

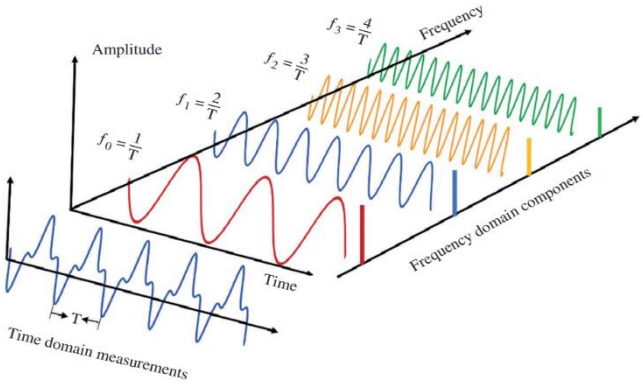


Figure 4: Measurements in the time domain against measurements in the frequency domain [46]

2.10 Fourier Analysis and its Applications in Vibration Signal Analysis

The Fourier Series decomposes periodic signals into sinusoidal components, aiding in identifying frequencies within a signal. However, for non-periodic vibration signals, other methods like the Fourier integral are more suitable [5]. The Discrete Fourier Transform (DFT) analyses discrete signals but is computationally intensive for large datasets. In contrast, the Fast Fourier Transform (FFT) efficiently computes the DFT and is widely used in machine diagnostics, especially for vibration signal analysis [5][10]. Prominent frequency peaks in the frequency spectrum often correspond to defective bearings, with the number of harmonics indicating fault severity [47]. Understanding vibrational patterns at varying speeds is crucial for accurate diagnostics, with distinct peaks at characteristic fault frequencies signalling bearing defects [48][49].

Signal Processing Techniques extract vital information from vibration data. The frequency spectrum displays signal energy distribution across frequencies, providing insights into individual frequency components [2][50]. While spectral analysis is standard for diagnosing gearboxes, its complexity can hinder fault detection in intricate gear systems [10]. The power spectrum, derived from the spectrum's quadratic amplitude, effectively detects machinery defects under varying conditions [51]. Lastly, Frequency Response Analysis (FRA) studies a system's response to different input signal frequencies, revealing resonance peaks and system characteristics [32].

2.11 Vibration Increment

offers real-time monitoring of machinery health by tracking changes in vibration patterns, rather than just static measurements. This method effectively indicates fault progression, especially in machinery prone to wear, allowing timely interventions. It enhances predictive maintenance accuracy by highlighting when faults need urgent attention [52]. Moreover, sensor placement is vital for data precision. For instance,

horizontal sensors excel in detecting intact bearings, while vertical ones are better for outer race anomalies [53].

2.12 Vibration Signatures of Bearing Failure Modes

Eccentricity: described as a condition where the centre of rotation of rotating elements deviates from the geometric centre [18]. This deviation results in vibrations that are similar to those caused by an unbalanced rotor. The most dominant vibration peak due to eccentricity is observed at 1X in the radial direction.
Looseness: When belts wear out or become loose, the frequency spectrum displays harmonics that correspond to the belt's frequency. Manufacturing defects, such as splices in the belt, can significantly contribute to these vibrations. Such defects can lead to increased dynamic stresses in the belt [54].
Bent Shaft: A bent, or misaligned shaft occurs when the rotating centres of machinery shafts don't align perfectly. Shafts can bend due to various reasons, often influenced by thermal gradients across the shaft. This bending can either be temporary or permanent [18].
Defected Bearings: Bearings, even when new, emit specific vibration frequencies. When a defect or damage occurs in the bearing components, the vibration spectrum undergoes changes, revealing the natural frequency of the bearing. This phenomenon is crucial for detecting bearing faults early on [55].
Cracked Shaft: Powerful machinery with rotating shafts often faces challenging working conditions, making them susceptible to cracks. It has been emphasized that these cracks introduce additional frequency components in the system's vibration response, which can be used for early detection [56].
Worn Gearboxes: Gear wear is a natural occurrence over time, impacting the durability and overall lifespan of the gear transmission system. Gear wear could lead to severe failures, emphasizing the importance of monitoring gear wear through vibration analysis for effective predictive maintenance [57].

2.13 Advantages and Limitations of Vibration Signal Monitoring

Vibration sensors offer a plethora of advantages. Research highlights that these sensors are non-intrusive and are particularly effective for monitoring the conditions of rotating machines [25]. They can detect up to 80% of common issues, such as misalignment and unbalance. Modern devices even come equipped with accelerometers, making vibration monitoring more accessible. However, these sensors come with a set of challenges. Different sensors have distinct operating characteristics, and improper mounting can lead to the masking of essential vibration signal features. It's crucial to understand the physical characteristics of vibration sources and review datasheets when selecting a sensor [25].

3 Methodology

In the conducted study, a positivist research philosophy was adopted, emphasizing controlled settings and statistical analyses to explore the relationship between vibration signatures and faults. The research approach and strategy were logical and experimental, utilizing established theories to guide data collection and hypothesis testing. Quantitative methods are primarily used, focusing on analysing vibration signal data to discern patterns among fault situations. The foundational literature review provided a comprehensive understanding of Maintenance Strategies, condition monitoring, and the intricacies of Vibration standards. This review also highlighted the potential of vibration analysis in predictive maintenance and early fault detection.

In the article, primary data collected through experimental work, focusing on analysing vibration signatures for fault diagnosis in engineering systems. The aim was to inform preventive maintenance strategies. Data collection was comprehensive, supported by an extensive literature review from notable sources like Springer, IEEE, and Elsevier. The Machinery Fault Simulator (MFS) was employed to gather genuine vibration signals from components. The recorded vibration data, especially vibration velocity (Vrms), was compared to ISO standards, specifically ISO10816 and ISO13373, to determine vibration severity levels. This comparison is pivotal for early defect detection and predictive maintenance. The

article utilized tools like Vibra-Quest (VQ), MATLAB (R2023a), and Excel for data analysis, aiming to offer a systematic approach to studying vibration signatures under different fault conditions, thereby enhancing the reliability of the findings.

3.1 Experiment Setup

On each of the 7 tests, the machine was reset with the necessary test components and began gathering vibration measurements via 4 sensors Channels 1 and 2 are horizontal, and channels 3 and 4 are vertical, at a shaft rotation speed of 10Hz, 20Hz, or 30Hz, which the testing centre offers.

3.2 Benchmark

A major part of this experimental work was the creation of a benchmark, which served as a standardised reference point for evaluating the performance of various malfunctioning systems. The machine was set up with fault-free components and then started collecting vibration readings to establish this benchmark. These meticulously documented data points depict the system's ideal, error-free behaviour. The benchmarked data is then kept for later use as a baseline for comparison when performing trials with broken parts. Figure 5 shows the Multi-Belt Drive, belt tensioner, ¾" bearings, ¾" shaft, router, ¾" coupler, and reciprocating mechanism included in the kit.

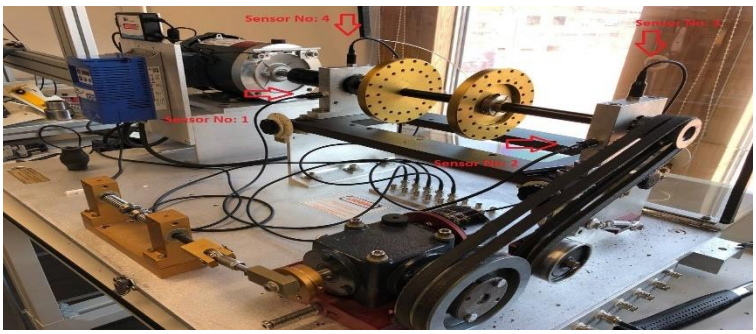


Figure 5: Experiment setup for obtaining vibration data of Benchmark

3.3 Coupling End Bent Shaft

The experiment seeks to analyse complex vibration patterns caused by gyroscopic effects. It will explore the difficulties of balancing an overhung rotor on a bent shaft and address alignment issues. The kit includes a ¾" shaft with a ~0.010" bend at the coupling end, along with a ¾" bore loader weighing 11lb (5kg) attached to the shaft, as shown in Figure 6.



Figure 6: Experiment setup for obtaining vibration data of Coupling End Bent Shaft

3.4 Cracked Shaft

The experiment explores crack effects on natural frequencies and vibration behaviour, developing an early-stage diagnostic technique. It investigates crack propagation, utilizing advanced signal processing. The experiment kit includes steel shafts: one with a 4 ½'' 4-bolt flange simulating crack, one with slit crack and filler, and one with a deep V-notch crack, as shown in Figure 7.

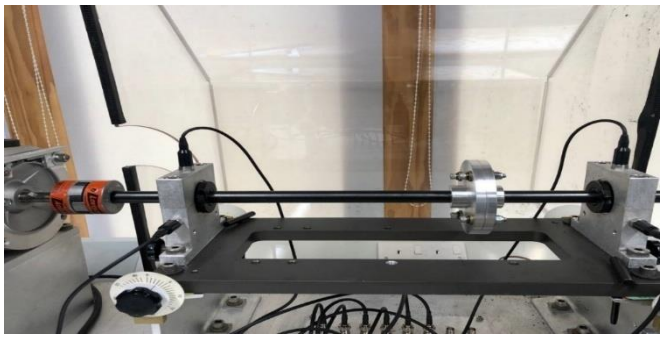


Figure 7: The experiment setup for obtaining vibration data of Cracked Shaft

3.5 Eccentric Sheave

This study examines eccentric sheave effects on vibrations of the bearings, as depicted in Figure 8. The kit includes a single eccentric sheave for experimentation. Along with multi-belt Drive, belt tensioner, ¾" bearings, ¾" shaft, router, coupling ¾" and reciprocating mechanism.

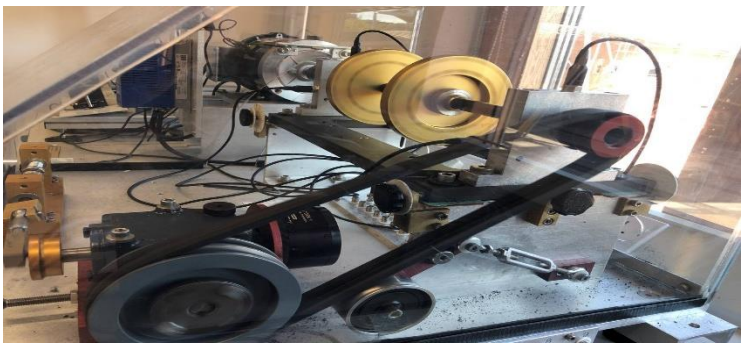


Figure 8: The experiment setup for obtaining vibration data of Eccentric Sheave.

3.6 Defected Bearing

This experiment involves conducting tests with progressively severe defects to understand the waveforms and spectra associated with common bearing issues. The kit includes specimens with distinct defects: inner race defect, outer race defect, ball defect, and a combination of defects, along with a 3/4" bore loader weighing 11lb (5kg) attached to the shaft, as shown in Figure 9.

3.7 Loose Belt

In this experiment, the belt tensioner will intentionally be set loose to assess the influence of a loose belt on vibration levels. The study aims to comprehend the effects of sheave misalignment and belt tension, including belt fault frequencies. The provided kit encompasses crucial components: two V-belts, two

double groove sheaves, a rolling tensioner, a three-way oil-lubricated gearbox with straight-cut bevel gears, cup and cone roller bearings, and a manually adjustable magnetic brake, as shown in Figure 10.

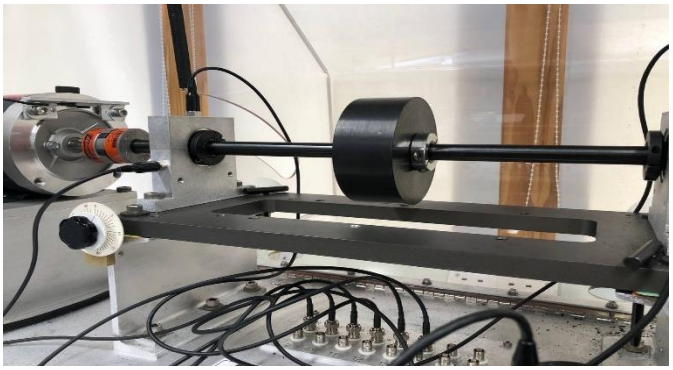


Figure 9: The experiment setup for obtaining vibration data of Defected Bearing.

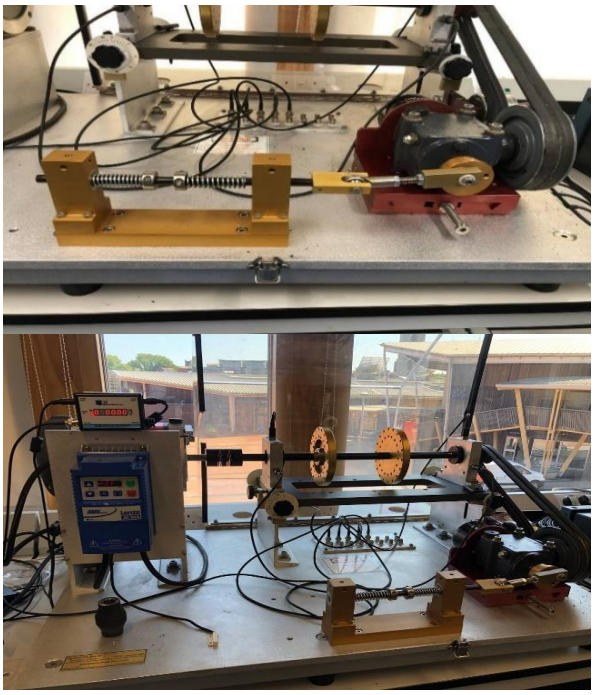


Figure 10: The experiment setup for obtaining vibration data of Loose Belt.

3.8 Direct Driven Gearbox

This research investigates vibration signatures of defective straight-tooth gearbox pinions, focusing on isolated analysis without belt, shaft, or bearing influences. The provided kit includes all essential hardware for direct mounting of the gearbox to the AC motor. Additionally, it features a reciprocating mechanism equipped with two springs, adjustable spring engagement timing, and two-stroke settings, Figure 11 illustrates the experiment setup.

3.9 Measurement

The MFS, equipped with a Multi-channel DAQ and VibraQuest system, was used for data acquisition

and analysis. Data was further analyzed using MATLAB. Vibration data was captured up to a frequency limit of 2000 Hz with 1600 lines of spectral resolution, focusing on various vibration characteristics like RMS and Vrms. The author set up the machine based on the MFS by Spectra-Quest manual, running it at speeds of 10Hz, 20Hz, and 30Hz, and collected data from four sensors placed at different positions. This helped in understanding the vibration patterns of different mechanical faults. Experiments were conducted at room temperature on separate days to minimize temperature effects on the data, as the temperature can influence vibration behaviors [58].

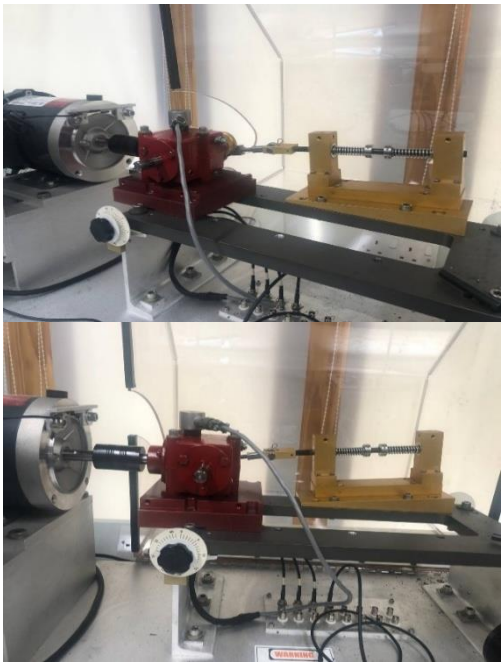


Figure 11: The experiment setup for obtaining vibration data of Direct direct-driven Gearbox with an internal worn gear tooth.

4 Results

This section presents experimental results.

4.1 Time domain analysis

This section delves into vibration attributes in machinery, emphasizing Time and Frequency Domain analyses. Tests in the Time Domain are conducted to identify vibration features and machine behavior in various configurations. Concurrently, dominant frequencies in vibration signals are identified in the Frequency Domain. Together, these evaluations offer comprehensive insights into machine performance and predictive maintenance opportunities.

MATLAB's Signal Processing Toolbox and Signal Analyzer were used to analyse vibration signals from the Machine Fault Simulator (MFS). Seven tests were conducted on various machine configurations. Before each test, the machine was adjusted and set in motion. Vibration data collected from four sensors at shaft rotation speeds of 10Hz, 20Hz, and 30Hz, starting from a standstill. The data is displayed in graphs, with the x-axis representing time and the y-axis indicating amplitude. Each graph, as shown in Figures 12 to 15, is labelled from "Channel 1" to "Channel 4" to identify the sensor. These visualizations provide a detailed view of the machine's vibration patterns under different conditions, aiding in further analysis and enhancements.

4.2 Extracted Vibration Characteristics in Time Domain

Fault feature extraction is essential for fault detection and classification [59]. These features are particularly effective in identifying issues in tapered roller bearings [60]. Analysis done on vibration signals, RMS, and peak values to detect defects, emphasized that monitoring RMS and extreme values can predictively detect issues [44]. This approach offers a concise representation of vibration data, enhancing analysis accuracy. A comparison of distinct fault features at three rotational frequencies is presented in the Tables below.

Channel 1	Vibrational Characteristics														
	Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Min Value	Max value	Mean	Peak to Peak	RMS	Min Value	Max value	Mean	Peak to Peak	RMS	Min Value	Max value	Mean	Peak to Peak	RMS
Benchmark	-0.00130	0.00239	0.00055	0.00369	0.00119	-0.00156	0.00282	0.00061	0.00437	0.00122	-0.00172	0.00304	0.00067	0.00476	0.00125
Coupling End Bent Shaft	-0.00147	0.00215	0.00036	0.00362	0.00112	-0.00152	0.00227	0.00041	0.00379	0.00114	-0.00143	0.00231	0.00042	0.00374	0.00114
Cracked Shaft	-0.00141	0.00220	0.00040	0.00361	0.00113	-0.00141	0.00226	0.00041	0.00367	0.00114	-0.00141	0.00235	0.00048	0.00376	0.00116
Eccentric Sheave	-0.00121	0.00241	0.00057	0.00363	0.00120	-0.00136	0.00242	0.00053	0.00378	0.00118	-0.00151	0.00274	0.00059	0.00425	0.00121
Fault Bearing	-0.00146	0.00247	0.00046	0.00393	0.00117	-0.00225	0.00335	0.00047	0.00560	0.00119	-0.00398	0.00425	0.00039	0.00823	0.00120
Loose Belt	-0.00124	0.00241	0.00059	0.00364	0.00121	-0.00132	0.00240	0.00055	0.00373	0.00119	-0.00156	0.00280	0.00064	0.00436	0.00124
Warn Gearbox Tooth	-0.00537	0.00594	0.00064	0.01131	0.00140	-0.00794	0.00915	0.00056	0.01709	0.00183	-0.01182	0.01102	0.00070	0.02284	0.00252

Figure 12: Vibrational Characteristics in time domain channel 1

Channel 2	Vibrational Characteristics														
	Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Min Value	Max value	Mean	Peak to Peak	RMS	Min Value	Max value	Mean	Peak to Peak	RMS	Min Value	Max value	Mean	Peak to Peak	RMS
Benchmark	-0.08875	0.09294	0.00257	0.18169	0.00992	-0.16341	0.20624	0.00095	0.36964	0.02478	-0.22926	0.29500	0.00261	0.52426	0.04142
Coupling End Bent Shaft	-0.01104	0.01254	0.00066	0.02358	0.00258	-0.02396	0.02534	0.00057	0.04930	0.00602	-0.04977	0.04447	0.00081	0.09424	0.01226
Cracked Shaft	-0.00835	0.01075	0.00077	0.01909	0.00221	-0.02767	0.03508	0.00076	0.06275	0.00733	-0.04775	0.07165	0.00076	0.11940	0.01225
Eccentric Sheave	-0.04600	0.04992	0.00271	0.09592	0.00760	-0.09324	0.08351	0.00122	0.17675	0.01951	-0.18855	0.16632	0.00142	0.35486	0.04572
Fault Bearing	-0.31434	0.34374	0.00095	0.65808	0.02482	-0.88807	0.60832	0.00135	1.49639	0.06910	-1.06700	2.14188	-0.00669	3.20889	0.30478
Loose Belt	-0.03404	0.03830	0.00169	0.07234	0.00615	-0.06208	0.08351	0.00227	0.14559	0.01487	-0.26418	0.21541	0.00215	0.47959	0.04201
Warn Gearbox Tooth	-0.39449	0.37128	0.00038	0.76577	0.04460	-0.64711	0.55211	0.00063	1.19922	0.10214	-0.65749	0.78344	0.00071	1.44093	0.14851

Figure 13: Vibrational Characteristics in time domain channel 2

Channel 3	Vibrational Characteristics														
	Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Min Value	Max value	Mean	peak to peak	RMS	Min Value	Max value	Mean	peak to peak	RMS	Min Value	Max value	Mean	peak to peak	RMS
Benchmark	-0.12940	0.11482	0.00233	0.24422	0.01085	-0.21465	0.16096	0.00082	0.37561	0.02486	-0.27413	0.20921	0.00171	0.48334	0.03914
Coupling End Bent Shaft	-0.01389	0.01307	0.00040	0.02696	0.00310	-0.03912	0.03339	0.00040	0.07251	0.00858	-0.07629	0.08443	0.00066	0.16072	0.01809
Cracked Shaft	-0.00870	0.01025	0.00053	0.01895	0.00216	-0.02689	0.02688	0.00066	0.05377	0.00580	-0.05048	0.05524	0.00072	0.10572	0.01073
Eccentric Sheave	-0.09672	0.06995	-0.00105	0.16667	0.00939	-0.11728	0.08960	0.00114	0.20687	0.01747	-0.22314	0.17728	0.00132	0.40042	0.03759
Fault Bearing	-0.11826	0.10181	0.00073	0.22007	0.01243	-0.28137	0.73944	0.01918	1.02081	0.08785	-0.41704	0.61395	0.06412	1.03098	0.12891
Loose Belt	-0.04729	0.04740	0.00170	0.09469	0.00584	-0.13647	0.10919	0.00193	0.24566	0.01698	-0.24970	0.21061	0.00172	0.46031	0.03675
Warn Gearbox Tooth	-0.18459	0.18595	0.00056	0.37054	0.02113	-0.30005	0.28126	0.00072	0.58130	0.05197	-0.45416	0.40954	0.00074	0.86370	0.07953

Figure 14: Vibrational Characteristics in time domain channel 3

Channel 4	Vibrational Characteristics														
	Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Min Value	Max value	Mean	peak to peak	RMS	Min Value	Max value	Mean	peak to peak	RMS	Min Value	Max value	Mean	peak to peak	RMS
Benchmark	0.00118	0.01744	0.00850	0.01626	0.00910	-0.00397	0.00441	0.00067	0.00837	0.00143	-0.00047	0.02744	0.00974	0.02790	0.01110
Coupling End Bent Shaft	-0.00226	0.00297	0.00064	0.00523	0.00136	-0.00262	0.00251	0.00029	0.00514	0.00123	-0.00279	0.00327	0.00078	0.00607	0.00144
Cracked Shaft	-0.00220	0.00396	0.00149	0.00615	0.00193	-0.00242	0.00356	0.00107	0.00598	0.00162	-0.00223	0.00358	0.00099	0.00581	0.00157
Eccentric Sheave	0.00197	0.07706	0.01913	0.07509	0.02491	-0.00242	0.00587	0.00200	0.00829	0.00240	-0.00271	0.00771	0.00278	0.01042	0.00318
Fault Bearing	-0.00263	0.00553	0.00092	0.00616	0.00154	-0.00289	0.00348	0.00059	0.00637	0.00135	-0.00294	0.00403	0.00088	0.00697	0.00152
Loose Belt	-0.00099	0.00815	0.00380	0.00913	0.00415	-0.00074	0.01405	0.00610	0.01479	0.00670	-0.00112	0.01416	0.00586	0.01529	0.00646
Warn Gearbox Tooth	-0.00223	0.00395	0.00138	0.00618	0.00185	-0.00214	0.00442	0.00154	0.00656	0.00200	-0.00221	0.00439	0.00144	0.00660	0.00193

Figure 15: Vibrational Characteristics in time domain channel 4

In Figure 12, at 10 Hz, the Benchmark recorded a mean vibrational value slightly higher than the Fault Bearing. However, the Fault Bearing showed more pronounced vibrational fluctuations, indicated by its higher Peak-to-peak value, even though its average vibration was lower. RMS values for both were closely matched. As the shaft's rotational speed increased, the amplitude and variability of its vibrations also rose. At 10 Hz, the loose belt and eccentric sheave fault were significant, but the worn gearbox surpassed them. The defective gearbox showed increasing vibrational disturbances with rising rotational speed, especially evident at 30Hz, where its RMS value was more than double compared to other defects. This trend suggests that the worn gearbox fault has a substantial impact on the machinery's vibration, aligning with findings from previous studies [61].

Figure 13 provides a comparative analysis of fault characteristics across three rotational frequencies for Channel 2. At 10Hz, the Fault Bearing and Warn Gearbox Tooth exhibit pronounced vibrations, with the Fault Bearing showing the highest values. As the speed doubles, the vibrational differences across fault types become more evident, with the Fault Bearing again displaying the most significant vibrations. At 30Hz, the Fault Bearing's vibrational measures are the highest. Excluding the Benchmark, Fault Bearing, and Warn Gearbox Tooth, the Eccentric Sheave and Loose Belt show the highest vibrational magnitudes at 10Hz and 20Hz, respectively. At the highest speed, the Loose Belt and Eccentric Sheave display notable vibrations, with the Cracked Shaft also showing an increase.

Figure 14 compares vibrational values of different faults at varying rotation speeds. At 10Hz, the Eccentric Sheave and Loose Belt have the highest vibrational magnitudes, with the Cracked Shaft showing notably lower values, consistent with research findings [62]. At 20Hz, the Eccentric Sheave's vibration remains high, while the Loose Belt's metrics are consistent with its 10Hz readings [54]. The Cracked Shaft displays minimal disturbance. At 30Hz, the Loose Belt's vibration is dominant, with the Eccentric Sheave slightly lower. Interestingly, the Cracked Shaft's vibration increases as speed rises, aligning with research observations [54].

Regarding Figure 15, at 10Hz rotational speeds, the Cracked Shaft and Eccentric Sheave show significant vibrational metrics, with the latter being the most dominant fault. The Loose Belt and Warn Gearbox Tooth also display notable vibrations. At 20Hz, the Loose Belt and Eccentric Sheave are prominent, while the Cracked Shaft and Coupling End Bent Shaft have moderate vibrations. At 30Hz, the Cracked Shaft's vibrations remain steady, but the Eccentric Sheave and Loose Belt dominate. The Warn Gearbox Tooth consistently presents the highest RMS across all speeds. The data underscores the importance of understanding vibrational characteristics at varying speeds for accurate machinery diagnosis and maintenance [63].

4.3 The Crest Factor (xCF)

To further substantiate the findings, both RMS and Peak values were extracted from the spectral data, and corresponding Crest Factor (xCF) values were calculated. These computations are detailed in Table 1. The first thing to notice is the variability of xCF values across different channels for each fault type. For example, the Coupling End Bent Shaft shows a minimum xCF value of 2.0249 in CH1 and peaks at 4.6664 in CH3. This represents more than a 100% increase from the minimum to maximum xCF values for this particular fault. The Cracked Shaft presents a similar but even more drastic scenario, with CH2 peaking at an xCF value of 5.8509 compared to 2.0140 at CH1, a nearly 190% increase.

Within this context, the author's observations, as seen in Table 1, of the Worn Gearbox Tooth fault are alarming. The fault consistently displayed x_{CF} values ranging from 4.3626 at CH4 to 5.2752 at CH2 across all monitored channels. Moreover, the Defected Bearing fault type stands out for its particularly high Crest Factor (x_{CF}) values, notably a reading of 7.0276 at CH2. As detailed in Table 1, this fault type displays a wide spectrum of x_{CF} values across channels, ranging from a low of 2.6506 at CH4 to a high at CH2.

Table 1: Crest Factors comparison

Fault type	XCF			
	CH1	CH2	CH3	CH4
Coupling end bent shaft	2.0249	3.6266	4.6664	2.2726
Cracked shaft	2.0140	5.8509	5.1496	2.2785
Eccentric sheave	2.2556	3.6376	4.7160	2.4251
Defected bearing	3.5474	7.0276	4.7626	2.6506
Loose belt	2.2517	5.1270	5.7304	2.1922
Worn gearbox tooth	4.3673	5.2752	5.1498	4.3626

4.4 Vibration Velocity (Vrms)

In light of the literature discussed previously, the research has focused on calculating, the Vibration Velocity (Vrms) has been calculated as shown in Table 2 and found that the defective bearing has the highest Vrms value of (0.3048 mm/s).

Table 2: Vibration velocity comparison

Fault type	Vrms			
	CH1	CH2	CH3	CH4
Coupling end bent shaft	0.0011	0.0123	0.0181	0.0014
Cracked shaft	0.0012	0.0122	0.0107	0.0016
Eccentric sheave	0.0012	0.1457	0.0376	0.0032
Defected bearing	0.0012	0.3048	0.1289	0.0015
Loose belt	0.0012	0.0420	0.0368	0.0065
Worn gearbox tooth	0.0025	0.1485	0.0795	0.0019

In addition to the Vrms values in Table 2, and compare with Figure 16, which illustrates the ISO108016 and ISO13373, Standards for monitoring non-rotating machines and machines' internal parts respectively. Which includes Vibration velocity range and classification of machine classes, providing further context for understanding the severity of machinery health conditions.

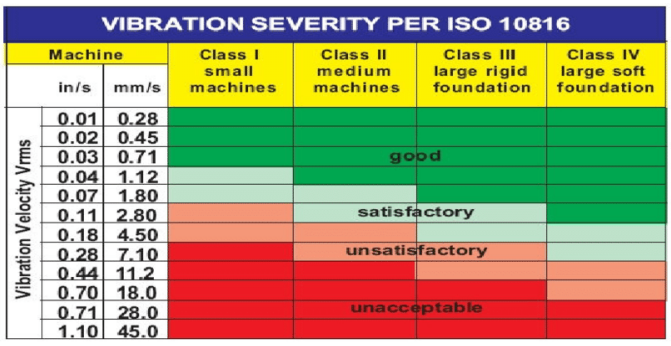


Figure 16: ISO10816 Standard for monitoring non-rotating machines and machines' internal parts, respectively

4.5 Frequency Domaine

To enhance the analysis, the five most pronounced peaks in the frequency spectrum for each channel were identified using the MATLAB Signal Processing Toolbox and the Signal Analyzer, which include performing Fast Fourier transform (FFT) for the vibration data. These peaks, based on amplitude, represent the dominant frequencies where the vibrations are most intense. For each channel, the frequencies and their corresponding amplitudes of these peaks are presented in a table. This organized representation simplifies the process of identifying and contrasting the predominant frequencies across channels as shown in Figures 17 to 20. Additionally, these pivotal peaks are demarcated with red circles on the frequency amplitude plot, offering a distinct visual reference of the primary frequencies.

4.6 Extracted Vibration Characteristics in Frequency Domain

Using the Statistics, Signal Processing, and Machine Learning Toolbox™ in MATLAB™, the Crest factor (x_{CF}) and Vibration Velocity (Vrms) were calculated for each channel at the highest speed of 30hz. This speed was chosen because of its alignment with typical machinery speeds and the speed limitation of the MFS.

Channel 1		Harmonic Vibrations														
		Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Frequency vs Amplitude	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X
Benchmark	Frequency(Hz)	50	149.84375	449.5313	349.6875	249.84375	50	149.84375	449.375	249.6875	349.5313	50	149.84375	449.5313	249.6875	349.6875
	Amplitude	0.0012179	0.0003997	8.25E-05	7.13E-05	5.688E-05	0.0009822	0.0001268	0.0001008	7.813E-05	6.65E-05	0.0011797	0.0004104	0.000104	7.069E-05	6.36E-05
Coupling End Bent Shaft	Frequency(Hz)	50	149.84375	449.5313	249.6875	849.0625	50	149.84375	249.84375	349.84375	449.6875	50	149.84375	449.375	249.6875	349.5313
	Amplitude	0.001151	0.0004068	9.12E-05	7.6E-05	4.949E-05	0.0013637	0.0002622	7.972E-05	6.601E-05	6.22E-05	0.0009664	0.00031	8.97E-05	6.847E-05	6.13E-05
Cracked Shaft	Frequency(Hz)	49.84375	149.6875	249.5313	449.0625	349.375	50	149.84375	249.6875	449.5312	349.5313	50	149.84375	449.5313	349.6875	249.6875
	Amplitude	0.0011006	0.0003917	7.99E-05	6.42E-05	6.026E-05	0.0010994	0.0003903	7.69E-05	6.471E-05	6.13E-05	0.0012016	0.0004047	9.12E-05	6.518E-05	5.85E-05
Eccentric Sheave	Frequency(Hz)	50	150.15625	450.4688	250.3125	350.46875	50	150	250.15625	350.15625	450.1563	50	150.15625	250.1563	450.3125	350.1563
	Amplitude	0.0011181	0.0003983	8.57E-05	8.11E-05	5.592E-05	0.0013544	0.0002644	8.408E-05	6.664E-05	6.57E-05	0.001351	0.000268	8.32E-05	7.129E-05	6.67E-05
Fault Bearing	Frequency(Hz)	49.84375	149.6875	449.0625	249.5313	349.21875	49.84375	149.6875	449.21875	249.5325	349.375	50	149.84375	449.6875	249.84375	349.6875
	Amplitude	0.0011656	0.0004147	9.33E-05	7.57E-05	4.794E-05	0.001084	0.0003844	8.071E-05	7.979E-05	7.09E-05	0.0013365	0.0003399	9.76E-05	8.171E-05	4.81E-05
Loose Belt	Frequency(Hz)	50	150	449.8438	349.8438	249.84375	50	150.15625	450.46875	250.3125	350.4688	50	150	450.1563	250.15625	350.1563
	Amplitude	0.001395	0.0003497	9.68E-05	6.26E-05	5.466E-05	0.0011257	0.0004033	8.571E-05	8.359E-05	5.17E-05	0.0013715	0.000295	8.59E-05	7.543E-05	7.1E-05
Worn Gearbox Tooth	Frequency(Hz)	50.15625	150.3125	450.9375	250.4688	1262.0313	50.15625	150.3125	1343.2813	1935.9375	1205	50.15625	150.3125	1270.625	1290.4688	1230.938
	Amplitude	0.0012244	0.0004132	7.71E-05	6.82E-05	5.703E-05	0.0010803	0.0006853	0.0001383	0.0001027	9.89E-05	0.0010063	0.0003861	0.00022	0.0001811	0.00018

Figure 17: Vibration characteristics in frequency domain channel 1

Channel 3		Harmonic Vibrations														
		Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Frequency vs Amplitude	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X
Benchmark	Frequency(Hz)	60	683.75	677.34375	665.15625	695.9375	676.09375	521.5625	540.9375	425	695.46875	469.21875	674.53125	498.59375	703.90625	557.1875
	Amplitude	0.0009493	0.0007101	0.0007014	0.0006969	0.0006941	0.00499	0.0043457	0.0039914	0.0039492	0.0039392	0.0138982	0.0122372	0.0103121	0.0098975	0.0095983
Coupling End Bent Shaft	Frequency(Hz)	60	80.9375	69.84375	201.5625	593.125	19.84375	79.375	586.71875	704.375	39.6875	29.84375	92.1875	294.0625	59.6875	1941.25
	Amplitude	0.0012389	0.0004227	0.0004058	0.0003922	0.000378	0.0040077	0.0019115	0.001464	0.0012726	0.0012298	0.0107838	0.0060283	0.0028098	0.0027125	0.0023167
Cracked Shaft	Frequency(Hz)	280.3125	55.9375	272.34375	63.75	241.40625	843.75	865.59375	902.03125	884.6875	882.1875	60	294.6875	29.84375	59.6875	302.1875
	Amplitude	0.0005331	0.0003946	0.0002571	0.0002417	0.0002254	0.000945	0.0008123	0.0007346	0.0006894	0.0006256	0.0031411	0.0028744	0.0022369	0.0021723	0.0013365
Eccentric Sheave	Frequency(Hz)	0.15625	9.21875	60	691.25	18.4375	384.84375	19.21875	654.21875	250.15625	673.4375	672.5	584.6875	497.03125	233.90625	29.21875
	Amplitude	0.0054046	0.0013204	0.0007977	0.0005288	0.0005149	0.0057772	0.005129	0.0044605	0.0032088	0.003043	0.0120727	0.00919	0.0091477	0.0086574	0.0078531
Fault Bearing	Frequency(Hz)	60	451.25	458.75	599.375	247.1875	0.3125	1.25	1.71875	2.5	4.6875	0.625	1.40625	2.65625	2.34375	4.0625
	Amplitude	0.0019238	0.0009025	0.0008513	0.0008268	0.0008238	0.0463904	0.0269136	0.0186715	0.0094059	0.0075778	0.0548378	0.0358901	0.0150771	0.014742	0.0106846
Loose Belt	Frequency(Hz)	60	28.90625	440.46875	435.3125	396.71875	382.65625	510.46875	528.90625	431.875	531.25	532.8125	473.59375	414.375	592.03125	505
	Amplitude	0.0015367	0.0009821	0.0005132	0.0004983	0.000459	0.0040167	0.0038008	0.0037051	0.0035872	0.0031493	0.0161247	0.0120864	0.0097261	0.0093793	0.0079005
Worn Gearbox Tooth	Frequency(Hz)	195.78125	352.34375	339.21875	1262.0313	332.8125	658.4375	1843.75	1856.875	1935.9375	1995.1563	19.84375	1230.9375	1270.625	1290.4688	1191.25
	Amplitude	0.0026304	0.0021472	0.0017283	0.0017211	0.0016969	0.0078536	0.0072782	0.0061799	0.0057016	0.0055953	0.011191	0.0107779	0.0087699	0.0077621	0.007366

Figure 18: Vibration characteristics in frequency domain channel 3

Channel 2		Harmonic Vibrations														
		Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Frequency vs Amplitude	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X
Benchmark	Frequency(Hz)	140.625	132.8125	102.34375	99.375	136.09375	136.40625	139.53125	734.0625	753.4375	133.28125	90	762.5	144.53125	674.53125	703.90625
	Amplitude	0.0015198	0.000771	0.00073	0.0006853	0.0006824	0.0036022	0.002943	0.0029155	0.0026571	0.0025782	0.009413	0.0079695	0.0072973	0.0072268	0.0071496
Coupling End Bent Shaft	Frequency(Hz)	9.84375	60	311.40625	301.40625	291.5625	19.84375	1133.75	1135.1563	139.84375	1095.4688	29.84375	151.875	244.21875	92.1875	184.53125
	Amplitude	0.0005699	0.0003875	0.0003782	0.0003559	0.0003544	0.002685	0.0013663	0.0012647	0.0009454	0.0008254	0.0074037	0.0044757	0.0032803	0.0028653	0.0021928
Cracked Shaft	Frequency(Hz)	280.3125	47.96875	151.875	241.40625	55.9375	1646.5625	1724.6875	1706.0938	1764.375	1822.5	238.75	1768.125	294.6875	147.34375	1945.3125
	Amplitude	0.0004202	0.0003655	0.0003502	0.0002996	0.0002464	0.0013064	0.0012116	0.0012286	0.0010152	0.000812	0.0027179	0.0024888	0.0022371	0.002198	0.0018812
Eccentric Sheave	Frequency(Hz)	9.21875	60	38.4375	91.875	76.875	38.21875	134.6875	38.4375	82.03125	57.65625	29.21875	87.65625	58.4375	89.84375	233.90625
	Amplitude	0.0029327	0.0015364	0.0010404	0.0008435	0.0008419	0.0119384	0.004938	0.0047791	0.0033908	0.0032282	0.0296888	0.0195642	0.0161502	0.0114478	0.0100598
Fault Bearing	Frequency(Hz)	726.84375	718.75	707.96875	713.28125	739.0625	705	720.3125	721.5625	726.25	717.65625	0.3125	11.09375	11.5625	725.3125	10.625
	Amplitude	0.0023632	0.0020115	0.0019676	0.0019141	0.0018856	0.0038211	0.0037949	0.0037411	0.0034878	0.0033757	0.1867574	0.0113586	0.0112937	0.0107483	0.0106874
Loose Belt	Frequency(Hz)	135.625	9.6875	136.5625	106.25	145.9375	235.625	286.25	19.6875	296.875	274.84375	88.75	90.625	59.21875	592.03125	128.28125
	Amplitude	0.0008009	0.0007273	0.0007144	0.0007106	0.0006259	0.0033374	0.0026938	0.0024807	0.0022469	0.0020217	0.0213249	0.0120661	0.0073317	0.0062619	0.004719
Worn Gearbox Tooth	Frequency(Hz)	195.78125	352.34375	215.3125	97.8125	189.21875	329.21875	355.625	105.3125	697.96875	737.5	99.21875	337.5	19.84375	635.3125	714.6875
	Amplitude	0.0100061	0.0078842	0.0056433	0.0054435	0.0054267	0.0192294	0.0155315	0.0139652	0.0116473	0.0105037	0.0276796	0.0201357	0.0198421	0.0164714	0.0148987

Figure 19: Vibration characteristics in frequency domain channel 2

Channel 4		Harmonic Vibrations														
		Shaft Rotational Speed 10Hz					Shaft Rotational Speed 20Hz					Shaft Rotational Speed 30Hz				
Fault Type	Frequency vs Amplitude	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X	1X	2X	3X	4X	5X
Benchmark	Frequency(Hz)	50	149.84375	99.84375	199.84375	349.6875	50	149.84375	99.84375	449.375	199.6875	50	149.84375	99.84375	199.84375	449.53125
	Amplitude	0.00122516	0.00074823	0.00057886	0.00059131	0.00024708	0.00088605	0.00061278	0.00038317	0.00021441	0.000198	0.00125675	0.00078463	0.00067886	0.0003786	0.0002492
Coupling End Bent Shaft	Frequency(Hz)	50	149.84375	99.84375	449.53125	349.6875	50	149.84375	100	349.84375	449.6875	50	149.84375	99.84375	449.375	349.53125
	Amplitude	0.00114585	0.00076939	0.00025821	0.00019289	0.00012659	0.00134971	0.0004959	0.00025488	0.00017577	0.00012239	0.0009638	0.00038479	0.00031523	0.00019091	0.00016529
Cracked Shaft	Frequency(Hz)	49.84375	149.6875	99.84375	349.375	449.0625	50	149.84375	99.84375	349.53125	449.53125	50	149.84375	99.84375	449.53125	349.6875
	Amplitude	0.00109575	0.00073121	0.00031958	0.00016891	0.00013449	0.0010994	0.00073479	0.00029882	0.00016815	0.00013743	0.00119559	0.00078019	0.00025414	0.00019345	0.00017448
Eccentric Sheave	Frequency(Hz)	50	100.15625	150.15625	200.15625	20	50	150.15625	100	200.15625	350.15625	50	150.15625	100	326.25	200.15625
	Amplitude	0.00122716	0.00096038	0.00074626	0.00038931	0.00025365	0.00137027	0.00049243	0.00031841	0.00020359	0.00019495	0.00136896	0.00051103	0.00045087	0.00031953	0.00022944
Fault Bearing	Frequency(Hz)	49.84375	149.6875	99.84375	449.0625	349.21875	49.84375	149.6875	99.84375	349.375	449.21875	50	149.84375	449.6875	100	199.84375
	Amplitude	0.00114946	0.00078225	0.00027007	0.00020401	0.00012817	0.00107	0.00072995	0.00027487	0.00018313	0.00017531	0.00132751	0.00065769	0.00021162	0.00020524	0.00016528
Loose Belt	Frequency(Hz)	50	150	100	449.84375	200	50	150.15625	100.15625	200.15625	450.46875	50	100	150	200.15625	350.15625
	Amplitude	0.00142674	0.00068929	0.00033564	0.00022342	0.00021325	0.00114842	0.00075098	0.00039817	0.00024738	0.00020875	0.00142477	0.00057382	0.00033148	0.00025807	0.00023244
Warn Gearbox Tooth	Frequency(Hz)	50.15625	150.3125	100.15625	350.78125	200.46875	50.15625	150.3125	100.15625	350.625	200.3125	50.15625	150.3125	100.15625	350.625	200.3125
	Amplitude	0.00122345	0.00077952	0.0002733	0.00016582	0.00010582	0.00108776	0.00072678	0.00034957	0.00017826	0.00015494	0.00109198	0.0007349	0.00034791	0.00017601	0.00014681

Figure 20: Vibration characteristics in frequency domain channel 4

Figure 17 presents the harmonic vibrations of channel 1 at different shaft rotational speeds (10hz, 20hz, 30hz) and highlights the distinct frequency and amplitude patterns for various faults. At 10hz, the benchmark fault shows a primary vibrational characteristic at 50hz, with amplitude readings decreasing for the loose belt as harmonics increase. The eccentric sheave, coupling end bent shaft, and cracked shaft all display strong amplitudes at the 1x harmonic, with a gradual decrease by the 5x harmonic. At 20hz and 30hz, similar patterns emerge, with each fault type starting strong at the 1x harmonic and tapering off by the 5x harmonic. The data underscores the unique vibrational signatures of different fault types across varying rotational speeds.

Figure 18 showcases harmonic vibrations of channel 2 at different shaft rotational speeds. At 10hz, the benchmark fault type displays a frequency of 140.625hz and an amplitude of 0.001519765, while the worn gearbox tooth fault has a frequency of 195.78125hz and an amplitude of 0.010006125. At 20hz, various faults, including the coupling end bent shaft and gearbox, exhibit distinct frequencies, with notable shifts at the 5x harmonic. The cracked shaft, eccentric sheave, fault bearing, and loose belt faults also have unique frequencies and amplitudes. At 30hz, the frequencies for the coupling end bent shaft and worn gearbox tooth fault further change, highlighting the significance of these metrics in machinery diagnostics.

For channel 3, at different shaft rotations in Figure 19, distinct vibrational patterns emerge for various fault types. At 10hz rotation, the benchmark has a 60hz frequency for its 1x harmonic with an amplitude of 0.000949. However, the coupling end bent shaft fault shows a change in amplitude at this rotation, but the frequency remains the same. At 20hz, the vibrational patterns vary significantly among faults. For instance, the coupling end bent shaft fault and worn gearbox tooth fault have frequencies of 19.84375hz and 658.4375hz at 1x harmonic, adjusting at 5x harmonic to 39.6875hz and 1995.15625hz. The cracked shaft fault ranges from 843.75hz to 882.1875hz from 1x to 5x harmonic. Other faults like eccentric sheave and loose belt also display unique frequencies. At 30hz rotation, the vibrational patterns further diversify, with each fault type presenting specific frequencies, underscoring the need for regular machinery monitoring.

For channel 4, Figure 20 illustrates that for various components and at different frequencies (10hz, 20hz, and 30hz), the amplitude consistently shows a declining trend from the 1x to the 5x harmonic. For instance, at 10hz, the loose belt starts with an amplitude of 0.001426742 and reduces to 0.000213255 by the 5x harmonic. This pattern of decreasing amplitude across harmonics is consistent for other components like the warn gearbox, fault bearing, coupling end bent shaft, cracked shaft, and eccentric sheave at all observed frequencies.

4.7 Vibration Increment

In Table 3, a comprehensive vibrational analysis across distinct fault types and channels, certain patterns distinctly emerged. Between the frequencies of 10Hz to 20Hz, the horizontal channels, CH1 and CH2, exhibited varied sensitivities to different fault types.

Table 3: Vibration increments for all channels between two shaft frequency speeds (10Hz and 20Hz)

Fault type	Vibration increment rate (10Hz and 20Hz)			
	CH1	CH2	CH3	CH4
Coupling end bent shaft	19%	103%	54%	-49%
Cracked shaft	5%	109%	169%	-2%
Eccentric sheave	1%	229%	184%	-3%
Defected bearing	4%	84%	24%	-89%
Loose belt	42%	127%	364%	3%
Worn gearbox tooth	2%	101%	159%	62%

For instance, the "Cracked Shaft" fault induced an increase of 228.64% in CH2, highlighting its sensitivity. On the other hand, the vertical channels, CH3 and CH4, demonstrated their unique patterns. Notably, the eccentric sheave fault caused CH4 to decrease by 88.96%, while CH3 increased by 24.12%.

As we shifted our focus to the 20Hz to 30Hz frequency range in Table 4, the dynamics altered. The benchmark data in CH4, a vertical channel, experienced a pronounced increment of 233.23%. Meanwhile, the fault bearing fault redirected its influence from CH3, where it had previously caused an increment of 363.86% in the 10Hz to 20Hz range, to the horizontal channels, CH1 and CH2, registering increases of 47.00% and 114.44% respectively.

The analysis of this data has yielded valuable insights into the relationship between various fault types and their optimal orientations and frequency ranges. It has become apparent that the frequency range of 10hz to 20hz holds the highest prevalence among the fault types.

Table 4: Vibration increments for all channels between two shaft frequency speeds (20Hz and 30Hz)

Fault type	Vibration increment rate (20Hz and 30Hz)			
	CH1	CH2	CH3	CH4
Coupling end bent shaft	9%	42%	29%	233%
Cracked shaft	-1%	91%	122%	18%
Eccentric sheave	2%	90%	97%	-3%
Defected bearing	12%	101%	94%	26%
Loose belt	47%	114%	1%	10%
Worn gearbox tooth	17%	229%	87%	3%

Table 5: Optimal Orientation and Frequency Range for Different Fault Types

Fault type	Optimal orientation and frequency range
Benchmark	Vertical (20 HZ – 30 HZ)
Coupling end bent shaft	Vertical (10 HZ-20 HZ)
Cracked shaft	Horizontal (10 HZ- 20 HZ)
Eccentric sheave	Horizontal (20 HZ- 30 HZ)
Bearing fault	Vertical (10 HZ- 20 HZ)
Loose belt	Horizontal (20 HZ- 30 HZ)
Warn gearbox tooth	Vertical (10 HZ- 20HZ)

Table 5 offers a clear overview of which sensor (horizontal or vertical) and frequency range are most effective for detecting each fault type, aiding in optimizing sensor placement for effective fault detection. It

specifies the ideal orientation and frequency range for diagnosing various mechanical faults. Vertical orientation is optimal for detecting issues such as a bent shaft at the coupling end, faulty bearing, and worn gearbox tooth, predominantly within the 10Hz to 20Hz frequency range. On the other hand, horizontal orientation is best for identifying a cracked shaft, eccentric sheave, and loose belt, generally within the 20Hz to 30Hz frequency range for the latter two and 10Hz to 20Hz for the cracked shaft. The frequency ranges for faults tend to cluster around either 10Hz to 20Hz or 20Hz to 30Hz, offering a focused approach for diagnostic efforts.

5 Discussion

This section highlights the main contributions and outcomes of the paper.

5.1 Vibration Signatures and Patterns

The detection of discrete vibration signatures linked with various mechanical flaws was one of the study's most noteworthy findings. This supports previous research that emphasises the significance of vibration analysis in machinery evaluation [5][10]. The data demonstrated that distinct defects manifest differently in both the time and frequency domains, resulting in a robust paradigm for fault identification. This is especially significant since it enables the construction of more precise predictive models for machinery maintenance.

The findings of this study are consistent with the literature discussed which emphasise the relevance of Condition Monitoring (CM) in improving productivity, reliability, and safety during operation [11]. The findings also support the proposal for an extensive machinery condition monitoring method in the literature [12]. This study makes a significant contribution by introducing the concept of multimodal approaches, which is supported by the literature [61]. As the literature emphasise, these methodologies provide a more thorough picture of machinery health, justifying the literature's emphasis on vibration monitoring as a non-invasive way of assessing machinery health [10].

5.2 Fault Severity and Rotational Speed

Another significant discovery was a clear relationship between the severity of the fault and the magnitude of vibrations, particularly at higher rotational speeds. This is consistent with prior research that has highlighted the importance of rotational speed in the emergence of mechanical flaws [64]. The consequence is that monitoring rotational speed can act as an early warning system for potential mechanical failures, allowing for timely intervention.

5.3 Time Domain

This research aligned with the literature, both emphasizing the critical role of vibration analysis in predicting machine faults [44]. The graphical representations of vibration signatures and extracting vibration parameters which are illustrated in the analysis developed on the conceptual groundwork from the literature [44], offering nuanced insights valuable for future endeavours in diagnostic and predictive maintenance. It would be beneficial for future studies to incorporate these specific metrics into the existing methodological framework. Such an integration would facilitate a more comprehensive approach to fault detection, marrying both graphical and numerical analyses for heightened accuracy.

The findings of the analysis are in alignment with the literature in the way that both underscore the efficacy of vibration analysis for diagnosing issues in rotating machinery [19][20]. Moreover, the research aligns well with the existing literature, which collectively emphasizes the importance of a multi-domain approach for effective machinery fault diagnosis. With the nature if this study utilizing time, frequency, and time-frequency domains in vibration analysis, it acknowledges the limitations of visual inspection, a concern highlighted in the literature [5][34]

This research also employs features-based inspection using machine learning, a method endorsed by different researchers [35]. Furthermore, it incorporates statistical techniques in the time domain, resonating with the insights shown in the literature [25]. Overall, the methodologies and observations in this study are in concordance with those in the existing body of literature, advocating for a comprehensive, multi-faceted approach for accurate fault diagnosis in machinery.

5.4 Crest Factor and Vibration Velocity

The Crest Factor (xCF) and Vibration Velocity (Vrms) were identified as reliable metrics for assessing machinery condition. This is consistent with the literature findings [20][43]. With regards to the crest factor, the findings of this research related to the Worn Gearbox Tooth fault align with existing literature on the subject as mentioned. Specifically, it was observed that the Crest Factor (xCF) values showed a clear trend across all monitored channels.

These values ranged from 4.3626 at Channel 4 (CH4) to 5.2752 at Channel 2 (CH2), consistently surpassing the critical threshold of 4.5. Such observations are not merely indicative of the existence of a fault; they underscore the severity of the issue, affecting multiple components within the gearbox. These findings are consistent with the literature, which posits that xCF values above 4.5 are generally indicative of damage in the mechanical system [20]. while gears in good condition typically show xCF values between 3.5 and 4.0, while those with impulsive defects can reach as high as 10.0. The alignment of our data with this existing research underscores the severity of the issue we have identified and suggests that it is systemic and potentially chronic within the gearbox.

Furthermore, the research investigation revealed that the Crest Factor values for the Worn Gear rose as rotational speed rose. These findings are consistent with earlier research. An experimental investigation on the vibration characteristics of regular gears, which presented that the Crest Factor changes with rotational speed [42]. Their findings, in particular, show that when the speed of the gear shaft increases, the Crest Factor values grow noticeably.

This research highlights significantly high Crest Factor (xCF) values for the Defected Bearing, with the highest x_{CF} reading of 7.0276 at CH2, as detailed in Table 12. Supported by literature that suggests that high crest factors can indicate emerging flaws in machinery bearings [36][39]. However, contrary views exist as at is argued that the crest factor is not reliable for early fault detection in bearings [37].

Concerning vibration velocity, the relevance of literature findings on vibration velocity, is closely aligned with this research [44]. the RMS value (Vrms) for a worn gearbox at 30Hz revealed that all problematic elements looked to be within operational limits. However, empirical evidence suggests that this worn gearbox is faulty and will fail soon. This disparity demonstrates the limits of using RMS data simply for condition monitoring.

In the context of assessing the severity of machinery health conditions, this research scrutinizes the high Vrms value of 0.3048 mm/s, which is attributed to the "Defective bearing." Upon comparison with international standards like ISO 10816 and ISO 13373, it was observed that the defected bearing, along with other fault types such as loose belts and cracked shaft parts, are still considered to be within operational limits. This methodology aligns with prior work in the field. Such an evaluation is essential for identifying potential equipment malfunctions or unsafe operating conditions.

5.5 Frequency Domain

When observing the analysis for the frequency domain. These unique signatures, observed at different harmonics and shaft rotational speeds, serve as valuable diagnostic criteria for machinery maintenance. The findings have practical implications for real-time monitoring and predictive maintenance, enabling timely interventions that can reduce operational costs and downtime. The variations in amplitude across different harmonics emerge as a crucial diagnostic indicator. Specifically, a Loose Belt consistently demonstrates a reduced amplitude across harmonics, while an Eccentric Sheave maintains a more stable profile. These unique vibration signatures are instrumental in identifying machinery faults. Such diagnostic criteria, when thoroughly analysed, offer critical insights into the machine's health, enabling more precise maintenance

decisions, as supported by literature [16][45].

The study's findings on vibration analysis are well-supported by existing literature that emphasized the effectiveness of vibration analysis for machine condition monitoring and adds to the literature by providing experiments, which demonstrates the effectiveness of vibration-based maintenance in a practical setting [27][65]. However, the study also aligns with literature in acknowledging that the effectiveness and accuracy of vibration analysis are highly dependent on skilled analysis and quality equipment [12][15][26].

5.6 Vibration Increment

In the context of Vibration Increment, which tracks the rate of change in vibration patterns, provides real-time insights [52]. The research further elaborates on the relationship between various fault types and their impact on different sensor channels. The research identifies that the frequency range from 10Hz to 20Hz is most prevalent for different fault types, thereby aiding in the optimization of sensor placement and frequency range selection for effective fault detection.

Regarding the sensor orientations and frequency ranges. Vertical sensors are best for faults like bent shafts and worn gearbox teeth in the 10-20Hz range, while horizontal sensors excel at identifying cracked shafts and loose belts, generally in the 20-30Hz range. These findings align with existing research strengthening the study's empirical basis [15][53].

5.7 Research Implications

The findings have far-reaching implications for the industry. The identification of distinct vibration signatures for different faults can lead to more targeted and efficient maintenance strategies. This is crucial for industries where machinery downtime can result in significant financial losses. Moreover, the study provides a scientific basis for the development of predictive maintenance algorithms, which could revolutionize industrial operations by minimizing unplanned downtimes.

The inclusion of Vibration increment in predictive maintenance algorithms could be a game-changer for industries reliant on heavy machinery. By providing a dynamic measure of machine health, it also allows for more nuanced and timely maintenance decisions, thereby reducing the risk of unplanned downtimes and associated costs [13].

From an academic standpoint, this study fills a crucial gap in the literature by providing a comprehensive framework for fault diagnosis through vibration analysis. It also opens avenues for further research, particularly in the area of machine learning algorithms for automated fault detection and the integration of multiple data sources for more robust predictive models.

This concept opens new avenues for academic research. Future studies could focus on developing machine learning algorithms that incorporate vibration increment as a key variable, thereby improving the predictive accuracy of these algorithms. Moreover, the research could delve into the optimal Vibration increment thresholds for different types of machinery and operational conditions [16][65].

5.8 Key findings

The research offers a comprehensive analysis of the role of vibration in machinery fault diagnosis, particularly in industrial settings. Key findings include:

- Identification of distinct vibration signatures for different mechanical faults, facilitating the development of more accurate predictive models for machinery maintenance.
- A direct correlation between the severity of a fault and the magnitude of vibrations, especially at higher rotational speeds, suggests that rotational speed can serve as an early warning system.
- Introduction of reliable metrics such as Crest Factor (xCF) and Vibration Velocity (Vrms) for machinery condition assessment.
- Emphasis on time and frequency domain analysis for real-time monitoring and predictive maintenance.
- Exploration of the concept of Vibration Increment, including optimal frequency ranges and sensor

placements for effective fault detection.

The highlighted findings hold significant implications within the context of existing literature. They extend and corroborate prior studies by offering a more robust framework for fault diagnosis. The study addresses a critical need for standardized assessment methods and adds practical depth to existing research. These contributions bridge the gap between theory and practice, benefiting both academia and industry.

5.9 Limitations

- Conducted under controlled conditions, not fully replicating real-world complexities.
- Focus on quantitative methods, lacking qualitative insights.
- Experiments are limited to a speed of 30Hz due to MFS system constraints, affecting the identification of issues at higher speeds.

While some researchers were able to run their experiments at a considerably higher speed of approximately 59.67Hz, they stated that it's still difficult to identify issues [59]. The frequency spectrum for bearing failures in vibration as stated in the literature divides into four areas, with increasing bearing wear indicating variations [66]. These regions are shown in Table 6. Despite these limitations, the research offers promising avenues for future investigation. This research leaves the potential for future investigation without straying into Area 2.

Table 6: Frequency spectrum regions

Area	Description	Rotational frequency speed
1	Machine rotational speed and harmonics	0-84 HZ
2	Bearing flaw frequency	84-500 HZ
3	Bearing part natural frequency	500-2000 HZ
4	High frequency recognition	Beyond 2000 HZ

6 Conclusion

This section presents the main conclusions and recommendations for future work.

6.1 Conclusions

The article provided a comprehensive overview of vibration analysis, its methodologies, and its role in assessing machinery health. Each objective outlined at the onset of this study has been meticulously addressed, culminating in a series of pivotal findings.

The literature focuses on the complex realm of vibration analysis. It introduces advanced methodologies and tools essential for fault detection and diagnosis. This article also highlights the practical uses of specialized platforms like the Machinery Fault Simulator (MFS) by Spectra-Quest and Vibra-Quest (VQ). The findings and analysis shift the discussion to fault feature extraction, a key step in fault identification and classification. The research extends from the identification of unique vibration signatures to the introduction of trustworthy metrics, it emphasizes the importance of Crest factor (xCF) and Vibration Velocity (Vrms) and demonstrates its utility through empirical data, offering significant contributions to the domain of machinery maintenance, especially within industrial environments. Furthermore, expanding the scope of the research to encompass diverse types of machinery and operational conditions could yield additional valuable insights.

To translate these findings into actionable insights, recommendations for real-time condition monitoring can be established. These would capitalize on the research's conclusions about the most effective frequency ranges and sensor placements for fault detection, thereby optimizing machinery health monitoring.

6.2 Recommendations for future research

While the study successfully addressed the research questions it initially posed, it also opened wide range of avenues for further exploration. Specifically, there are multiple areas ripe for future research including:

- Testing at higher speeds to investigate the prospect of greater precision.
- One promising direction involves developing predictive maintenance algorithms that utilize the metrics introduced in this study.
- While the study used platforms like MFS and VQ for experimental analysis, applying the learned methodologies in real-world industrial settings would provide invaluable insights.

Acknowledgments

The authors wish to acknowledge the invaluable contributions of technical staff at the School of Computer Science & Engineering, Faculty of Business & Technology, University of Sunderland.

References

[1] Senapaty, G. and Rao, U.S., "Vibration based condition monitoring of rotating machinery," in *MATEC Web of Conferences*, vol. 144, p. 01021. EDP Sciences, 2018.

[2] Amarnath, M., Shrinidhi, R., Ramachandra, A. and Kandagal, S.B., "Prediction of defects in antifriction bearings using vibration signal analysis," *Journal of the Institution of Engineers (India), Part MC*, vol. 85, p. 88, 2004.

[3] Zaman, I.B. et al., "Study of passive vibration absorbers attached on beam structure," *Applied Mechanics and Materials*, vol. 660, pp. 511–515, 2014.

[4] Mobley, R.K., *An Introduction to Predictive Maintenance*. Elsevier, 2002.

[5] Ahmed, H. and Nandi, A.K., "Compressive sampling and feature ranking framework for bearing fault classification with vibration signals," *IEEE Access*, vol. 6, pp. 44731–44746, 2018.

[6] Basri, E.I. et al., "Preventive maintenance (PM) planning: a review," *Journal of Quality in Maintenance Engineering*, vol. 23, no. 2, pp. 114–143, 2017.

[7] Ahuja, I.P.S. and Khamba, J.S., "Total productive maintenance: literature review and directions," *International Journal of Quality & Reliability Management*, vol. 25, no. 7, pp. 709–756, 2008.

[8] Prajapati, A., Bechtel, J. and Ganesan, S., "Condition based maintenance: a survey," *Journal of Quality in Maintenance Engineering*, vol. 18, no. 4, pp. 384–400, 2012.

[9] Kumar, S. et al., "Condition based maintenance of bearings and gears for fault detection–A review,"

Materials Today: Proceedings, vol. 5, no. 2, pp. 6128–6137, 2018.

[10] Goyal, D. and Pabla, B.S., "Condition based maintenance of machine tools—A review," *CIRP Journal of Manufacturing Science and Technology*, vol. 10, pp. 24–35, 2015.

[11] Tiboni, M. et al., "A review on vibration-based condition monitoring of rotating machinery," *Applied Sciences*, vol. 12, no. 3, p. 972, 2022.

[12] Teixeira, H.N., Lopes, I. and Braga, A.C., "Condition-based maintenance implementation: a literature review," *Procedia Manufacturing*, vol. 51, pp. 228–235, 2020.

[13] Cınar, Z.M. et al., "Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0," *Sustainability*, vol. 12, no. 19, p. 8211, 2020.

[14] Selcuk, S., "Predictive maintenance, its implementation and latest trends," *Proc. Inst. Mech. Eng. Part B*, vol. 231, no. 9, pp. 1670–1679, 2017.

[15] Teng, W. et al., "Vibration analysis for fault detection of wind turbine drivetrains—A comprehensive investigation," *Sensors*, vol. 21, no. 5, p. 1686, 2021.

[16] Meyer, A., "Vibration fault diagnosis in wind turbines based on automated feature learning," *Energies*, vol. 15, no. 4, p. 1514, 2022.

[17] Ekeocha, R.J.O., "Vibration in systems," *J. Mech. Eng. Res.*, vol. 10, no. 1, pp. 1–6, 2018.

[18] Jauregui, J.C. and Lozano Guzman, A.A., *Mechanical Vibrations and Condition Monitoring*, 2020.

[19] Althubaiti, A., Elasha, F. and Teixeira, J.A., "Fault diagnosis and health management of bearings in rotating equipment based on vibration analysis–a review," *J. Vibroeng.*, vol. 24, no. 1, pp. 46–74, 2022.

[20] Altaf, M. et al., "A new statistical features based approach for bearing fault diagnosis using vibration signals," *Sensors*, vol. 22, no. 5, p. 2012, 2022.

[21] Vishwakarma, M. et al., "Vibration analysis & condition monitoring for rotating machines: a review," *Materials Today: Proceedings*, vol. 4, no. 2, pp. 2659–2664, 2017.

[22] Zhang, S. et al., "Deep learning algorithms for bearing fault diagnostics—A comprehensive review," *IEEE Access*, vol. 8, pp. 29857–29881, 2020.

[23] Nithyavathy, N. et al., "Vibration monitoring and analysis of ball bearing using GSD platform," *Materials Today: Proceedings*, vol. 43, pp. 2290–2295, 2021.

[24] Randall, R.B. and Antoni, J., "Rolling element bearing diagnostics—A tutorial," *Mech. Syst. Signal Process.*, vol. 25, no. 2, pp. 485–520, 2011.

[25] Nandi, A.K., Liu, C. and Wong, M.D., "Intelligent vibration signal processing for condition monitoring," *Proc. Int. Conf. Surveillance*, vol. 7, pp. 1–15, 2013.

[26] Al-Najjar, B. and Alsyouf, I., "Enhancing a company's profitability and competitiveness using integrated vibration-based maintenance: A case study," *Eur. J. Oper. Res.*, vol. 157, no. 3, pp. 643–

657, 2004.

[27] Malla, C. and Panigrahi, I., “Review of condition monitoring of rolling element bearing using vibration analysis and other techniques,” *J. Vib. Eng. Technol.*, vol. 7, pp. 407–414, 2019.

[28] Vavra, C., “Machine vibration analysis benefits for manufacturers,” 2021.

[29] Saucedo-Dorantes, J.J. et al., “Multiple-fault detection methodology based on vibration and current analysis applied to bearings in induction motors and gearboxes on the kinematic chain,” *Shock and Vibration*, 2016.

[30] Lacey, S.J., “Using vibration analysis to detect early failure of bearings,” *Insight-Northampton*, vol. 49, no. 8, pp. 444–446, 2007.

[31] Giagopoulos, D. et al., “Optimal modeling of an elevator chassis under crash scenario,” *Applied Mechanics*, vol. 3, no. 1, pp. 227–243, 2022.

[32] Matsushita, O. et al., “An overview of vibration problems in rotating machinery,” in *Vibrations of Rotating Machinery*, Springer, pp. 1–18, 2019.

[33] Kumar, J.P., Chauhan, P.S. and Pandit, P.P., “Time domain vibration analysis techniques for condition monitoring of rolling element bearing: A review,” *Materials Today: Proceedings*, vol. 62, pp. 6336–6340, 2022.

[34] Fickling, S.D. et al., “The EEG quality index for automated assessment of signal quality,” in *IEEE IEMCON*, pp. 219–229, 2019.

[35] Altaf, M. et al., *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms*. Wiley, 2020.

[36] Li, S. et al., “An ensemble deep convolutional neural network model with improved DS evidence fusion for bearing fault diagnosis,” *Sensors*, vol. 17, no. 8, p. 1729, 2017.

[37] Tandon, N., “A comparison of some vibration parameters for the condition monitoring of rolling element bearings,” *Measurement*, vol. 12, no. 3, pp. 285–289, 1994.

[38] Dube, A.V. et al., “Vibration based condition assessment of rolling element bearings with localized defects,” *Int. J. Sci. Technol. Res.*, vol. 2, no. 4, pp. 149–155, 2013.

[39] Kondhalkar, G.E. and Diwakar, G., “Crest factor measurement by experimental vibration analysis,” in *ICRRM 2019*, Springer, pp. 133–138, 2020.

[40] Zhang, X. et al., “Early fault detection method of rolling bearing based on MCNN and GRU network,” *Shock and Vibration*, 2021, pp. 1–13.

[41] Boudiaf, A. et al., “A comparative study of various methods of bearing faults diagnosis,” *J. Failure Analysis and Prevention*, vol. 16, no. 2, pp. 271–284, 2016.

[42] Alshammari, S.A. et al., “Fault diagnosis of helical gear through various vibration techniques,” *J. Mech. Des. Vib.*, vol. 7, no. 1, pp. 21–26, 2019.

[43] Wang, K. et al., “Mechanical fault prognosis through spectral analysis of vibration signals,” *Algorithms*, vol. 15, no. 3, p. 94, 2022.

[44] Igba, J. et al., “Analyzing RMS and peak values of vibration signals,” *Renewable Energy*, vol. 91, pp. 90–106, 2016.

[45] Souza, R.M. et al., “Deep learning for diagnosis and classification of faults,” *Computers & Industrial Engineering*, vol. 153, p. 107060, 2021.

[46] Brandt, A., *Noise and vibration analysis: signal analysis and experimental procedures*, Chichester, UK: John Wiley & Sons, 2023.

[47] Kamiel, B.P., “Demodulation of vibration signal based on envelope-kurtogram,” *JMPM*, vol. 4, no. 2, pp. 115–123, 2020.

[48] Stack, J.R. et al., “An amplitude modulation detector for fault diagnosis,” *IEEE Trans. Ind. Electron.*, vol. 51, no. 5, pp. 1097–1102, 2004.

[49] Galar, D. and Kumar, U., *eMaintenance: Essential Electronic Tools for Efficiency*. Academic Press, 2017.

[50] Ju, R. et al., “Comparison between the incremental harmonic balance method,” *J. Vibration and Acoustics*, vol. 143, no. 2, p. 024501, 2021.

[51] Yang, D.M., “The detection of motor bearing fault with maximal overlap discrete wavelet packet transform and teager energy adaptive spectral kurtosis,” * Sensors *, vol. 21, no. 20, p.6895, 2021.

[52] Song, Q. et al., “Accuracy of fault diagnosis for rolling element bearings,” *Sensors*, vol. 20, no. 22, p. 6465, 2020.

[53] Rahi, S., “Research design and methods: a systematic review,” *Int. J. Econ. & Mgmt. Sci.*, vol. 6, no. 2, pp. 1–5, 2017.

[54] Li, S. et al., “Deep convolutional neural network for bearing fault diagnosis,” *Sensors*, vol. 17, no. 8, p. 1729, 2017.

[55] Tchomeni Kouejou, B.X. and Alugongo, A.A., “Cracked cardan shaft analysis,” *Fluids*, vol. 8, no. 7, p. 211, 2023.

[56] El-Thalji, I. and Jantunen, E., “Summary of fault modelling and predictive health monitoring,” *Mech. Syst. Signal Process.*, vol. 60, pp. 252–272, 2015.

[57] Mustafa, A.M. and Nayfeh, J.F., “Experimental research on machinery fault simulator,” in *PHM-Besançon*, IEEE, pp. 72–78, 2020.

[58] Ashwear, N. and Eriksson, A., “Influence of temperature on the vibration properties of tensegrity structures,” in * International Journal of Mechanical Sciences *, vol. 99, pp.237-250, 2015.

[59] Pacheco-Chérrez, J. et al., “Bearing fault detection with vibration and acoustic signals,” *Engineering Failure Analysis*, vol. 139, p. 106515, 2022.

[60] Scheffer, C. and Girdhar, P., *Practical Machinery Vibration Analysis and Predictive Maintenance*, Elsevier, 2004.

[61] Dyer, D. and Stewart, R.M., “Detection of rolling element bearing damage,” 1978.

[62] Lee, C.Y. et al., “Data science for vibration heteroscedasticity,” *Energies*, vol. 12, no. 5, p. 801, 2019.

[63] Ibarra-Zarate, D. et al., “Bearing fault diagnosis using cepstrum pre-whitening,” *Int. J. Adv. Manuf. Technol.*, vol. 104, pp. 4155–4168, 2019.

[64] Usgame Sandoval, H.M. et al., “Acoustic emission-based early fault detection,” *Ingeniería e Investigación*, vol. 33, no. 3, pp. 5–10, 2013.

[65] Zhang, X. et al., “Early fault detection with MCNN and GRU,” *Shock and Vibration*, 2021.

[66] Wu, K. et al., “Torsional vibration responses of the engine crankshaft-gearbox system,” *Mech. Syst. Signal Process.*, vol. 173, p. 109052, 2022.

Chapter 41

Rethinking Hospital Asset Maintenance: Integrating Artificial Intelligence into CMMS and EAM Systems

Contribution #79
Cite as:
António Seiça, Mateus Mendes. Rethinking Hospital Asset Maintenance: Integrating Artificial Intelligence into CMMS and EAM Systems. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-79,
  title ={Rethinking Hospital Asset Maintenance:
     Integrating Artificial Intelligence into CMMS and
     EAM Systems},
  author={António Seiça and  Mateus Mendes},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Rethinking Hospital Asset Maintenance: Integrating Artificial Intelligence into CMMS and EAM Systems

António Seiça[1], Mateus Mendes[1,2]

[1] Polytechnic of Coimbra, ISEC, 3045-093 Coimbra, Portugal
[2] RCM2+ - Research Centre in Asset Management and System Engineering, Portugal

aseica@gmail.com, mmendes@isec.pt

Abstract

Hospitals increasingly depend on complex, high-value physical assets, making effective maintenance essential for both patient safety and operational efficiency. However, traditional maintenance systems often fail to meet the demands of growing technological complexity, workforce shortages, and the need for real-time decision-making. Computerized Maintenance Management Systems (CMMS) and Enterprise Asset Management (EAM) systems play a critical role when effectively implemented. Nevertheless, staff frequently struggle to use them as meaningful tools, often perceiving these systems as burdensome administrative tasks.

This paper draws on foundational concepts in maintenance management and relevant industry standards to introduce key definitions, identify the limitations of conventional approaches, and propose a layered, AI/ML-enhanced system architecture for the next generation of CMMS/EAM platforms. It explores areas where AI and machine learning can bring significant improvements, including asset lifecycle management, work order optimization, predictive maintenance, inventory control, cost analysis, and access to technical documentation.

Additionally, the paper presents a high-level architectural framework as a blueprint for the future implementation of AI-assisted maintenance systems, emphasizing the importance of interoperability to ensure a cohesive, efficient, and proactive hospital maintenance ecosystem. Future directions, such as cloud-edge orchestration and advanced predictive maintenance strategies, are also explored

Keywords: Computerized Maintenance Management System (CMMS); Enterprise Asset Management (EAM); Artificial Intelligence (AI); Machine Learning (ML); Hospital Asset Management; Predictive Maintenance

1 Introduction

Hospitals are complex ecosystems that rely heavily on the seamless operation of a wide range of physical assets—from sophisticated electromedical equipment to essential support infrastructure such as heating, ventilation, and air conditioning (HVAC) systems, medical gas networks, and power generators. In addition to these major systems, healthcare professionals interact daily with a multitude of smaller devices and tools distributed across various hospital units. The effective management and maintenance of this diverse asset base are critical to ensuring patient safety, operational efficiency, and compliance with regulatory standards.

However, hospital asset management is undergoing a rapid transformation. Modern healthcare facilities increasingly depend on advanced technologies that incorporate microelectronics, robotics, smart sensors, and Internet of Things (IoT) connectivity. This surge in equipment complexity, combined with rising patient demand and an aging population, places significant strain on traditional maintenance practices.

Conventional maintenance systems often rely on manual data entry, reactive workflows, and siloed information, leading to inefficiencies, increased equipment downtime, and escalating operational costs. Moreover, the rapid pace of technological advancement frequently outpaces the training and expertise available

to maintenance personnel, widening the skills gap.

This paper argues that Artificial Intelligence (AI) and Machine Learning (ML) technologies offer transformative solutions to these challenges. Techniques such as Large Language Models (LLMs) for natural language understanding, Long Short-Term Memory (LSTM) recurrent networks for time-series analysis and prediction, classification algorithms, and computer vision deep-learning models like You Only Look Once (YOLO) can enable a paradigm shift from reactive to proactive and predictive maintenance.

By integrating AI/ML capabilities into existing Computerized Maintenance Management Systems (CMMS) and Enterprise Asset Management (EAM) platforms, hospitals can significantly enhance asset performance, improve reliability, reduce downtime, and enable data-driven decision-making. This paper explores the foundational concepts of maintenance management, outlines current challenges, presents detailed AI/ML-driven strategies, and proposes a high-level framework for the future of hospital physical asset maintenance.

2 Background: Evolution of Maintenance Management Systems in Hospitals

Effective maintenance is broadly defined as "the combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function", as stated at EN13306 Maintenance Terminology [1]. To manage these complex activities, organizations have increasingly relied on specialized information systems.

Computerized Maintenance Management Systems are software platforms designed to support the planning, execution, and tracking of maintenance activities. Their primary functions include facilitating technician workflows, monitoring equipment performance, logging interventions, and assisting maintenance teams in data-driven decision-making. CMMS platforms can also provide evidence of regulatory compliance, which is particularly critical in healthcare settings [2].

Building upon the capabilities of CMMS, Enterprise Asset Management systems offer a more comprehensive approach. EAM systems manage the full lifecycle of an organization's physical assets—from acquisition and commissioning to operation, maintenance, and eventual decommissioning or replacement. The "enterprise" scope reflects the ability to manage assets across multiple departments, facilities, and locations. EAM aims to optimize asset performance, reduce both capital and operational costs, and extend asset life, thereby improving return on investment (ROI) [3]. While CMMS and EAM are sometimes used interchangeably, EAM represents a broader and more strategic asset management approach. This is aligned with the ISO 55000 series (e.g., ISO 55000:2014, ISO 55001:2014, ISO 55002:2014), which provides a formal framework for physical asset management at an organizational level [3].

Within hospital environments, maintenance and infrastructure responsibilities are typically assigned to departments known as the Service of Installations and Equipment (SIE) or the Service of Electromedical Equipment (SEE). These units oversee a wide and heterogeneous portfolio of assets, including:

- Electromedical equipment: Devices such as infusion pumps, vital signs monitors, ophthalmic microscopes, portable ventilators, electrosurgical units, and defibrillators.
- Support equipment: Infrastructure-related systems including HVAC units, medical gas networks, water heating and cooling systems, elevators, and power backup systems.
- Building infrastructure: The structural and functional elements of the hospital itself.

To ensure regulatory compliance and operational safety, SIEs must maintain organized and up-to-date records for a variety of elements: technical datasheets, scheduled maintenance plans, completed preventive maintenance records, fault and repair histories, budgets, maintenance requests, and calibration certificates for testing devices, also known as Resources for Monitoring and Measurement (RMM).

A crucial component of electromedical maintenance is the regular calibration and safety verification of devices, especially those involved in measuring biological signals. These procedures are typically carried out by accredited third-party laboratories certified under standards such as NP EN ISO/IEC 17025, which ensures

traceable calibration documentation. In parallel, electrical safety is a mandatory requirement governed by standards including IEC 60601-1 and IEC 62353. The former outlines safety and performance criteria for medical electrical equipment, including classifications (e.g., Class I, Class II, Type B, BF, CF) and protective measures. The latter, IEC 62353, focuses on periodic testing and post-repair inspections and defines four primary test types: Earth Resistance, Leakage Currents, Insulation Resistance, and Functional Resistance.

Hospitals typically adopt one of several maintenance models: in-house technical teams, manufacturer-specific service contracts, third-party multi-brand service providers, or a hybrid strategy combining multiple approaches. Each model offers distinct advantages and trade-offs in terms of control, specialization, responsiveness, and cost.

Prior work such as "CMMS – An integrated view from maintenance management to online condition monitoring" [3] underscores the need for CMMS and EAM systems to evolve in response to emerging operational requirements. These include multi-platform access, multilingual interfaces, and the ability to integrate data streams from diverse sensors—wired or wireless, local or remote. The study also anticipates the integration of advanced technologies such as Augmented Reality (AR), Artificial Vision, and holography to enhance technician performance and maintenance efficiency.

In another work, "Augmented Reality Maintenance Assistant Using YOLOv5" [4], the authors propose a deep learning-based assistant tool capable of recognizing components of complex equipment such as car engines, using computer vision. This model leverages the YOLOv5 architecture for object detection and represents a practical example of how AI technologies can assist real-time maintenance workflows.

3 Current Challenges in Hospital Asset Maintenance

Despite the advancements in maintenance management systems, hospitals continue to face a range of complex challenges in managing their physical assets effectively. These challenges stem from increasing technological complexity, workforce limitations, and system interoperability issues, all of which hinder optimal maintenance practices:

a) Increasing Equipment Complexity and Technological Sophistication

Modern healthcare institutions are increasingly more equipped with advanced technologies, including microelectronics, robotics, high-precision sensors, imaging systems, lidar and radar units, IoT-enabled devices, and AI-powered subsystems. These technologies demand complex tasks such as configuration, programming, tuning, calibration, and metrology, which often exceed the capabilities of conventionally trained maintenance personnel. This raises a critical issue: are new graduates and technicians adequately prepared for these tasks, or are they primarily trained to rely on OEMs (Original Equipment Manufacturers) for support. The growing dependency on external services may delay interventions and increase operational costs.

b) Manpower Shortages and Skills Gaps

The rapid evolution of intelligent medical equipment often outpaces the learning and adaptation capacity of technical staff. This leads to a widening knowledge gap, where in-house maintenance personnel struggle to keep up with newly introduced systems. Additionally, many organizations continue to rely on outdated resources such as paper-based manuals and undocumented experiential knowledge, which inhibits standardized diagnostic procedures and the identification of degradation patterns. The lack of scalable training approaches and knowledge repositories also contributes to inconsistency in maintenance practices across departments.

c) Aging Infrastructure and Rising Demand

Demographic trends—particularly the aging population as confirmed at population-pyramid.net [5]—are expected to drive up demand for healthcare services. This translates into heavier usage of complex medical devices and infrastructure, accelerating wear and increasing maintenance burdens.

Hospitals must therefore not only plan for increased equipment acquisition but also prepare for all associated lifecycle processes: procurement, installation, training, commissioning, operation, maintenance, and eventual decommissioning or replacement.

d) Data Silos and Interoperability Constraints

Modern CMMS and EAM systems often incorporate data from condition monitoring systems (e.g., temperature, vibration, and fluid sensors). However, these subsystems frequently operate on proprietary protocols or closed architectures, complicating seamless integration across platforms.
This lack of interoperability can lead to fragmented data flows, duplicated manual entries, and incomplete asset histories. Moreover, integration challenges are exacerbated in multi-provider environments, where different stakeholders (e.g., OEMs, third-party contractors, and hospital departments) operate their own CMMS platforms without standardized interfaces for data exchange.

e) Limitations of Traditional Maintenance Practices

Conventional maintenance strategies often emphasize corrective (reactive) or scheduled (preventive) maintenance, both of which have notable shortcomings.
For example, the common workflow of printing work orders (WOs) for field technicians and then documenting the intervention afterward in the office leads to significant time lags. This gap increases the risk of incomplete or inaccurate records and prevents real-time access to critical tools such as fault diagnostic systems.
Furthermore, the lack of adaptive, data-driven scheduling mechanisms means maintenance tasks may either be performed too late (leading to failures) or too early (leading to unnecessary interventions).

These multidimensional challenges underscore the urgent need for a more intelligent, integrated, and proactive approach to hospital asset maintenance—one that leverages AI and ML technologies to fill expertise gaps, automate routine decision-making, and enable predictive capabilities.

4 Proposed Solutions: AI and Machine Learning Integration into CMMS/EAMS

Addressing the previously outlined challenges requires a paradigm shift—one that leverages Artificial Intelligence (AI) and Machine Learning (ML) to fundamentally enhance maintenance management. As the Software as a Service (SaaS) model continues to evolve, we are witnessing a transition from specialized, task-specific platforms to AI-driven systems where intelligent agents assume much of the decision-making and operational control. As referred by Nandella [6], traditional SaaS applications risk becoming relegated to the role of passive data repositories. This evolution presents a strategic opportunity to rethink and upgrade hospital asset management practices.

A comprehensive AI-enhanced maintenance system should adopt a multi-layered architecture, enabling improvements across the entire asset lifecycle—from data acquisition and analysis to predictive decision-making and task execution.

AI and ML can profoundly improve the functionality of traditional CMMS core modules:

a) Asset Management:

AI can significantly enhance asset management by enabling natural language queries to retrieve complex information. For example, users can ask, “What is the failure history of the blood pressure monitor used in the cardiology ward?” AI systems can also summarize lifecycle cost reports from historical data and generate repair guides based on recent failure reports, detailing necessary tools, procedures, and required technician skillsets. Modern Large Language Models systems have this capacity, once full integration with the CMMS or EAM systems provide the necessary data.

b) Work Order Management (WO):

AI can streamline work order processes through natural language interfaces—for instance, "Create a WO to replace HVAC filters in Surgery A." It can autonomously assign appropriate tools, spare parts, and technicians based on their qualifications, experience, and availability. Furthermore, AI can suggest repair actions based on reported symptoms and recommend checks on potentially affected nearby equipment. This can be achieved again through integration of LLM systems.

Dynamic Work Order Prioritization Agents can rank tasks based on predicted service impact, equipment criticality, and technician availability. Additionally, LLM-powered systems can automatically summarize technician voice notes into structured WO close-out reports, improving documentation quality and reducing administrative burden.

c) Corrective, Preventive, and Predictive Maintenance:

AI can optimize maintenance schedules by analyzing equipment histories and recommending ideal maintenance intervals. It can translate complex sensor readings and anomalies into actionable insights expressed in clear, accessible language.

A Maintenance Recommender Agent can dynamically adjust preventive maintenance plans on a monthly basis by incorporating variables such as equipment usage patterns, asset behavior, and weather data. Predictive Maintenance, enabled by AI models, plays a critical role in reducing downtime and avoiding critical failures by estimating the Remaining Useful Life (RUL) of systems using real-time condition monitoring data.

d) Root Cause Analysis (RCA):

AI systems can assist in RCA by proposing likely failure causes based on logs, alerts, and historical reports. They can help generate visual aids such as fishbone (Ishikawa) diagrams and perform 5-Whys analysis automatically. LLM-driven causal agents can prompt users, process structured and unstructured data, and generate RCA documentation autonomously, significantly reducing diagnostic time and increasing consistency in failure interpretation.

e) Facilities, Tools, Spare Parts & Inventory Management:

AI can recommend spare part reorders based on historical usage patterns and forecast needs using predictive analytics. It can identify equivalent substitutes for discontinued parts and track tool utilization per intervention.

Furthermore, AI can recommend optimal "intervention pathways" based on facility layouts and technician proximity. An Autonomous Spare Part Forecasting Agent can anticipate stockouts by analyzing asset failure rates, supplier lead times, and aging profiles, even employing supply-demand models inspired by financial arbitrage to optimize procurement timing and prices.

f) Cost Control & Key Performance Indicators (KPIs):

AI can enhance financial oversight by offering natural language interfaces for dashboards and auto-generating executive summaries. It can also forecast maintenance costs based on equipment performance data and reliability indices.

Using the NP EN 15341:2009 framework [7]—which defines KPIs across economic, technical, and organizational dimensions—AI agents can track and evaluate performance trends. For example, an AI-driven dashboard might answer: "Why did maintenance costs spike in March?" by correlating WO data, parts usage, and equipment reliability. An AI Auditor Agent can continuously monitor KPI targets, detect anomalies, and recommend corrective actions.

g) Procurement Management:

AI can facilitate procurement processes by collecting staff input to draft tender documents, then evaluating and summarizing proposals from multiple vendors. It can assist in supplier selection by comparing offers against past performance metrics, pricing trends, and delivery reliability.

h) 3D, Virtual & Augmented Reality, Holography, and Digital Twins:

Advanced visualization technologies offer considerable value in enhancing maintenance performance. Augmented Reality (AR), in particular, supports real-time, immersive guidance by overlaying digital instructions onto physical components. AR can also store digital equipment metadata, making operational status and history accessible at a glance. Image recognition [4], Digital Twins [3] and holographic interfaces provide deeper, simulation-based insights into asset behavior and system interactions.

i) Interfaces & Communication:

AI can support technicians with intelligent voice assistants or chatbot interfaces, facilitating hands-free operation and instant troubleshooting. Automated helpdesk functions can guide users through problem-solving workflows.

Moreover, Just-In-Time Learning Modules powered by LLMs can deliver short, tailored training sessions immediately before unfamiliar tasks, drawing from manuals, past interventions, and expert knowledge.

5 Case Examples and Scenarios

The practical application of AI-powered maintenance systems in hospitals can be illustrated through the following real-world-inspired scenarios:

a) Proactive Ventilator Maintenance

A nurse reports an unusual noise coming from an ICU ventilator during a routine check. Instead of waiting for a critical failure, an AI assistant analyzes sensor data and logs, detects anomalies in airflow and pressure regulation, and preemptively generates a work order (WO). It suggests the necessary tools, retrieves the digital manual section relevant to the issue, and notifies the on-duty biomedical technician. This AI-driven, preemptive action prevents equipment downtime and maintains continuous availability of life-critical systems.

b) Dynamic Operating Room (OR) Environmental Control

During surgery, the AI system detects slight deviations in OR temperature and humidity, which could affect sterility conditions. Rather than relying on scheduled checks or post-procedure reports, the system autonomously issues a WO to adjust the HVAC settings. It identifies which technician is available and nearby, sends real-time instructions to their mobile device, and logs the event. This immediate response ensures compliance with surgical environment standards, reducing infection risk and improving patient safety.

c) Smart Spare Part Forecasting in Dialysis Units

A hospital's dialysis unit frequently experiences delays due to unavailable filter cartridges. Using historical data, patient scheduling, and machine usage rates, the AI system forecasts future demand and detects that stock levels are insufficient for the upcoming two weeks. It automatically generates a purchase requisition, suggests a reliable vendor based on previous delivery times and cost performance, and flags the request to procurement. This proactive measure avoids service interruptions and optimizes inventory turnover.

d) Technician Onboarding with Just-in-Time Learning

A new maintenance technician is assigned to service an unfamiliar robotic surgery system. As they approach the equipment, the AI interface detects their presence via NFC badge and delivers a short, targeted instructional video tailored to that model and its known failure modes. The AI also preloads the latest repair logs, lists tools needed for the intervention, and offers AR-based visual guidance through smart glasses. This scenario highlights how AI can reduce onboarding time and elevate the skill ceiling of junior staff.

e) Root Cause Analysis of Infusion Pump Failures

Several infusion pumps report intermittent failures during medication delivery. The AI system clusters fault logs and user error reports across devices and correlates them with power fluctuation data and humidity levels in storage areas. It determines that faulty power connectors in pumps stored in higher-humidity zones are more prone to early degradation. The AI assistant automatically builds a fishbone diagram illustrating the contributing factors and suggests mitigation: relocating equipment and replacing a specific batch of connectors.

6 High-Level System Architecture for AI-Enhanced CMMS/EAM

To effectively integrate AI and Machine Learning into Computerized Maintenance Management Systems (CMMS) or Enterprise Asset Management (EAM), a layered architectural model is proposed. This architecture ensures scalability, modularity, and interoperability across healthcare maintenance environments. Each layer plays a critical role in transforming raw equipment data into actionable maintenance intelligence.

a) Data Acquisition Layer

Components: IoT sensors, OEM interfaces, legacy equipment connectors, SCADA systems, APIs, MQTT brokers.
Function: This layer is responsible for acquiring real-time equipment data, including telemetry from ventilators, infusion pumps, HVAC systems, and diagnostic machines. It aggregates sensor data—such as temperature, vibration, pressure, and voltage—using industrial protocols and OEM APIs. Legacy systems may require custom bridges or adapters to participate in the ecosystem.
Example: A ventilator equipped with pressure sensors and an MQTT-enabled microcontroller sends data on airflow patterns and anomalies to the CMMS every 10 seconds.

b) Core CMMS Modules Layer

Components: Asset Management, Work Order (WO) Management, Preventive/Predictive Maintenance modules among others
Function: These are the backbone systems where core maintenance logic resides. With AI augmentation, these modules are enhanced to dynamically generate WOs, adjust preventive schedules, and log asset lifecycle events more intelligently.
Example: A predictive maintenance module alerts a technician to service a dialysis machine based on runtime hours and recent anomaly patterns, bypassing fixed schedule constraints.

c) AI/ML Layer

Components: Large Language Model Interfaces, AI Agents, ML Services
Function: This layer introduces intelligence into decision-making. AI agents can prioritize work orders, perform root cause analysis (RCA), generate reports, and offer contextual recommendations. It includes both pretrained models (e.g., LLMs) and custom-trained models (e.g., RUL predictors).
Example: When a technician describes an issue in natural language, an LLM interprets the input,

correlates it with historical failure data, and recommends corrective actions—including links to procedures or AR-based visual guides.

d) Data Access Layer

Components: APIs, Database Systems (SQL/NoSQL), Data Warehouses
Function: This layer ensures secure and efficient data exchange between CMMS, hospital Electronic Health Records (EHR), ERP systems, and other clinical or administrative platforms. It standardizes access through RESTful APIs or FHIR-based interoperability.
Example: Maintenance data for surgical robots is accessed alongside clinical usage logs from the hospital ERP, enabling correlation between usage patterns and wear rates.

e) User Interface Layer

Components: Mobile apps, Voice Assistants, AR/VR Interfaces, Dashboards
Function: This layer provides multimodal interaction tailored to different user roles (e.g., technician, manager, OEM). It supports mobile work order updates, voice commands, AR-based repair guides, and executive dashboards.
Example: A manager uses a dashboard to ask, "Why did asset downtime increase last week?", and receives a natural language explanation backed by AI correlation analysis.

7 System Architecture Overview

Figure 1 illustrates the high-level system architecture designed to support AI-augmented maintenance operations in hospital environments. This architecture embraces a layered, modular structure that enables seamless integration of real-time data sources (e.g., IoT, SCADA, MQTT), traditional CMMS functionalities, and cutting-edge AI/ML capabilities. At its core, the system promotes:

- Modularity, by clearly separating traditional maintenance workflows, AI services, and user interaction layers;
- Real-time integration, through standardized data acquisition APIs connected to diverse equipment sources;
- AI-centric design, featuring autonomous planning agents, predictive analytics, and self-healing IoT routines;
- User-centric interfaces, including voice assistants, augmented reality (AR), and mobile dashboards tailored for technicians, managers, and OEMs.

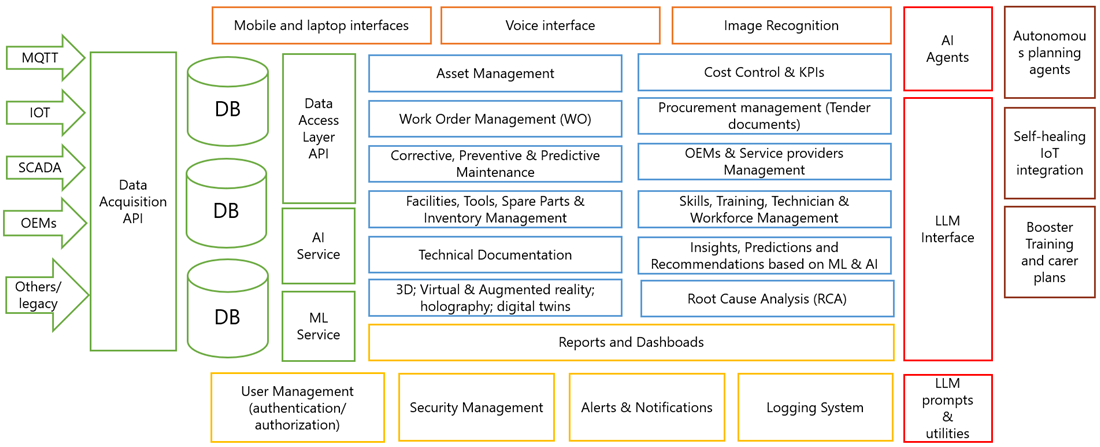


Figure 1: System Architecture overview

The architecture also incorporates robust data access layers, secure authentication, alerting systems, and support for large language models (LLMs), ensuring the system is interoperable, scalable, and resilient. It provides the foundation for transforming hospital maintenance from reactive to intelligent, proactive, and context-aware management.

8 Conclusion

The integration of Artificial Intelligence and Machine Learning into hospital maintenance systems represents a pivotal shift in how healthcare facilities manage their critical infrastructure. By augmenting traditional CMMS/EAM platforms with intelligent agents, real-time data processing, and predictive analytics, hospitals can move from reactive maintenance to truly proactive and optimized asset management.

The proposed multi-layered architecture ensures scalability, modularity, and interoperability, aligning with Industry 4.0 and human-in-the-loop AI principles. It empowers technicians and managers alike with natural language interfaces, real-time insights, and intelligent automation across every core function—from asset tracking and work order management to inventory forecasting and root cause analysis.

Beyond operational efficiency, this AI-driven approach enhances patient safety, reduces equipment downtime, optimizes resource utilization, and supports better clinical outcomes. The shift is not only technological but strategic, paving the way for resilient, self-aware, and data-driven healthcare infrastructure. As hospitals evolve to meet increasing demands and complexity, AI-powered maintenance systems will become an essential pillar of sustainable, high-performance healthcare delivery.

References

[1] E. m. Standard, NP EN 13306 - 2007 - Terminologia Da Manutenção.

[2] J. T. Farinha, Manutenção – A Terologia e as Novas Ferramentas de Gestão, Almedina, 2011.

[3] Farinha, José Torres, et al. "CMMS–An integrated view from maintenance management to on-line condition monitoring." Colecao: http://hdl. handle. net/10316.2/33309 (2014).

[4] Malta, Ana, Mateus Mendes, and Torres Farinha. "Augmented reality maintenance assistant using yolov5." Applied Sciences 11.11 (2021): 4758..

[5] “Pirâmide populacional do Portugal,” 2025. [Online]. Available: https://population-pyramid.net/pt/pp/Portugal, [Last access on 2025-06-23].

[6] S. Nadella, “Microsoft CEO’s Shocking Prediction: “Agents Will Replace ALL Software",” [Online]. Available: https://www.youtube.com/watch?v=uGOLYz2pgr8. [Last access on 2025-06-23].

[7] European maintenance Standard, “NP EN 15341:2009 Manutenção. Indicadores de desempenho da manutenção (KPI),” 2009.

Chapter 42

Impact of 5S tool on reducing setup times - case study in a metalworking industry

Contribution #81
Cite as:
Ana Maia, Ana Vieira. Impact of 5S tool on reducing setup times - case study in a metalworking industry. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-81,
  title ={Impact of 5S tool on reducing setup times -
     case study in a metalworking industry},
  author={Ana Maia and  Ana Vieira},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Impact of 5S tool on reducing setup times - case study in a metalworking industry

Ana Isabel Duarte Maia[1*], Ana V Vieira[1,2,3,4]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[2] RCM2+ Research Centre for Asset Management and Systems Engineering, Polytechnic Institute of Coimbra, Coimbra Institute of Engineering, Rua Pedro Nunes - Quinta da Nora, 3030-199 Coimbra, Portugal

[3] CISE - Electromechatronic Systems Research Centre, Universidade da Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal

[4] Ci2 - Smart Cities Research Center, Polytechnic Institute of Tomar, Portugal

[*]Corresponding author
maia.anamaia022@gmail.com, avieira@isec.pt

Abstract

The growing need to enhance the efficiency of production processes has driven organisations to adopt continuous improvement methodologies. This article examines the impact of implementing the 5S tool in reducing setup times, highlighting its role in the successful deployment of SMED (Single-Minute Exchange of Die).

Initially, the principles of 5S, its benefits, and the main challenges associated with its application are presented, emphasising its contribution to the standardisation of the working environment. Subsequently, the SMED methodology is explored as a strategy for minimising the setup times of industrial equipment and production processes. A connection between the two tools is also established, demonstrating how the effective application of 5S can directly influence the success of SMED. Finally, a practical case study of 5S implementation in a manufacturing setting is presented, in which the reorganisation of drawers for materials, tool shelves, and files containing technical drawings required for production contributed to a reduction in setup times.

The results indicate that, through small improvements, combined with organisation, discipline, and the standardisation of tasks, it is possible to achieve operational gains, reinforcing the importance of fostering and sustaining a culture of continuous improvement within industry.

Keywords: Lean, 5S, SMED, Process Optimization, Quality Control

1 Introduction

The quest to be a little better every day and to achieve high operational efficiency is a priority in contemporary industries. In the field of production, reducing machine set-up times is key to increasing flexibility and productivity. Among the various methodologies applied, 5S and SMED stand out for their effectiveness, simplicity and complementarity.

5S is an approach that promotes organisation and cleanliness in the workplace, facilitating quick access to tools and materials, while SMED's main objective is to reduce tool changeover times between operations and different production batches. Although each tool can bring benefits individually, applying them together proves to be particularly advantageous, since the organisation provided by 5S supports and speeds up the setup

change process envisaged by SMED.

This article presents the integrated implementation of these methodologies in a metalworking company, detailing the steps taken and the results obtained, with a focus on improving setup times and the organisation of the workspace. It also reinforces the idea that small improvements, implemented gradually and with little investment, can, in the long term, generate significant gains and advantages in process optimization

2 State of the art

Of Japanese origin, 5S is a methodology structured into five main stages - Seiri, Seiton, Seiso, Seiketsu and Shitsuke - which seeks to establish standards of organisation, cleanliness and discipline. At the same time, this practice is part of a management philosophy that helps transform organisational culture [1] [2] [3] [4].

2.1 Brief description of each S

The first S, Seiri, or sense of use, is related to organising the workplace to keep only the items that are really necessary, according to their frequency of use, eliminating anything unnecessary. This practice reduces clutter and obstacles to productivity by facilitating quick and efficient access to the resources needed for production.

The second S, Seiton, or sense of order, seeks to ensure that everyone knows where to look for and store objects, by standardising names, using labels and colours, separating different objects and defining specific locations, avoiding obstructions in passageways. This reduces searching time, improves logistics, avoids unnecessary purchases and damage, as well as facilitating cleaning and organisation. The third S, Seiso, sense of cleanliness, highlights the importance of keeping everything clean as a synonym for quality and safety, encouraging each person to clean up after themselves, conserve equipment, dispose of rubbish and keep areas under defined responsibility, which increases productivity, minimises losses and favours good relations.

The fourth S, Seiketsu, a sense of standardisation, involves hygiene and maintaining cleanliness and order, reflected in the employees' way of life, requiring application of the previous three Ss, training, elimination of unsafe conditions and promotion of a cooperative environment, bringing safety, damage prevention, improvement of the organisational image and greater motivation. Finally, the fifth S, Shitsuke, a sense of self-discipline, represents commitment to continuous improvement, through creativity, good communication, alignment with the organisation's values, continuous training and evaluation of the programme, facilitating operations, avoiding losses and providing predictability in results [1] [2] [3] [4].

All these pillars combined help to maintain an organised, clean and functional working environment, promoting gains in both productivity and quality.

2.2 Benefits of applying the 5S

The methodology under study has several advantages, particularly in terms of productivity, safety, quality, elimination of waste and employee involvement. It is associated with increased product quality and the creation of a safer and more organised working environment. Better organisation and cleanliness help to minimise accidents at work and improve the organisational climate, generating more favourable working conditions and higher performance [2] [3] [4] [5] [6] . This tool contributes to the elimination of hazards and risks in the workplace, enhancing worker safety regardless of the complexity of the tasks performed. Typically, following the implementation of 5S, a reduction in overall operational time is observed. Process standardisation and waste elimination become evident, leading to a decrease in unnecessary movements and optimisation of quality and productivity [2] [3] [4] [5] [6].

By creating a standardised workplace, companies are better able to effectively apply Lean principles aimed at waste reduction. The implementation of 5S fosters a more efficient and organised working environment, reducing the likelihood of errors and defects in products, such as weld joint failures, which can lead to high rework costs and compromise the safety of the final product [2] [3] [4] [5] [6].

Furthermore, this methodology contributes to sustainability by optimising the use of materials and resources. The elimination of excess and proper organisation help reduce waste and improve resource

efficiency, resulting in a lower environmental impact.

Notably, Japanese companies focus on engaging individuals and workgroups at all levels, both in the development and execution of 5S, promoting an inclusive approach that strengthens commitment and ownership of the processes. The methodology fosters a culture of discipline, enhancing workers' dedication and commitment, which in turn leads to a more efficient working environment.

Both 5S and Kaizen value employee involvement in process improvements. The implementation of 5S encourages workers to take ownership of their workspaces and contribute ideas for further enhancements, thereby empowering the workforce. The discipline aspect (shitsuke) involves training and education aimed at boosting morale and work quality, promoting a culture of self-discipline and continuous improvement [2] [3] [4] [5] [6].

2.3 Integration with other Lean tools

Just as a production process is composed of various stages and interconnected with other processes, Lean tools can also be integrated with one another. In this regard, the 5S methodology is often regarded as a fundamental component of Lean Manufacturing. 5S provides a systematic approach to organizing and standardizing the workplace, which is essential for the effective implementation of Lean principles. By establishing order and cleanliness, 5S helps to minimize waste and enhance efficiency, thereby supporting Lean objectives [5].

Alongside 5S, there are also tools such as TPM, Kaizen, the Kanban System, SMED, among others. Integrating 5S with TPM can result in more efficient and productive maintenance, as an organized and clean environment makes it easier to identify problems and carry out preventive maintenance activities. This, in turn, reduces machine downtime and improves equipment availability [6].

On the other hand, the Kaizen philosophy, which focuses on continuous improvement, complements 5S by promoting incremental, efficiency-driven changes. The implementation of 5S can establish a solid foundation for applying Kaizen practices, where employees are encouraged to consistently identify and implement improvements [6].

The implementation of 5S can also facilitate the adoption of the Kanban system, which is used to manage the flow of materials and production. An organized and clean environment helps to provide a clearer view of stock levels and operations, thereby enhancing the efficiency of Kanban utilization [6].

The 5S tool is also related to the Lean tool used for reducing setup times, SMED (Single Minute Exchange of Die). Several studies have been conducted in which each activity involved in the tool or production changeover process was recorded, detailed, and timed to comply with the application rules of the SMED methodology. All process activities were categorized into two groups: external activities (those that can be performed while the machine is running) and internal activities (those requiring the machine to be stopped). Following this stage, all subsequent phases of SMED were implemented with the aim of reducing setup time and ensuring the process is organized and ready at the moment of changeover [3] [4] [7].

With the combined implementation of SMED and 5S, changeover time is effectively reduced, demonstrating that these methodologies are complementary and interdependent within the Lean environment. Certain activities, such as the removal of unnecessary, damaged, or unused materials, contribute a significant proportion to time reduction in the search for tools or materials. Alongside SMED, the 5S methodology is applied in areas where disorganisation leads to time wasted searching for needed items, whether in product and parts warehouses, workbenches, shelves, or drawers. In this context, 5S has made a significant contribution to overall efficiency by eliminating waste related to transportation, waiting time, and inventory, providing not only productivity gains but also better space utilization, enhanced safety, and greater worker engagement [7].

2.4 Challenges faced in the application of 5S

There are several factors that influence the unsuccessful implementation of the 5S method. One of the main obstacles is the lack of adequate knowledge, training, and awareness among professionals regarding the methodology to be implemented. Furthermore, there is resistance to change from both workers and management, who often prefer to maintain traditional practices, even if disorganised and inefficient. The

traditional organisational culture and the mindset that old methods are sufficient hinder the introduction of 5S, as does the lack of pursuit of daily perfection and continuous improvement. This resistance may be influenced by perceptions of increased workload, the level of autonomy at work, and employees' self-efficacy. Other contributing factors include limited financial resources to conduct training and maintain system standards, as well as the absence of consistent implementation and a lack of support and commitment from senior management. These combined obstacles make effective application of 5S difficult, undermining the potential benefits of improved organisation, safety, and efficiency in construction processes [8] [9].

Another important contribution is the suggestion that improving workplace autonomy and promoting employees' self-awareness can help reduce resistance by providing practical guidance for change management strategies. Studies have shown that resistance is not merely a theoretical barrier but an empirical reality with practical effects that senior managers and supervisors must consider. When workers feel valued and supported by the organisation, they develop a positive attitude towards organisational initiatives, including change processes. This perception fosters feelings of loyalty, commitment, and responsibility towards the organisation, making it more likely that employees will actively participate in change initiatives and prepare themselves mentally and emotionally for implementation, thereby reducing resistance and facilitating smoother transitions. Therefore, management that promotes fairness and establishes high-quality relationships between leaders and subordinates can diminish employees' resistance to organisational change [9] [10].

2.5 Evolution from 5S to 6S and 7S

To meet the need for ensuring adequate levels of occupational health and safety, especially in industrial environments with a high risk of accidents, a sixth "S" was introduced, dedicated to Safety. Its origin stems not only from concerns about the organisation and efficiency of the workspace but also from compliance with safety requirements, machine certifications, and the use of personal protective equipment (PPE) by workers [11].

The 6S procedure involves three main stages: workspace organisation, process standardisation, and the adoption of specific safety measures. These actions enable workers to have access to personalised protective resources, thereby strengthening accident prevention. The Safety phase is considered essential to ensure that all operations comply with occupational health and safety regulations.

In this way, the application of 6S has a positive impact on risk reduction and the promotion of a safety culture, contributing to the goal of achieving zero accidents. This evolution of 5S has become an essential strategy for companies seeking to optimise their production processes and protect their employees [11].

The Lean methodology, particularly the 5S and 6S tools, is already widely used by companies across various sectors to improve organisation, productivity, safety, and quality in the workplace. As a natural evolution of this process, the 7S methodology was proposed, incorporating Sustainability as the seventh S. This expansion aims to consolidate management practices that combine operational efficiency, ethics, social responsibility, and environmental commitment. The transition from Lean 5S to 7S reflects the growing needs of organisations regarding safety, sustainability, and positive social impact [1] [11].

3 Methodology

In a factory environment, in a metalworking industry, a 5S programme was implemented in a sector of the factory with the aim of reducing the setup time of a sheet metal forming machine, a bending machine.

After analysing the tasks and the corresponding times for each one, it was concluded that applying 5S in certain areas would make it easier to find the tools and materials needed for production, reduce the time taken to carry out some external tasks and, consequently, reduce the total setup time.

Initially, the average setup time was 30 minutes, with a large part of this time corresponding to setting up the machine and another significant part to finding and placing the tools (dies and punches) in the machine. Although the 5S activities do not focus on the activity that requires the most time, a reduction in setup time is expected, albeit modest. All small improvements are accounted for and should be valued.

Firstly, the 5S method was applied to the four files containing the technical drawing sheets needed for production. To carry out the first S, repeated sheets were removed, as well as old sheets relating to discontinued products or prototypes that were no longer needed.

The technical drawings were then organised by product category, and there are now five files. Each one now has a reduced number of sheets compared to the initial situation, which are ordered numerically according to the part code. To complete this second stage of organisation, an index was created for each file, making it easier to identify the drawing corresponding to the part to be produced.

As part of the third S, the dirty and deteriorated micas were replaced with new, clean and transparent micas, and new, clean dossiers were added.

Before moving on to the last stage of 5S, the procedure was standardised by placing different coloured labels on each file, each colour corresponding to a category of part. This measure makes it quicker to select the correct dossier, using visual memory.

Finally, to support the last S, regular audits will be carried out, in conjunction with the monthly sector audits, with the aim of ensuring that order and cleanliness are maintained and that there is an ongoing commitment to continuous improvement.

The pictures below represent the before, Figure 1, and the after of 5S application, Figure 2.



Figure 1: File shelves before 5S tool was applied



Figure 2: File shelves after applying the 5S tool

After the new organisation of the files and technical drawings, it is expected that this task will take only 30 seconds. Instead of a minute and a half as in the initial situation.

As well as organising the files with the technical drawings for each part to be produced, the layout and arrangement of the bending tools, punches (normal punches, knife punches and special punches) and dies was also organised.

Initially, the tools were not labelled or given a defined place to be stored. The machine technician would spend some time looking for the right tool for production and, once he had used it, he would place it in a free space, which may or may not correspond to where he had taken it from.

To begin with, tools that were not being used or were damaged were removed, in accordance with the first S of 5S (Seiri), which consists of selecting and eliminating unnecessary material.

Subsequently, an identification was defined for all the tools, including dies and punches. The technique selected for engraving the tool identification was laser engraving, due to its high durability, as can be seen in Figure 3 and Figure 4.

Once identified, the tools were organised vertically using welded stainless-steel rails specifically designed for this purpose. This measure corresponds to the second S (Seiton), related to ordering and defining specific locations for each item. This arrangement allows each tool to be stably supported and separated from the others, making it easier to see immediately and to handle. In addition, this solution contributes to better utilisation of the available space, since previously the tools were stored horizontally on a shelf, occupying a significantly larger area and making it more difficult to access the desired tool.

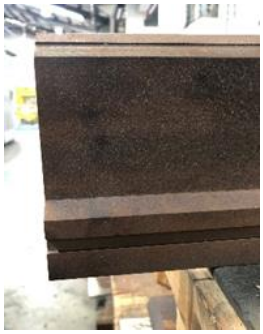


Figure 3: Standard punch, without identification



Figure 4: Standard punch with laser engraving

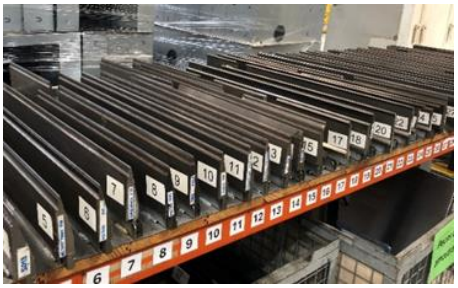
To complement this organisation, once the space had been cleaned, numbered labels were applied to each rail, and to each tool with the number of the rail it corresponded to, reinforcing storage standardisation, quick identification and storage in the correct place. A list was also drawn up of all the existing tools and their location. The aim of this labelling system is to reduce search time and minimise the likelihood of errors or exchanges between tools.

Finally, the new organisational system and the procedures for using it were explained to the employees. The check will be included in the monthly continuous improvement audits, ensuring that 5S is maintained over time.

In this way, based on the analyses carried out, it is expected that the average time taken to find a tool will drop from approximately 2 minutes to just 30 seconds.

Table 1 below shows the comparison between the initial situation and the result after implementing these improvements.

Table 1: Initial situation vs final situation

Before		
After		

4 Results

The results obtained confirm the success of the implementation of these tools. In the first activity, related to the organisation of the folders, the average search time was reduced from one and a half minutes to just 30 seconds. In the second activity, concerning the tidying and identification of the tools, the search time decreased from two minutes to 30 seconds as well. Overall, there was a reduction of approximately two and a half minutes per setup.

5 Discussion

The combined implementation of SMED and 5S has demonstrated, as highlighted in the state of the art, that these methodologies are complementary and interdependent within the context of Lean production. The 5S method created the necessary foundation by eliminating wastes related to transport, waiting time, inventory, and the search for materials, thereby enhancing the effectiveness of SMED actions aimed at reducing setup time. In this specific case, the application of 5S involved activities such as the removal of unnecessary, damaged, or unused materials, which significantly contributed to the reduction in the time spent searching for tools and technical drawings.

With the combination of the two methodologies, particularly the 5S tool, it was possible to save two and a half minutes per setup. Although this gain may appear modest on the scale of an isolated operation, it is important to emphasise that, considering the execution of six setups per day, the accumulated impact over the course of a week, a month, and a year results in a significant saving of time, with direct effects on productivity, efficiency, and resource utilisation. In addition to the reduction in time, further benefits can be observed, such as better use of space, increased safety, and greater worker engagement.

6 Conclusion

The results obtained show that combining various Lean tools is an asset for the success of companies' improvement projects. Applying the 5S tool to achieve SMED objectives is an effective approach to reducing setup times and optimising production processes in a shop floor environment. In this work, it was found that seemingly small improvements, which are quick to implement and require low investment, can produce significant gains and create the necessary conditions for more complex and large-scale interventions.

In this way, it can be concluded that it is essential to begin continuous improvement activities with the most immediate aspects that can be addressed swiftly, building a solid foundation that will later allow more structural problems with greater operational impact to be tackled. This gradual path not only promotes the achievement of measurable short-term results but also supports the development of a culture of continuous improvement and commitment among the teams involved.

Acknowledgments: We appreciate the support provided by the RCM2^{+} (Research Centre for Asset Management and Systems Engineering) and by ISEC (Coimbra Institute of Engineering). Our thanks are extended to the company that provided the opportunity to develop and execute the project, fundamental to the development of the work, as well as all the personnel involved.
This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Projects UIDB/04131/2020 and UIDP/04131/2020.

References

[1] J. F. Carrera, A. A. Del Olmo, M. R. Cuadrado, M. D. M. E. Escudero, and L. R. Cuadrado, "From lean 5s to 7s methodology implementing corporate social responsibility concept,"

Sustainability (Switzerland), vol. 13, no. 19, Oct. 2021, doi: 10.3390/su131910810.

[2] R. Gapp, R. Fisher, and K. Kobayashi, "Implementing 5S within a Japanese context: An integrated management system," *Management Decision*, vol. 46, no. 4, pp. 565–579, 2008, doi: 10.1108/00251740810865067.

[3] J. M. T. Farinha, *Asset Maintenance Engineering Methodologies*. 2018.

[4] J. M. T. Farinha, *Manutenção - A terologia e as Novas Ferramentas de Gestão*, 1st ed. Monitor, 2011.

[5] J. R. Correa Castañeda and G. A. Montoya Cárdenas, "5S Methodology: literature review and implementation analysis," *Journal of Scientific and Technological Research Industrial*, vol. 3, no. 2, pp. 37–45, Dec. 2022, doi: 10.47422/jstri.v3i2.30.

[6] C. Manzanares-Cañizares, A. Sánchez-Lite, V. F. Rosales-Prieto, J. L. Fuentes-Bargues, and C. González-Gaya, "A 5S Lean Strategy for a Sustainable Welding Process," *Sustainability (Switzerland)*, vol. 14, no. 11, Jun. 2022, doi: 10.3390/su14116499.

[7] M. Bukhsh *et al.*, "Productivity Improvement in Textile Industry using Lean Manufacturing Practices of 5S & Single Minute Die Exchange (SMED)," 2021.

[8] S. Ahmed, M. M. Hossain, and I. Haq, "Implementation of lean construction in the construction industry in Bangladesh: awareness, benefits and challenges," *International Journal of Building Pathology and Adaptation*, vol. 39, no. 2, pp. 368–406, Mar. 2021, doi: 10.1108/IJBPA-04-2019-0037.

[9] F. Jaramillo, J. P. Mulki, V. Onyemah, and M. R. Pesquera, "Salesperson resistance to change: An empirical investigation of antecedents and outcomes," *International Journal of Bank Marketing*, vol. 30, no. 7, pp. 548–566, 2012, doi: 10.1108/02652321211274318.

[10] N. Rehman, A. Mahmood, M. Ibtasam, S. A. Murtaza, N. Iqbal, and E. Molnár, "The Psychology of Resistance to Change: The Antidotal Effect of Organisational Justice, Support and Leader-Member Exchange," *Front Psychol*, vol. 12, Aug. 2021, doi: 10.3389/fpsyg.2021.678952.

[11] M. Jiménez, L. Romero, J. Fernández, M. del M. Espinosa, and M. Domínguez, "Extension of the Lean 5S methodology to 6S with an additional layer to ensure occupational safety and health levels," *Sustainability (Switzerland)*, vol. 11, no. 14, 2019, doi: 10.3390/su11143827.

Chapter 43

Towards an Automated System for Pig Aggression Detection and Tracking

Contribution #82
Cite as:
Roberto Costa, Ricardo P.M. Cruz, Bruno Ferreira, Nuno Lavado, Luísa Martin, Isabel Sa, Bruno Ferreira, Lusa Martin, Amelia Ramos, Roberto Costa, Ricardo P. M. Cruz, Nuno Lavado. Towards an Automated System for Pig Aggression Detection and Tracking. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-82,
  title ={Towards an Automated System for Pig
     Aggression Detection and Tracking},
  author={Roberto Costa and  Ricardo P.M. Cruz and
     Bruno Ferreira and  Nuno Lavado and  Luísa Martin
     and  Isabel Sá and  Bruno Ferreira and  Luísa
     Martin and  Amélia Ramos and  Roberto Costa and
     Ricardo P. M. Cruz and  Nuno Lavado},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Proceedings of PAMDAS 2025 - International Conference on Physical Asset Management and Data Science.
Coimbra, Portugal, 17/18 July 2025. https://pamdas.rcm2.pt/event/1/

Towards an Automated System for Pig Aggression Detection and Tracking

Isabel Sá[1], Bruno Ferreira[2], Luísa Martin[3,4], Amélia Ramos[3,4], Roberto Costa[3,4], Ricardo P. M. Cruz[2,5], Nuno Lavado[6,7,*]

[1]University of Porto, Faculty of Sciences, Portugal
[2]University of Porto, Faculty of Engineering, Portugal
[3]Polytechnic University of Coimbra, Coimbra Agricultural School, Portugal
[4]Research Center for Natural Resources, Environment and Society (CERNAS), Polytechnic University of Coimbra, Portugal
[5]Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), Porto, Portugal
[6]Polytechnic University of Coimbra, Coimbra Institute of Engineering, Portugal
[7]Research Centre for Asset Management and Systems Engineering (RCM2+), Polytechnic University of Coimbra, Portugal
[*]Corresponding author: nlavado@isec.pt

Abstract

Aggressive behavior in pigs within intensive farming systems poses significant welfare and productivity challenges, yet manual monitoring remains laborious and inconsistent. To address this, we developed an automated deep learning approach combining a VGG16 convolutional backbone with a bidirectional LSTM to detect aggression from video sequences, enhanced by interpretability tools (Grad-CAM, Captum) for model diagnostics. While classification performance was limited (43.9% accuracy, 1.22% aggression recall), mainly related to insufficient variability and quantity of annotated aggressive behaviour episodes at segmentation level, saliency maps confirmed the model's focus on biologically relevant features (e.g., head contacts). We further adapted the dataset for multi-object tracking (MOT) using the MOTChallenge framework, achieving moderate tracking performance (54.4% IDF1, 69.1% MOTA) via CSTrack integration. Despite current limitations, this work establishes a critical foundation for precision livestock monitoring by (1) demonstrating the value of explainable AI in behavior analysis and (2) providing a standardized pipeline bridging action recognition with identity-aware tracking—key advancements toward scalable, automated welfare assessment in real-world farming conditions.

Keywords: automated livestock behavior monitoring; VGG16 LSTM animal behavior classification; explainable AI Grad-CAM livestock farming; multi-object tracking (MOT) pigs precision farming; MOTChallenge adaptation animal tracking.

1 Introduction

Understanding and monitoring the behaviour of pigs in livestock environments is essential to ensure animal welfare, prevent injuries, and optimise management practices. Aggressive interactions among pigs, in particular, can lead to stress, wounds, and reduced productivity. However, manually identifying such behaviours from video recordings is time-consuming and prone to inconsistencies.

In this context, the present work proposes an automated pipeline for detecting aggressive behaviours in pigs using deep learning techniques applied to short video sequences. The approach builds upon the work of Chen et al. [1], who demonstrated the effectiveness of combining convolutional neural networks (CNNs) with recurrent layers to classify aggression in pig pens. Our model adopts a similar structure, using a pre-trained VGG16 network to extract frame-level visual features, which are then processed by a bidirectional Long Short-Term Memory (LSTM) layer to capture the temporal dynamics of behaviour.

To enhance the reliability and interpretability of the model, we integrated diagnostic tools such as Grad-CAM and prediction entropy analysis. These tools provide insights into the decision-making process of the neural network and help identify potential learning failures or biases in the data. Despite the implementation of early stopping and data augmentation strategies, our results indicate that the model struggled with generalisation, particularly in recognising aggressive events—highlighting the challenges of learning complex behavioural patterns from limited annotated data.

In a subsequent phase of the project, we explored multi-object tracking (MOT) techniques as a complementary strategy for behaviour analysis. Models such as FairMOT [2] and CSTrack [3] combine object detection and re-identification to track individual pigs over time. However, directly applying these models to farm scenarios requires careful adaptation of data and processing pipelines to account for the specific conditions of rural environments, such as occlusions, low contrast, and variable lighting.

This document is structured as follows:

- Section 2 presents a literature review, addressing works related to behaviour recognition, multi-object tracking, and model interpretability in computer vision, as well as the evaluation metrics used to assess classification and tracking performance.
- Section 3 details the adopted methodology, including data preparation, model architecture, the training process, and the interpretability techniques applied. It also discusses the results obtained, including classification performance, Grad-CAM visualisations, and tracking performance using MOT-compatible models.
- Section 4 presents the conclusions of the work and proposes future directions for improving models and their applications in production environments.

2 State of the Art

Automated video analysis for monitoring pig behaviour in intensive farming systems has received increasing attention due to its potential to enhance animal welfare monitoring, reduce labour costs, and provide continuous behavioural insights. Computer vision techniques, especially those based on deep learning, allow for the detection of behaviours such as aggression, lethargy, or excessive movement—critical indicators in modern livestock management.

2.1 Classical Approaches

Early works primarily employed handcrafted features and rule-based methods. For instance, Kashiha et al. [4] applied grayscale thresholding combined with blob detection to identify pigs, fitting ellipses to represent their bodies. Tracking relied on manual identity assignment at the start of each video sequence. These detections facilitated the extraction of metrics such as movement trajectories and feeding durations.

In related works, Chen et al. [5, 6] proposed motion-based analysis methods for detecting aggression, extracting features such as optical flow and velocity. These techniques provided the foundation for behaviour quantification, but their generalisation was limited. Our system

extends these approaches by employing learnable temporal models (LSTMs) to better capture dynamic behavioural patterns.

2.2 Deep Learning-Based Behaviour Recognition

To overcome the limitations of classical approaches, recent methods have turned to deep learning for more robust and scalable behaviour recognition. With the widespread adoption of deep learning, architectures based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly LSTMs, became the standard for behaviour recognition. Chen et al. [1] introduced a model that combines VGG16 with LSTM to classify aggressive behaviour in 30-frame video sequences, integrating Grad-CAM to visualise decision-relevant regions.

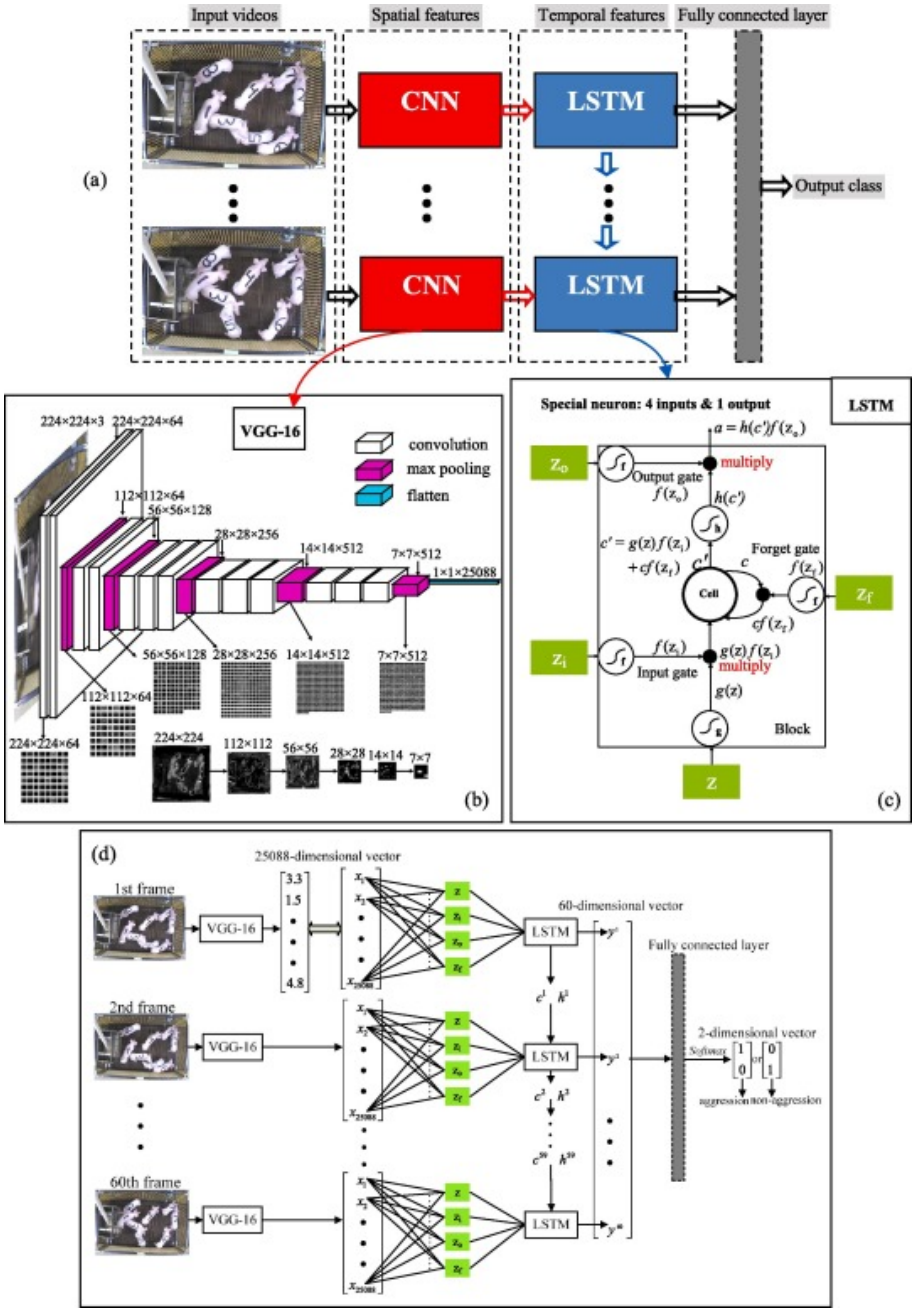


Figure 1: Schematic diagram of using CNN and LSTM to recognise aggressive episodes, adapted from Chen et al. [1].

Figure 1 presents a schematic diagram of the hybrid CNN-LSTM architecture used for recognizing aggressive pig behaviors, adapted from Chen et al. [1]. The model combines a VGG16 convolutional neural network (CNN) for spatial feature extraction from individual video frames with a bidirectional LSTM layer to analyze temporal dynamics across sequences. This architecture addresses both the visual cues (e.g., body posture, head contacts) and motion patterns characteristic of aggression.

Building on this paradigm, we implement a VGG16+LSTM model trained on balanced datasets of aggressive and non-aggressive video clips. To enhance interpretability, we integrated a custom Grad-CAM technique alongside the Captum library, enabling both spatial and temporal visualizations of the model's attention. This combined approach proved crucial for diagnosing misclassifications and understanding the model's focus on interaction zones and pig postures during aggression events.

2.2.1 Explainability in Deep Learning: Grad-CAM and Captum

As deep learning models become increasingly complex, understanding their decision-making processes has become crucial, especially in sensitive applications such as animal welfare monitoring. To address this, several explainability techniques have been developed to highlight the regions of an input that contribute most to a model's prediction.

One of the most prominent methods is **Grad-CAM** (Gradient-weighted Class Activation Mapping), introduced by Selvaraju et al. [7]. Grad-CAM generates coarse localization heatmaps by combining the gradients flowing into the last convolutional layers with their feature maps. This allows for the visualization of class-discriminative regions in an image, helping researchers and practitioners assess whether the model is focusing on semantically relevant areas—such as the heads or interaction zones between pigs when detecting aggression.

Building on this foundation, the **Captum** library was introduced as a unified framework for model interpretability in PyTorch. It includes implementations of various attribution methods such as Saliency Maps, Integrated Gradients, and Layer Grad-CAM. Captum abstracts the complexity of gradient and activation management, enabling more reliable and reproducible explainability workflows across diverse model architectures.

These explainability tools have been increasingly adopted in behaviour recognition pipelines, providing insights not only into model accuracy but also into the reliability and trustworthiness of predictions. In the context of animal behaviour, where wrong decisions may lead to welfare issues or misinterpretation of aggression, such visual diagnostics are essential for model validation and refinement.

2.3 Tracking and Individual Behaviour Monitoring

Accurate tracking of individual animals over time is essential for behaviour modelling and welfare analysis. Classical tracking-by-detection methods often failed under occlusion or lighting variations. To address this, recent approaches such as [3] introduced CSTrack, which combines detection, motion prediction, and re-identification (re-ID) for robust identity preservation.

In our implementation, we adopted the **CSTrack** framework and adapted it for pigs. We integrated detections from **YOLOv8**, a state-of-the-art object detector, and enhanced them with **appearance embeddings** — vector representations of individual visual features extracted from detected objects to help maintain consistent identity assignment across frames. For data annotation, we developed a conversion tool that takes as input **SAM-generated masks** (segmentations produced by the Segment Anything Model) and **COCO-style annotations** (a widely-used format for object detection datasets), outputting identity-preserving tracking files in **LCFormat** — a labeling convention compatible with multi-object tracking benchmarks such as the **MOTChallenge**.

This workflow enabled us to quantify complex behaviours such as repeated chasing, tail-biting, or isolation by associating actions with specific animals over time. Additionally, we extracted spatio-temporal statistics including time spent in different areas, social proximity metrics, and trajectory patterns.

2.3.1 Segmentation-Based Annotation Tools and COCO Format

Recent advances in object detection and tracking have increasingly relied on COCO-style annotations, which offer a standardised and flexible structure for representing segmentation masks, bounding boxes, and class labels. The pycocotools library has become a de facto standard for interacting with such annotations, enabling efficient manipulation of segmentation data and extraction of instance-level information. Originally developed for the MS COCO dataset [8], this toolkit is widely adopted in research and practical applications involving instance segmentation, particularly when converting pixel-wise masks into bounding boxes or preparing annotations for downstream tasks such as multi-object tracking (MOT). Its integration into pipelines facilitates compatibility with benchmark formats like MOTChallenge [9], bridging the gap between raw annotated data and tracking frameworks.

2.3.2 Evaluation Metrics

To comprehensively analyze the tracking accuracy of our model across different scenarios, we utilized the metrics derived from the MOT challenge based on pedestrian datasets [3]. The key metrics include MOTA, IDF1, MOTP, MT, ML, IDS, FP and FN.

Among these, **Multi-Object Tracking Accuracy (MOTA)** is considered the most critical metric in multi-object tracking, as it evaluates the overall accuracy of a multi-object tracking algorithm. The calculation formula for MOTA is:

$$\text{MOTA} = 1 - \frac{FN + FP + IDSW}{GT} \tag{1}$$

where GT represents the number of ground truth objects. The maximum value of MOTA is 1, while its minimum value can reach negative infinity. FN and FP denote the numbers of missed detections and false positives, respectively. $IDSW$ counts the instances of ID switches, indicating how often the tracker incorrectly assigns or changes object IDs. The smaller the $IDSW$, the better the tracker maintains object identity.

IDF1 measures the tracker's consistency in preserving object identity by combining ID precision and ID recall. The formula for calculating IDF1 is:

$$\text{IDF1} = \frac{2 \cdot IDTP}{2 \cdot IDTP + IDFP + IDFN} \tag{2}$$

where $IDTP$ represents the number of correctly matched objects, $IDFP$ denotes the number of incorrectly matched objects, and $IDFN$ refers to the number of missed objects.

Multiple Object Tracking Precision (MOTP) is a metric used to measure the positional error in tracking. It is defined as:

$$\text{MOTP} = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \tag{3}$$

where c_t represents the number of detection boxes that successfully match with ground truth in frame t, while $d_{t,i}$ measures the distance between matched pairs.

3 Methodology

This section details the methodology developed for detecting and analysing aggressive behaviours in pigs using deep learning techniques. The process encompasses dataset preparation, model development, training procedures, and the application of interpretability methods to evaluate the model.

3.1 Dataset Preparation

The dataset was compiled from video recordings of groups of pigs housed in pens, where episodes of both aggressive and non-aggressive behaviour were manually annotated by experts [1].

The initial dataset preparation followed an approach inspired by Chen et al., who similarly utilised short video sequences to classify aggressive behaviour in pigs. Each episode consists of a continuous sequence of 30 frames, corresponding to approximately 2 seconds of footage at 15 frames per second. The dataset was organised to ensure that training, validation, and testing sets were mutually exclusive, preventing overlap of episodes derived from the same video source. Classes representing aggressive and non-aggressive behaviours were balanced to avoid bias during training and evaluation. Data augmentation techniques such as horizontal and vertical flipping were applied to enhance model generalisation.

3.2 Model Architecture and Training

The implemented model employs a hybrid architecture combining a deep convolutional neural network for visual feature extraction with a recurrent layer to capture temporal dynamics of the episodes.

Specifically, a pre-trained VGG16 network serves as the feature extractor for individual frames. The extracted features are then fed into a bidirectional Long Short-Term Memory (LSTM) layer, which models temporal dependencies across the sequence.

The LSTM output is passed through fully connected layers to produce a binary classification distinguishing aggressive from non-aggressive behaviour.

Training was conducted using a cross-entropy loss function, optimised with the Adam algorithm with an initial learning rate of 1e-4. A learning rate scheduler (ReduceLROnPlateau) was employed to reduce the learning rate by a factor of 0.1 when the validation loss plateaued, with a patience of 10 epochs to allow recovery. Early stopping was implemented based on validation loss with a patience of 20 epochs to prevent overfitting, stopping training if no improvement was observed within this window. Training were executed using the high-performance server provided by INESC-TEC, equipped with NVIDIA A100 GPUs featuring 40 GB of VRAM.

[1]This work utilized expert-annotated videos of pig behavior from the R&D project 'PIGS+CARE' (2017–2019), funded by COMPETE 2020/PORTUGAL 2020 (FEDER) under grant POCI-01-0247-FEDER-017626 (Business R&D Incentive System).

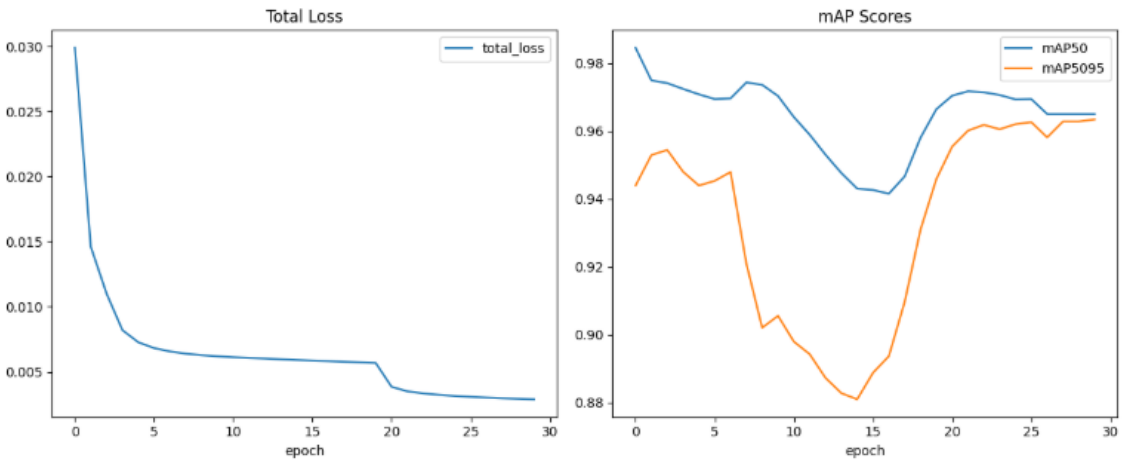


Figure 2: Evolution of total loss and mAP metrics during training.

Figure 2 illustrates the evolution of the total loss and mean average precision (mAP) metrics during the training process of the VGG16+LSTM model. The loss curve shows a steady decline, indicating effective learning, while the mAP curve reflects the model's improving detection performance over epochs. This visualization confirms that the training process achieved convergence, though the final performance on the test set (as detailed in Section 4.1) highlights limitations in generalization. The alignment of these metrics with early stopping (patience = 20 epochs) underscores the balance between preventing overfitting and maximizing model capability.

3.3 Explainability and Diagnostic Tools

To ensure that the model learns meaningful representations of behaviour, interpretability techniques based on attention mapping were applied, notably Gradient-weighted Class Activation Mapping (Grad-CAM), which visualises the regions of the video influencing the model's decisions. Additionally, gradients across layers were analysed throughout training epochs to detect issues such as vanishing or exploding gradients. Prediction entropy was also computed to assess the stability and confidence of the model's outputs during training. These interpretability tools enabled the generation of heatmaps over frames, highlighting areas receiving the greatest attention from the network and facilitating qualitative validation of the model's effectiveness.

These tools allow us to:

- Generate **Grad-CAM heatmaps** for individual frames, showing which regions the model attends to.
- Analyse **layer-wise gradients** to detect vanishing/exploding gradients or dead filters.
- Compute the **entropy of predictions** over time, identifying unstable sequences.
- Save **example visualisations** of both correctly and incorrectly classified sequences.

This interpretability layer was crucial for verifying whether the model focuses on relevant regions (e.g., heads, bites, collisions) when identifying aggression. To generate the Grad-CAM heatmaps, two different implementations were used: a custom method specifically adapted to our model's architecture, and the Captum library's abstraction for layer-wise attributions.

The custom Grad-CAM procedure operates by manually identifying the last convolutional layer of the VGG16 backbone and registering its activations and gradients during the forward and backward passes, respectively. The method bypasses the LSTM layers and focuses purely on the convolutional spatial features extracted from each frame. This approach enables us to

generate spatial saliency maps per frame, before temporal integration via the LSTM layer.
In parallel, the Captum library was used to verify the results using its LayerGradCam module. This approach automates much of the gradient handling but requires careful specification of the target layer and model handling for temporal sequences. Captum was applied frame-by-frame to isolate the convolutional attention per image prior to LSTM integration. While more abstract and modular, Captum may not capture the temporal dynamics present in the full sequence, which justifies the inclusion of a custom approach for complete interpretability.

3.4 Tracking Data Preparation

To enable multi-object tracking (MOT) evaluation under a standardised benchmark, we converted the annotated data from our pig behaviour videos into the MOTChallenge format, as described in A. Milan et al. [9]. Since the MOTChallenge benchmark and its associated code are primarily designed for human detection, we had to modify the codebase to allow detection and tracking of pigs moving in any direction during testing. This adaptation was essential to accurately represent the pig movements and orientations in our dataset. This format is widely used in the MOT community and allows consistent evaluation across datasets.

The conversion process involved parsing frame-level segmentation annotations (stored in COCO-style JSON files) and extracting bounding boxes via binary masks using the pycocotools library. Using the pycocotools library, we extracted object bounding boxes from binary segmentation masks. Each annotated object was associated with a unique identity (track_id) and class label (category_id), both preserved in the output. In accordance with the MOTChallenge specifications [9], each sequence was structured as follows:

- A folder named after the session, containing subfolders img1/, gt/, and det/.
- An img1/ directory containing sequentially named images (e.g., 000001.jpg).
- A gt.txt file with one line per ground truth bounding box in the format:

  <frame>, <id>, <bb_left>, <bb_top>, <bb_width>, <bb_height>, 1, 1, 1

- A det.txt file following the same structure, but with the ID set to -1 and containing the detection confidence:

  <frame>, -1, <x>, <y>, <w>, <h>, 1

- A seqinfo.ini file describing the sequence metadata (e.g., resolution, number of frames).

The dataset was split into training and testing sessions by evenly dividing the available annotated sessions. Two specific sessions, 09:17–09:30 and 09:30–09:42,were manually assigned to the training and test sets, respectively, to ensure that the training and test sets contained videos where episodes of violence occurred. The remaining sessions were divided evenly to maintain a balanced distribution across the sets. To enable seamless integration with CSTrack, we adapted the CSTrack.yaml configuration file to reflect the structure and location of our converted dataset. This included specifying the paths to training and testing sequences, adjusting dataset names, and ensuring compatibility with the MOTChallenge directory structure. By aligning the configuration with our data, we ensured that CSTrack correctly loaded the sequences, annotations, and detection hypotheses for training and evaluation. Figure 3 visualizes the discriminative ability of ID embeddings used in the tracking pipeline, which is critical for maintaining consistent identities across frames. The figure includes (a) an example frame with ground truth detections, (b) the corresponding ID embeddings, and (c) a cosine similarity matrix comparing embeddings within the frame. Additionally, (d–f) show template embeddings from previous tracklets and their similarity matrices, with red indicating high correlation. This analysis confirms that the

model learns distinct feature representations for individual pigs, though challenges in occlusion and appearance variation (discussed in Section 4.2) still lead to ID switches. The figure underscores the importance of robust re-identification features for improving tracking continuity in complex farming environments.

This structured approach enabled compatibility with tracking frameworks such as FairMOT and CSTrack, facilitating experiments focused on identity preservation and tracking performance.

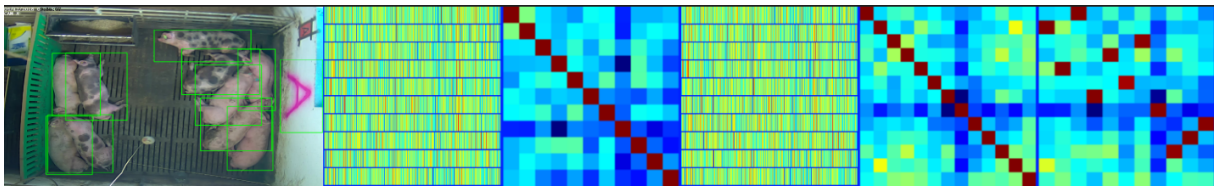


Figure 3: Visualization about the discriminative ability of ID embeddings: (a) Example Image with ground truth detection results; (b) ID embeddings of the current example frame; (c) Cosine metric matrix between two ID embeddings of the current example frame; (d) Templates (i.e., ID embeddings of the previous tracklets); (e) Cosine metric matrix between two ID embeddings of the template; (f) Cosine metric matrix between ID embeddings of the current example frame and the template. Red color indicates a higher correlation between two features.

Reference Format Specification

The MOTChallenge format requires all input images to be JPEG files named sequentially using six-digit filenames (e.g., 000001.jpg). Ground truth annotations and detection hypotheses are stored as comma-separated value (CSV) files, where each line corresponds to one object instance. As detailed in [9], the ground truth format is as follows:

> "The first number indicates the frame number, followed by a unique object ID. The next four numbers represent the top-left bounding box position and its width and height. The seventh value (1 or 0) acts as a flag to indicate whether the instance should be considered for evaluation. The last two values represent object class and visibility ratio."

This format enables detailed performance evaluation across identity preservation, detection accuracy, and occlusion handling.

4 Results

4.1 Model Performance on the Test Set

The quantitative evaluation results for the proposed VGG16+LSTM model on the test set are presented in Table 1. Despite the training process achieving convergence with early stopping (patience = 20), the model exhibited limited generalisation capability.

The model achieved an overall accuracy of 43.9%, with a notably high false negative rate (48.78%) for the aggression class. This indicates that the model failed to correctly identify the majority of aggressive episodes, misclassifying them as non-aggressive. The low true positive rate (1.22%) further emphasises the model's difficulty in recognising this behaviour pattern.

Such limitations directly affect the practical applicability of the model, particularly in scenarios where accurate detection of aggressive episodes is crucial for timely preventive or corrective actions. The high false negative rate may result in underreporting these events, reducing the effectiveness of automated behavioural monitoring systems.

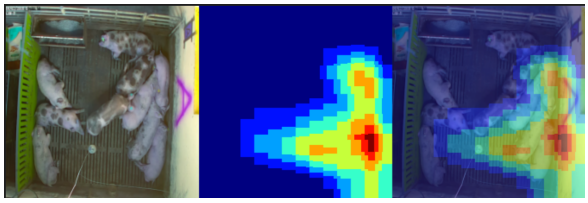
Table 1: Model performance on the test set

Metric	Value
Accuracy	43.9%

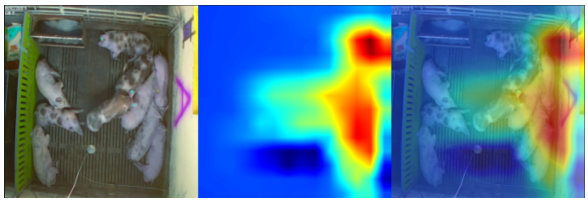
	Predicted Positive	Predicted Negative
Actual Positive	1 (1.22%)	40 (48.78%)
Actual Negative	6 (7.32%)	35 (42.68%)

In practice, the current model performance suggests it is not yet suitable for autonomous use in real environments, as failure to detect aggression could compromise swift interventions and animal welfare. It is recommended to explore strategies to improve model sensitivity, such as architecture adjustments, data balancing techniques, increasing the number of aggressive samples, or incorporating additional data modalities to enrich aggressive behaviour recognition.

4.1.1 Interpretability Analysis with Grad-CAM and Captum



(a) Grad-CAM visualisation generated using the custom method.



(b) Grad-CAM visualisation generated using the Captum library.

Figure 4: Interpretability methods applied to the aggressiveness classifier. Each image is composed of three parts: the original input frame (left), the standalone heatmap (centre), and the original image overlaid with the corresponding heatmap (right). This layout allows for a direct comparison of the saliency regions highlighted by the Grad-CAM methods.

To better understand the model's decisions, visual explanation techniques were employed. Grad-CAM was initially used to generate attention maps from the convolutional layers. Although some highlighted regions corresponded to pig interactions, in several cases, irrelevant areas—such as the food zone—were wrongly emphasised, even when no pigs were present.

Figure 4 demonstrates the interpretability methods applied to the aggression classifier, showcasing how Grad-CAM highlights regions of interest in input frames. Each panel displays the original frame (left), the standalone heatmap (center), and an overlay of the heatmap on the original image (right). This visualization reveals whether the model focuses on biologically relevant features (e.g., head contacts during aggression). These insights align with the quantitative results and underscore the value of explainability tools in diagnosing model behavior, even when classification performance is suboptimal.

Captum, a more granular attribution tool, yielded better consistency across a broader set of frames. In particular, it more reliably highlighted regions containing pigs, especially during aggressive events. This indicates that the model, albeit quantitatively weak, had some capacity to focus on semantically meaningful areas.

4.2 Tracking Results and Evaluation

To evaluate the performance of our tracking pipeline, we employed standard metrics from the MOTChallenge benchmark suite, including Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), ID F1 score (IDF1), number of ID switches (IDSW),

and mostly tracked targets (MT). The evaluation was conducted using the TrackEval toolkit, configured according to the MOT16 settings [9].

The tracking system recorded a total of 1705164 detections across sessions (ranging from 169884 to 274053 per session) and identified 9232 unique IDs, with session-specific counts varying from 504 to 1733.

Table 2 presents an evaluation of CSTrack's multi-object tracking performance on pig behavior analysis, using manually annotated ground truth data for validation. Focusing on key metrics, the table reveals that while the system achieved a moderate overall MOTA of 50.0%, indicating reasonable detection accuracy, it struggled significantly with identity preservation (IDF1 of just 29.0%) and accumulated 7883 identity switches across sessions. The data shows particular challenges in crowded scenarios, evidenced by high false negatives (507.9K) and false positives (394.3K), suggesting difficulties with occlusions and similar-appearing pigs. Notably, session 09:30 performed best (MOTA 69.1%), while 10:45 showed the weakest results (MOTA 37.9%), highlighting variability under different conditions. These results underscore the need for improved re-identification features and occlusion handling in agricultural computer vision applications.

Table 2: Tracking performance of CSTrack per session using manual annotations.

Session	**MOTA**	**IDF1**	**Recall**	**Precision**	**MT**	**FP (K)**	**FN (K)**	**IDS**	**FM**
09:30	**69.1**	54.4	86.3	83.6	7	38.4	31.1	538	2364
10:33	38.9	32.6	78.2	66.8	6	88.2	49.6	934	2623
10:45	37.9	19.4	79.5	65.9	6	93.4	46.7	1147	2856
10:58	51.2	20.9	68.4	80.2	5	38.3	71.8	971	2428
11:10	49.3	18.4	65.6	80.5	5	36.1	78.3	1055	2736
11:23	45.2	22.4	60.3	80.8	3	32.7	90.4	1605	3247
11:36	41.8	17.9	59.9	77.4	3	39.9	91.3	1319	3217
11:48	**66.4**	43.5	78.6	86.7	8	27.4	48.7	314	1237
Overall	**50.0**	29.0	72.1	76.9	43	394.3	507.9	7883	20708

It is worth emphasising that, despite having only 10 pigs in the experiment, the complexity of their interactions and close proximity complicate continuous and correct identity tracking. The detection component performs reasonably well, as indicated by recall and precision values of 72.1% and 76.9%, respectively, although false positives and false negatives still affect tracking quality. Out of a total of 181872 frames processed, the mostly tracked targets (MT = 43 out of 85 trajectories) show that around half of the objects are well tracked throughout their trajectories, nevertheless, there is room for improvement in recovering partial or lost tracks.

Behavioural Patterns and Tracking Robustness in Multi-Animal Pig Monitoring

The dataset comprises a total of 181872 frames after merging, captured from 10 pigs under observation. Among these, 153920 frames correspond to non-violent behaviours (label = 0), while 27809 frames correspond to violent behaviours (label = 1).

Behavioural distribution across the dataset indicates that the most frequently observed activity is lying (98325 frames), followed by social interaction (46244 frames) and head body knock (21603 frames). Less frequent behaviours include exploration, nosing, and interaction with objects.

Analysis of the average number of interactions per behaviour (Fig. 5) reveals that social interaction (11.41) and head body knock (10.30) are the most interactive behaviours, suggesting they play a significant role in the social dynamics of the pigs. Conversely, behaviours such

as nosing (7.43) and interaction with objects (8.26) show lower interaction averages, possibly indicating more solitary or less frequent activities.

When comparing violent and non-violent behaviours, the average number of interactions per instance is slightly higher in violent behaviours (9.66) than in non-violent ones (9.32), which may reflect the increased activity or complexity associated with aggression.

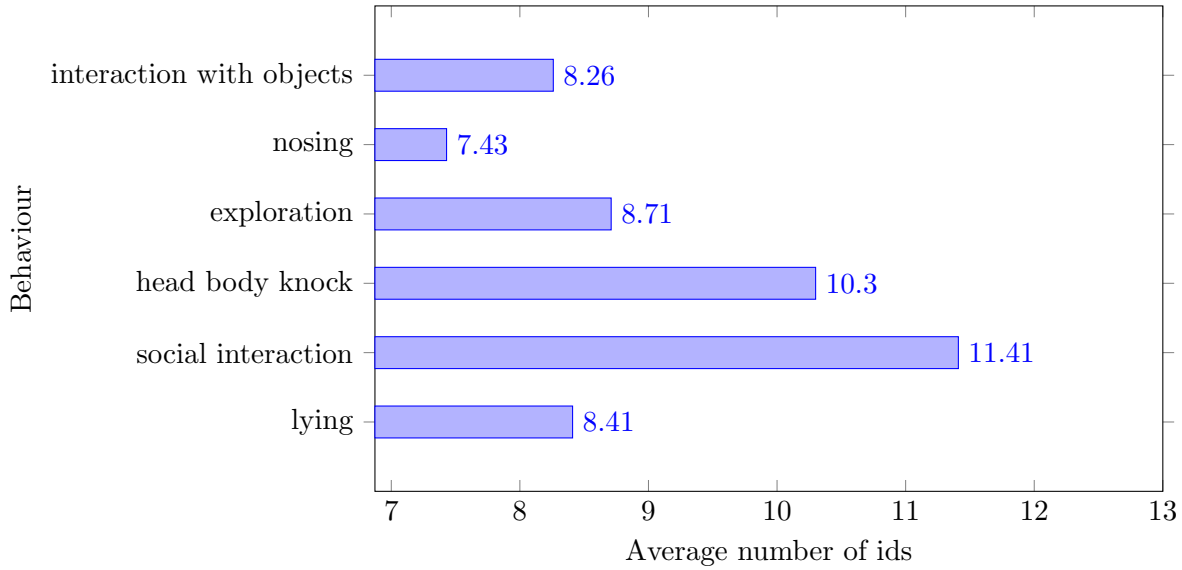


Figure 5: Average number of ids by behaviour type.

It is important to note that, although there are only 10 pigs in the dataset, the tracking system occasionally detects more or fewer targets than this number. Overestimations occur mainly due to false positives, such as the food area being mistakenly detected as a pig, likely caused by similar textures or shapes in the scene. Conversely, underestimations arise when pigs are partially or fully occluded by other animals, or when the detection fails due to motion blur or poor lighting conditions. These challenges contribute to identity switches and track fragmentation, underscoring the necessity for enhanced detection accuracy and more robust re-identification features to maintain consistent tracking identities.

These findings highlight the complex social behaviour of a relatively small group of pigs and the challenges posed to automated tracking systems. Improving both detection accuracy and identity association will be essential to enhance tracking continuity and reduce identity switches, thus enabling more precise behavioural quantification in future research. Overall, these results highlight the importance of enhancing both the detection robustness and the data association steps in the tracking pipeline to improve identity continuity and reduce fragmentation. Future work could focus on integrating better re-identification features and temporal consistency models.

5 Discussion

Despite the methodological rigor and comprehensive evaluation, the model's performance on the test set revealed significant limitations. The final accuracy was 43.9%, with a particularly low true positive rate (1.22%) for detecting aggression. Most aggressive episodes were misclassified as non-aggressive, resulting in a high false negative rate (48.78%).

These results suggest that, while the model learned to distinguish some features relevant to aggression, it failed to generalise effectively. Possible contributing factors include:

- **Insufficient variability and quantity** of annotated aggressive behaviour episodes, as

the dataset mainly contains interactions from the same pigs within the same pen, limiting the model's exposure to diverse aggression patterns.

- **Imbalance in intra-class complexity**, where non-aggressive behaviour was more homogeneous and thus easier to model, whereas aggression exhibited greater variation and subtlety.
- **Loss of temporal resolution**, as episodes were limited to short sequences (2 seconds), potentially missing longer-term behavioural cues needed for robust classification.
- **Underutilisation of motion cues**, with no optical flow or motion-specific features included in the model, reducing its ability to detect dynamic actions characteristic of aggression.

Our **PigsMot** dataset, which consists of carefully annotated pig movement sequences in a real farm environment, poses unique challenges such as occlusions, varying illumination, and complex interactions among animals. This dataset serves as a critical testbed to evaluate the adaptability of each tracking algorithm in practical, non-ideal conditions.

Results highlight the strengths and limitations of each approach, revealing that while some methods excel in controlled or synthetic environments, their performance may degrade in real-world scenarios like those captured in **PigsMot**. This comparison underscores the necessity for robust tracking solutions tailored to agricultural applications and animal behavior monitoring.

Table 3 presents a comparative evaluation of multiple state-of-the-art multi-object tracking (MOT) methods across various scenes, including our manually annotated dataset, **PigsMot**. The aim is to assess how well each tracker performs under distinct environmental conditions and levels of complexity.

The evaluated methods include classical and deep learning-based trackers widely adopted in animal monitoring and MOT benchmarks, as described by Liu et al. [10]. Metrics such as *MOTA* (Multiple Object Tracking Accuracy), *IDF1* (ID F1 Score), *FP* (False Positives), *FN* (False Negatives), and *IDs* (ID switches) are reported to provide a comprehensive overview of tracking quality.

Table 3: Comparison of tracking methods in different scenes.

Method	**Scenario**	**MOTA↑**	**IDF1↑**	**MT↑**	**ML↓**	**FP↓**	**FN↓**	**IDS↓**
Liu et al.- Deep-OC-SORT (YOLOv8)	night-in	64.1	71.9	5	3	0	6436	25
	night-out	28.2	38.5	1	9	0	22759	298
	daytime-out	90.0	78.2	13	0	575	527	244
Liu et al.- FairMOT	night-in	68.8	56.8	6	1	760	4780	85
	night-out	33.5	45.5	7	4	6266	14972	92
	daytime-out	90.5	70.7	12	0	535	585	161
Liu et al.- CSTrack	night-in	64.7	48.8	5	3	1	6191	79
	night-out	52.8	44.4	10	4	1142	13849	174
	daytime-out	91.1	69.4	11	0	333	679	194
Liu et al.- SDGTrack	night-in	78.1	81.9	9	0	1558	3371	11
	night-out	69.6	82.6	17	2	1678	9342	10
	daytime-out	95.0	90.8	12	0	176	500	3
Ours - CSTrack (PigsMot)	daytime-out	50.0	29.0	43	6	394300	507880	7883

Across public benchmark scenarios, particularly in controlled "daytime-out" settings, most methods (e.g., FairMOT, CSTrack, Deep-OC-SORT) demonstrate high MOTA scores above 90%. However, performance degrades notably in low-light or occlusion-heavy scenes, such as "night-out", where MOTA drops to as low as 28.2% (Deep-OC-SORT) or 33.5% (FairMOT).

In contrast, SDGTrack consistently outperforms others in all conditions, especially under day scenes, achieving both high MOTA (78.1–95.0)% and IDF1 (81.9–90.8)%, with very low ID switch rates.

When applying CSTrack to our PigsMot dataset, the performance drops substantially, with a MOTA of 50.0% and an IDF1 of only 29.0%. The tracker registered a very high number of false positives (394300) and ID switches (7883), which underscores the unique challenges of the dataset and suggests that existing tracking models, while effective in public benchmarks, struggle to generalize to high-density, unstructured animal behaviour scenes. This highlights the need for more specialized detection-reID integration and better temporal consistency models tailored for livestock behaviour analysis.

6 Conclusion

This work explored the use of deep learning techniques to automatically detect and interpret aggressive behaviour in pigs, with the overarching aim of improving animal welfare monitoring in intensive farming environments. The study covered the end-to-end pipeline from data collection and labelling to model design, training, evaluation, and interpretability. The main contributions of this work are:

- A behaviourally annotated dataset of pig interactions was curated, containing balanced episodes of aggressive and non-aggressive actions in video form. These episodes were standardised to 30-frame sequences at 15 FPS to capture short-term behaviour dynamics.
- A deep learning model architecture was implemented, combining a pre-trained convolutional neural network (VGG16) for spatial feature extraction with a bidirectional LSTM layer for temporal modelling. This hybrid structure was tailored to capture both the appearance and evolution of behaviours across time.
- A suite of interpretability and diagnostic tools was integrated into the training and evaluation pipeline. These included Grad-CAM (custom and Captum-based), gradient flow analysis, prediction entropy tracking, and visual inspection of both successful and failed predictions. This facilitated qualitative validation of the model's decision-making process.
- A methodology was developed to convert annotated behavioural sequences into the MOTChallenge format, enabling the application of multi-object tracking frameworks for future spatiotemporal analysis and integration with tracking-based approaches such as CSTrack.

The interpretability component of this work proved essential for understanding the model's behaviour. Grad-CAM maps showed that, in several cases, the model focused attention on semantically relevant regions such as pigs' heads and contact zones during aggressive events. However, there were also instances where attention was misdirected to irrelevant areas, such as empty feeding stations or walls, indicating incomplete learning.

Captum's more granular attribution approach produced more stable and focused visualisations, suggesting its value as a debugging and validation tool. Nonetheless, even with these insights, interpretability alone could not compensate for the lack of model generalisation, though it did offer valuable diagnostic information.

Given the limitations observed, several avenues for future research are proposed:

- **Data Augmentation and Expansion**: Increasing the number and diversity of aggressive behaviour samples, potentially through synthetic data generation or active learning techniques, could improve the model's generalisability.
- **Multimodal Integration**: Including additional data streams such as multiple cameras, depth maps, optical flow, or audio could enrich the feature space and provide stronger cues for aggression detection.

- **Model Architecture Refinement**: Exploring attention-based transformers, 3D CNNs, or temporal convolution networks (TCNs) may yield better representations of complex behaviour over time.
- **End-to-End Spatiotemporal Detection**: Integrating the classification model with a robust multi-object tracking pipeline would enable continuous, real-time aggression monitoring with spatial context.
- **Explainability-Driven Training**: Incorporating attention supervision or interpretability constraints during training might guide the model to focus on behaviourally relevant regions more consistently.
- **Spatiotemporal Integration with Tracking Pipelines**: Building upon the MOT Challenge annotation pipeline developed in this work, future research can incorporate CSTrack or similar object tracking methods to maintain individual pig identities across frames, such as the method SDGTrack used in [10]. This would allow for continuous behaviour monitoring, event localisation, and improved robustness in aggression detection by leveraging motion continuity and identity-preserving temporal context.

This work provided a solid foundation for the automatic detection of aggression in pigs using deep learning, combining both classification and interpretability in a single pipeline. While the results did not reach deployment-ready performance, the tools, analysis, and insights presented here serve as a valuable basis for future improvements.

Acknowledgments: This work utilized expert-annotated videos of pig behavior from the R&D project 'PIGS+CARE' (2017–2019), funded by COMPETE 2020/PORTUGAL 2020 (FEDER) under grant POCI-01-0247-FEDER-017626 (Business R&D Incentive System).

References

[1] C. Chen, W. Zhu, J. Steibel, J. Siegford, K. Wurtz, J. Han, and T. Norton, "Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory," *Computers and Electronics in Agriculture*, vol. 169, p. 105166, 2020.

[2] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International journal of computer vision*, vol. 129, pp. 3069–3087, 2021.

[3] C. Liang, Z. Zhang, X. Zhou, B. Li, S. Zhu, and W. Hu, "Rethinking the competition between detection and reid in multiobject tracking," *IEEE Transactions on Image Processing*, vol. 31, pp. 3182–3196, 2022.

[4] M. Kashiha, C. Bahr, S. Ott, C. P. Moons, T. A. Niewold, F. O. Ödberg, and D. Berckmans, "Automatic identification of marked pigs in a pen using image pattern recognition," *Computers and Electronics in Agriculture*, vol. 93, pp. 111–120, 2013.

[5] C. Chen, W. Zhu, C. Ma, Y. Guo, W. Huang, and C. Ruan, "Image motion feature extraction for recognition of aggressive behaviors among group-housed pigs," *Computers and Electronics in Agriculture*, vol. 142, pp. 380–387, 2017.

[6] C. Chen, W. Zhu, Y. Guo, C. Ma, W. Huang, and C. Ruan, "A kinetic energy model based on machine vision for recognition of aggressive behaviours among group-housed pigs," *Livestock Science*, vol. 218, pp. 70–78, 2018.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAMs: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[8] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[9] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: a benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831*, 2016.

[10] T. Liu, D. Jie, J. Zhuang, D. Zhang, and J. He, "Sdgtrack: A multi-target tracking method for pigs in multiple farming scenarios," *Animals*, vol. 15, no. 11, 2025. [Online]. Available: https://www.mdpi.com/2076-2615/15/11/1543

Chapter 44

Performance of Projects of Construction Companies in Portugal

Contribution #100
Cite as:
Caropul Mendes, Hugo Raposo, João Pedro Couto, J. Edmundo de-Almeida-e-Pais, Jorge Raposo, Torres Farinha. Performance of Projects of Construction Companies in Portugal. In *Proceedings of PAMDAS 2025 - International conference on Physical Asset Management and Data Science*, 17-18 Jul. 2025, Coimbra, Portugal. ISBN 978-989-8331-19-9.

```
@inproceedings{pamdas2025-100,
  title ={Performance of Projects of Construction
     Companies in Portugal},
  author={Caropul Mendes and  Hugo Raposo and João
     Pedro Couto and J. Edmundo de-Almeida-e-Pais and
     Jorge Raposo and Torres Farinha},
  booktitle={Proceedings of PAMDAS 2025 - International
     Conference on Physical Asset Management and Data
     Science},
  address = {Coimbra, Portugal},
  month = {July 17/18},
  year={2025}
}
```

Performance of Projects of Construction Companies in Portugal

Caropul Mendes[1,2*], Hugo Raposo[2,3], João Pedro Couto[4], J. Edmundo de-Almeida-e-Pais[2,3], Jorge Raposo[2,3,5], Torres Farinha[2,3]

[1] Escola Superior de Tecnologia e Gestão de Oliveira do Hospital - ESTGOH –; GRUDIS – Portuguese Network for Research in Accounting.

[2] Polytechnic University of Coimbra, Instituto Superior de Engenharia de Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.

[3] RCM[2+] Research Centre in Asset Management and System Engineering, Polytechnic University of Coimbra, Coimbra, 3030-199 Coimbra, Portugal

[4] EENG – Escola de Engenharia - Universidade do Minho; CTAC – Centre for Territory, Environment and Construction, Minho, Portugal.

[5] Ci2 –Smart Cities Research Center, Polytechnic University of Tomar, Quinta do Contador | Estrada da Serra, 2300-313, Tomar, Portugal.

*Corresponding author

* caropul.mendes@estgoh.ipc.pt; hugo.raposo@isec.pt; jpc@civil.uminho.pt; edmundo.pais@isec.pt; jorge.raposo@isec.pt; tfarinha@isec.pt

Abstract

In this article, we sought to evaluate the different studies related to the budget and its techniques, cost overruns, budget overruns and failure to meet deadlines. In this way, the aim was to obtain a better explanation of the factors that determine the causes of unfavourable budget deviations in the projects of companies in the Portuguese construction sector and their impact on performance. The work developed adopts quantitative research methods and techniques. Based on the sample obtained from 218 companies, corresponding to a response rate of 14.55%, the results obtained from structural equation modelling point to a high correlation between the six first-order constructs (internal environment, external environment, time/delay, process/monitoring, project complexity, project financing) that legitimize the perception of the main factors that determine the causes of unfavourable budget deviations in the projects of Portuguese companies in the construction sector. On the other hand, it appears that the factors, internal environment, time/delay, process/monitoring and complexity of the project are not related to the impact on the performance of Portuguese companies in the construction sector. In this article, it was concluded that the external environment and financing were statistically corroborated as causes of unfavourable budget deviations that negatively influence the impact on the performance of projects of Portuguese companies in the construction sector. However, it is worth highlighting that the objectives set for this study were achieved by making complementary knowledge available in this area of research, not only in scientific terms, but also in practical terms.

Keywords: Budget, Impact on Performance, Unfavourable Budget Deviations, Structural Equation Model

1 Introduction

The dynamics of the global market have led construction companies to increasingly focus on new strategic directions and management approaches to improve their competitiveness and sustainability. Portuguese

construction companies are following this trend and are playing a significant role in the national economy, both in terms of employability and in terms of recognition for their various activities and business areas, such as construction, civil engineering, and specialized activities.

When comparing the Portuguese construction sector with its European counterparts, there is a lack of competitiveness due to the recognition of signs such as failure to meet deadlines, performance, labor shortages, and a clear lack of workplace safety.

The literature review involves compiling knowledge developed by different authors, identifying the subjects investigated and their results, and is also a crucial way of building a foundation for the dissemination of knowledge.

From this perspective, the aim is initially to contextualize the construct of the determinants of the unfavourable budget deviation, to establish the foundations of the theoretical relationships between the study variables and the analysis, to be carried out later.

The reason for understanding the determinants that lead to unfavourable budget deviations is due to the fact that the forecasts contained in the budget differ from the results achieved *a posteriori*. It is not enough to simply calculate unfavourable budget deviations, but, above all, to understand the causes and assumptions behind the differences found.

Understanding the determinants that lead to unfavourable budget deviations in construction companies' projects is due to the fact that their complexity is motivated by structural and circumstantial reasons of their projects.

To achieve this goal, and without losing sight of the focus of the work, it will be necessary to consider a series of influences and contexts, namely, cultural and political issues that could help to understand the determining factors of the main causes of unfavourable budget deviations. The perception of the causes of unfavourable budget deviations will be analysed through the relationship of the study variables, namely in the projects of Portuguese construction companies.

The next section will present a literature review, with the purpose of presenting and analysing the main contributions of the studies developed on the theories and problems that contextualize the theme under study.

1.1 Article Structure

This article is structured into the following sections:

- Section 1 presents the introduction, the objectives of the study and a brief characterization of the Sector Construction Portuguese, finally the structure of the article;
- Section 2 presents the state of the art of the topic under study, a description of the historical context on the budget and also some studies that describe the determinants that lead to unfavourable budget deviations in the projects of companies in the construction sector;
- Section 3 describes the methodology adopted for this study and also presents a new conceptual model developed according to the analysis of the relationship between the main constructs of the factors that determine the causes of unfavourable budget deviations and their impact on the organization's performance;
- In Section 4 can be observed the results obtained from the model and a brief discussion of them;
- In Section 5, the conclusions are presented, as well as the main contributions made in this study.

2 State of the art

At this point, we opted for a historical framework on the budget and to present some studies that describe the determinants leading to unfavourable budget deviations in the projects of companies in the construction sector.

Since ancient times, the budget has been known as the traditional or classic budget, characterized by being

a future-oriented document that only presented the forecast of revenue and authorization of expenses for a certain period of time. The budget focused more closely on expenditure, without showing any concern for the real needs of management (Wildavsky, 1986).

In this context, and according to Akintoye (2008), it can be said that the budget began with Joseph of Egypt when he made a seven-year forecast for the sale of cereals during the famine period when "nothing was taken out of the treasury without a written order", in order to facilitate the control of consumption and the achievement of the desired objective.

Initially, the budget was prepared in public entities for the purposes of forecasting and presenting reliable information, legitimizing accountability and facilitating performance evaluation by comparing the predicted result with the actual one (Khan and Hildreth, 2002).

However, over several centuries, and until the Industrial Revolution, the function and focus of the budget underwent notable changes. In addition to its use in public entities, as seen in the time of Joseph of Egypt during the Pharaoh's empire, it began to be used in business organizations.

These brief historical meanings also allow us to invoke the approach to budget expansion in companies and their respective projects. For Akintoye (2008), the budget in business organizations is formally associated with the advent of industrial capitalism in the 18th century, when industries began to concern themselves with the production of goods on a large scale.

The use of budgeting at the level of business projects gained relevance when it began to be used by large companies, also, through the publication of works such as Budgeting: profit planning and control (Welsch, 1957) and Management Accounting Principles (Antony, 1970), these works being considered pioneering references in studies in the area of business budgets.

The topic of budgeting, namely the evolution of budgetary theory, budgetary techniques, failure to meet deadlines, cost overruns, and budgetary overruns have been studied by several authors such as Welsch (1957), Kaming et al. (1997), Hansen et al. (2003), Calisir and Gumussoy (2005), Blumentritt (2006), Couto (2007), King et al., (2010), Nasser et al. (2011), Jennings (2012), Raghunandan et al. (2012), Zidane and Andersen (2018) and Zhao et al. (2019).

According to Hansen et al. (2003) the budget also comprises one of the most investigated themes in the area of Management Accounting, so that organizations can use this tool for the most diverse purposes, including planning, coordination of activities, control, comparison of results obtained with those predicted and performance evaluation.

The assessment of budgetary performance is based on the achievement of companies' financial goals. However, over time, the evaluation of budgetary performance sought to ensure that companies met established objectives efficiently and effectively, focusing on innovation (Johnson and Kaplan, 1987).

However, budgetary performance for Beyond Budgeting researchers is considered inefficient in an unstable environment, so they propose replacing budgetary performance based on the performance of retrospective contracts (Hansen et al., 2003).

In this context, the authors emphasize that although there may be problems with budgetary performance, it continues to be used for control purposes. According to Libby and Lindsay (2010), companies are in a process of increasing adaptation regarding the use of the budget. Hansen et al. (2003) obtained two different conclusions regarding their studies on the development of budgetary practice, although the conclusions presented concern the same institution, one being based in the United States of America (USA) and the other in Europe.

For US-based organizations, improving the budgeting system was designed to unite operational models with the detailed financial model. However, the same organization in Europe presents a more radical view, recommending a two-phase approach: the first phase advocates eliminating the budgeting process, and the second phase prioritizes decentralization and empowerment of managers and employees. Both conclusions, although different, share the same point of view, revealing that traditional budgeting is incompatible with the rapidly changing environment.

However, it can be seen from the previous paragraphs that there is a considerable level of concern regarding budgetary practice, in which, on the one hand, abandonment is recommended and, on the other, adjustments and improvements are recommended.

This reveals the enormous importance of the budget in organizations, helping those responsible to monitor its management, and on the other hand, it limits them in creating a single model to address existing gaps in its application.

However, recognizing the main causes that determine unfavorable budget deviations in construction companies' projects makes it possible, in some way, to systematize the knowledge necessary to build a guiding principle that helps to overcome flaws in the budgetary execution of these projects.

On the other hand, the budget is used in companies as a way to verify changes in expected results and, therefore, define administrative tools that can allow corrections and adjustments in the current and future activities of companies.

To understand or identify the factors that determine the causes of unfavorable budget deviations in construction projects, it is interesting to address studies related to cost overruns, budget overruns, and failure to meet deadlines.

Odeck (2004) analyzed the magnitude of cost overruns with the purpose of verifying the relationship between actual costs and estimated costs and revealing their causes in the Norwegian highway construction sector, using data related to road construction in the period between 1992 and 1995.

The data revealed that projects built by the Norwegian Public Road Administration (NPRA) showed a significant difference between estimated costs and actual costs. Through the statistical analysis carried out, it was found that the significance of the excess cost was predominant compared to the cost reduction.

The average cost presented for projects completed between 1992 and 1995 is 7.88%, varying between 183% and 259%. Through these results, it becomes possible to identify the magnitude of the exposed costs and the impacts that the different variables pointed to in the implication of the excess cost. Although not relevant in terms of percentage, the excess in real terms for the project portfolio, considered for this study, reaches approximately 55 million euros. This is a large amount and enough to cause concern.

However, the positive side resulting from this study reveals the existence of a significant number of projects that were completed with actual costs lower than estimated.

According to Odeck (2004), the results confirm previously prepared studies, such as the case of Flyvbjerg et al. (2002) and Skamris and Flyvbjerg (1996), highlighting the difference not only between estimated costs and those obtained through road design, and revealing how political implications include factors that contributed to explaining these differences.

Calisir and Gumussoy (2005) sought to understand which factors could influence budget overruns in information technology projects, having analysed the frequency of problems encountered, the impact on project performance, the personal characteristics of the project manager, the conflict between managers and the project management style. The variables presented in this research explain approximately 73% of budget overruns in information technology projects. In this context, they discovered that delays in the execution of information technology projects are the main determinant of budget overruns found in the study, which generally leads to an increase in the amount forecast in the budget. In this context, the same study confirmed the existence of a relationship between the manager's experience and budget overruns.

Although the aforementioned study refers to the technology sector, it could be illustrated with another example/study: For example, the study by Kaming et al. (1997), who analysed the relationship between factors leading to time and cost overruns in infrastructure construction companies in Indonesia. The results obtained through factor analysis indicated that the factors that influenced the causes of delay in the expected deadline for those types of projects in Indonesia include project changes, low productivity levels, inadequate planning and scarcity of resources. In the case of cost overruns, the highlights are the increase in material costs, the degree of complexity of the project and inaccurate budget estimates.

Generally, all delays lead to unfavourable budget deviations in construction companies, making it crucial to try to assess the causes that led to these unfavourable budget deviations in construction projects. As such, it is essential to mention the studies by Flyvbjerg et al. (2004) whose objective was to identify the factors that lead to excess costs in 258 infrastructure projects (railways, bridges, tunnels and highways) in 20 different countries.

The authors revealed that the main causes of cost overruns include failure to meet the deadline for project implementation, its size and the type of construction project.

These authors went even further by stating that cost overruns are highly dependent on meeting the project implementation deadline, presenting a high level of statistical significance ($p < 0.001$). However, they indicate that decision-makers should be concerned about meeting the project implementation deadline and planning large infrastructure projects. Delays can become quite expensive, causing unfavorable budget deviations.

Creedy et al. (2010) studied the factors that lead to substantial cost overruns in road projects. The study produced interesting results by revealing that during the development phase of these projects, changes in project design and changes in the field of action produce factors that affect cost overruns in road projects.

However, for Okpala and Aniekwu (1988) the determining causes for cost overruns in the construction sector in Nigeria include material shortages, financing methods, contract deficiencies and fluctuations in material prices.

According to Arditi et al. (1985) the effect of cost overruns is not limited to the construction industry, but to the global economy of a country. The same authors point out that in Turkey, between 1970 and 1980, inflationary pressures, increases in material prices, material shortages, construction delays and errors in estimates were the main causes of cost overruns in school and hospital construction projects, resulting in costs exceeding the forecast by around 386% and 258%.

As can be seen, the relevance and interest of the topic has justified the development of several studies. The methodology adopted for this study will be presented below.

3 Methodology and Case Study

For the work developed, quantitative research techniques and methods are adopted that use the positivist argument for the development of knowledge characterized in different perspectives, notably in the ontological, epistemological and axiological perspective (Creswell et al., 2007; Saunders et al., 2009).

According to the ontological perspective, it is the reality of nature or being, considering the researcher's point of view; the epistemological perspective focuses on providing credible data through observable phenomena, taking into account the researcher's point of view; the axiological perspective, also from the researcher's point of view, addresses the issue of the researcher's independence in relation to the data, with the possibility of focusing on the research objective. However, it should be noted that the use of the quantitative research method allows the application of specific variables, the formulation of hypotheses, the use of measurement, and the evaluation of statistical data (Creswell et al., 2007), which makes it possible to investigate the factors that determine unfavourable budget deviations in the projects of Portuguese companies in the construction sector, as well as to assess their impact on the organization's performance.

Throughout the literature review, we were concerned with determining the effects of the causes that determine unfavourable budget deviations, through the various variables that allow for the evaluation of different approaches regarding their impact on the performance of the construction companies' projects, as well as the existence of models and theories, considering the inference of other variables.

From a methodological point of view, the exploratory and confirmatory factor analysis technique was used, seeking to identify the factors that determine the causes of unfavourable budget deviations in companies in the Portuguese construction sector and confirm the viability of the model proposed in Figure 4, page.

By using exploratory factor analysis, data reduction and increased agreement of measurements will be achieved by extracting the main components and factors with eigenvalues (own values) greater than 1 that were removed. According to the Kaiser criterion, the aim is to extract the factors using the varimax rotation method to facilitate the explanation of the components through the total variability of the variables (Willams et al., 2010).

Finally, primary sources were used through research. The following presents the conceptual model of the study.

3.1 Conceptual Model Proposed for the Study

This study aims to present a model that seeks to fill the gaps in the literature and discuss the hypotheses formulated in the work.

For Hoyle (1995), the conceptual research model aims to expose and clarify the relationships between the constructs that will be operationalized in empirical research. The author highlights the importance of using the figure to facilitate understanding of the interrelationship between the constructs, using specific terminology according to the objective of the study.

In this context, it is worth highlighting that the present study aims to investigate the model from the point of view of structural equation modelling that will be carried out later in AMOS (Analysis of Moment Structures).

The proposed model is based on testing and evaluating its constructs as first-order factors. Given the differences in the use of indicators presented by different authors (Sharp, 2003; Couto, 2007; Odeck, 2004; Remington et al., 2009; Shehu et al., 2014), comparing the results achieved in different studies becomes complex, and none of these studies used confirmatory factor analysis.

However, it appears in the literature that the studies found throughout the work do not investigate the determinants that cause unfavourable budget deviations in a large-scale context, although they present numerous empirical approaches relevant to the present investigation. However, it becomes dynamic to seek out the differences and compare the theoretical and practical implications of these studies, in particular, cost overruns, non-compliance with deadlines, budget overruns, the determinants that cause unfavorable budget deviations, among others.

The research seeks to analyse the interrelationship between two main concepts, that is, the determinants that lead to unfavourable budget deviations and their impact on the organization's performance.

To achieve the objective of this study, an extensive literature review was carried out to enable understanding of the conceptual model, developed according to the analysis of the relationship between the main constructs of the factors that determine the causes of unfavourable budget deviations and their impact on the organization's performance.

The structural relationship of the theories between the constructs is represented in Figure 1. In the conceptual model, the impact of performance on the organization is considered as an endogenous variable together with the project time/delay factor, project process/monitoring factor, project complexity factor and the project financing factor. On the other hand, the internal and external environment factors of the project organization are considered as exogenous variables. The measurement of the first-order constructs, presented here, will be adapted according to the instrument developed by several authors after the operationalization of the variables. Figure 1 presents the theoretical model and hypotheses proposed in this study.

The conceptual model proposed here underwent adjustments and changes aiming at the most appropriate direction according to the objectives of the present research. After presenting the conceptual model that summarizes the entire analysis that is intended to be carried out, the next section will present all the study hypotheses.

4 Results and Discussion

This section seeks to present the results of the research, supporting the methodological procedures presented in the previous point.

It is worth noting that a sample of 218 cases was considered for the analysis, corresponding to a response rate of 14.55%. This level of response cannot be considered high, but it falls within the required dimension since the response rate is higher than 10% (Meng, 2006 p. 63, our translation).

4.1 Structural Model of the Case Study

The structural equation model seeks to present the relationships between constructs, intending to assess to what extent these relationships are valid, as well as it being essential that the measurement of each construct is psychometrically sound (Byrne, 2010). In this study, the structural relationships of the theory between the constructs were considered using the maximum likelihood estimation method with the following results of the model fit index: $x^{\wedge 2} = 683.315$; gl = 309; p-value = 0.000; $x^{\wedge 2}/gl = 2.221$; RMSA = 0.075; PCLOSE = 0.000; IFI = 0.916; TLI = 0.904; CFI = 0.915; PCFI = 0.806; MECVI = 3.879.

It can be seen that the adjustment indices shown are almost identical to those presented by the adjusted measurement model, enabling the acceptance of the proposed structural model. However, the results of the structural model hypotheses are shown in Table 1.

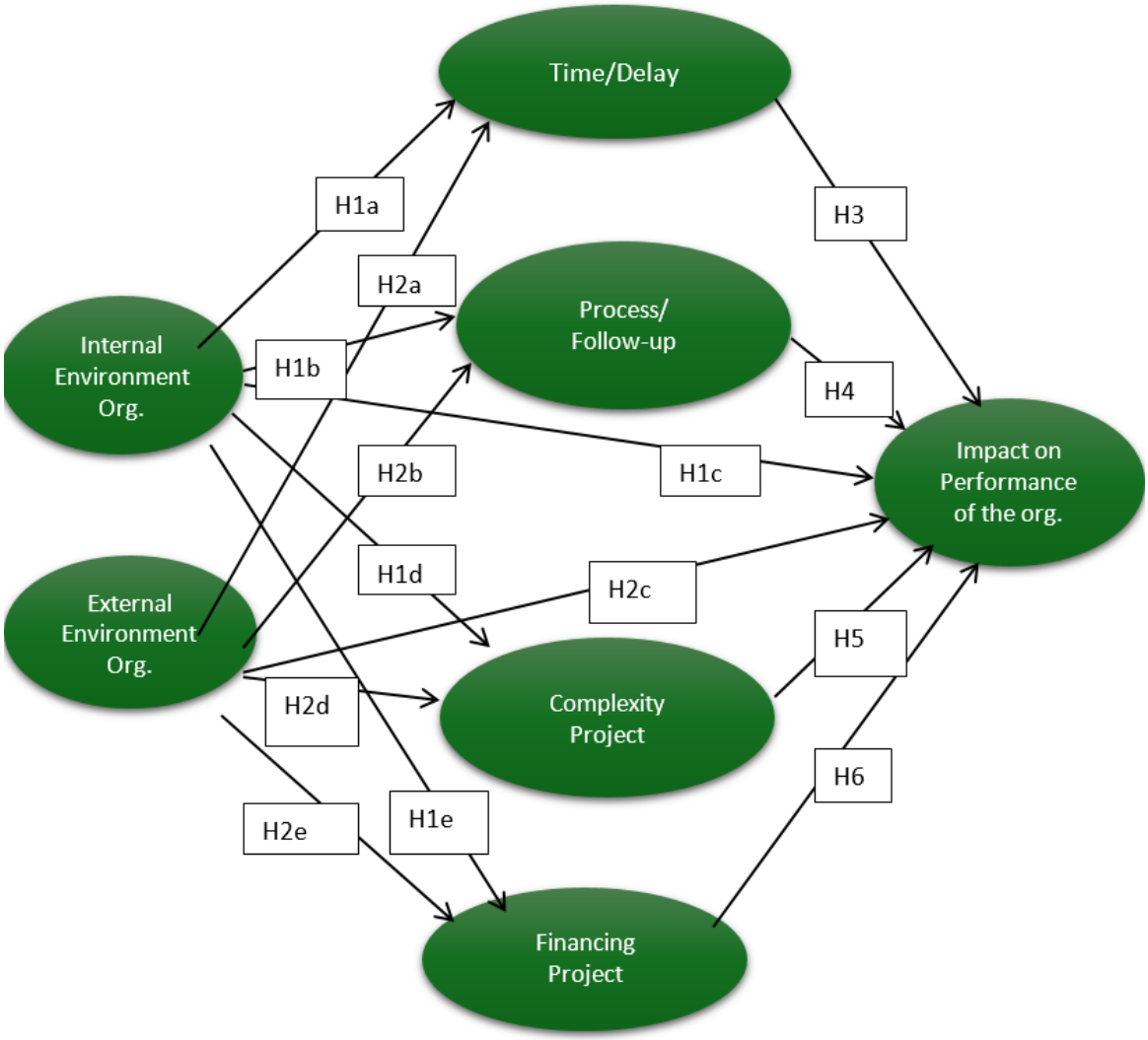


Figure 1: Conceptual Model

According to the results presented in Table 1, it appears that of the 14 hypotheses tested, only 10 are empirically corroborated by this article. However, the information found in the study reveals that the six constructs are the main causes that determine unfavourable budget deviations in Portuguese construction companies and two of the six constructs negatively influence the impact on the performance of these companies.

The following table presents a summary of the corroborated and uncorroborated hypotheses, in order to facilitate a better understanding of the study results.

Table 1: Result of the structural model

N.	Hypotheses	Structural relationship	*Factor Loading*	P	R^2
1	H1a	Amb_int → Temp_atr	0,541	***	0,293
2	H1b	Amb_int → Proc_aco	0,702	***	0,493
3	H1c	Amb_int → Imp_perf	0,049	n.s	0,002
4	H1d	Amb_int → Comp_pro	0,417	***	0,174
5	H1e	Amb_int → Finan_LCP	0,421	***	0,177
6	H2a	Amb_ext → Temp_atr	0,254	***	0,065
7	H2b	Amb_ext → Proc_aco	0,188	**	0,035
8	H2c	Amb_ext → Imp_perf	0,316	***	0,100
9	H2d	Amb_ext → Comp_pro	0,429	***	0,184
10	H2e	Amb_ext → Finan_LCP	0,346	***	0,120
11	H3	Temp_atr → Imp_perf	0,076	n.s	0,006
12	H4	Proc_aco → Imp_perf	0,024	n.s	0,001
13	H5	Comp_pro → Imp_perf	-0,068	n.s	0,005
14	H6	Finan_LCP → Imp_perf	0,353	***	0,125

Obs: *** p<0.001; ** p<0.01; * p<0.05; n.s. – não sig

Table 2: Summary of confirmed hypotheses

Hypotheses		Results
H1a	The causes of unfavourable budget deviations due to the internal project environment positively affect the time/delay factor;	Corroborated
H1b	The causes of unfavourable budget deviations due to the project's internal environment positively affect the project process/monitoring factor;	Corroborated

H1c	The causes of unfavourable budget deviations due to the project's internal environment negatively influence the impact on the organization's performance;	Not corroborated
H1d	The causes of unfavourable budget deviations due to the internal environment of the project positively affect the complexity factor of the project;	Corroborated
H1e	The causes of unfavourable budget deviations due to the internal environment of the project positively affect the financing factor of the project;	Corroborated
H2a	The causes of unfavourable budget deviations due to the external project environment positively affect the time/delay factor;	Corroborated
H2b	The causes of unfavourable budget deviations due to the project's external environment positively affect the project process/monitoring factor;	Corroborated
H2c	The causes of unfavourable budget deviations due to the project's external environment negatively influence the impact on the organization's performance;	Corroborated
H2d	The causes of unfavourable budget deviations due to the external environment of the project positively affect the complexity factor of the project;	Corroborated
H2e	The causes of unfavourable budget deviations due to the external environment of the project positively affect the financing factor of the project;	Corroborated
H3	The causes of unfavourable budget deviations due to project time/delay negatively influence the impact on the organization's performance;	Not corroborated
H4	The causes of unfavourable budget deviations due to the project's budget execution process/monitoring negatively affect the impact on the organization's performance;	Not corroborated
H5	The causes of unfavourable budget deviations due to project complexity negatively influence the impact on the organization's performance;	Not corroborated
H6	The causes of unfavourable budget deviations due to project financing negatively influence the impact on the organization's performance.	Corroborated

It is worth noting that, although the hypotheses formulated in the study refer only to direct effects between the latent variables, it is important to verify the existence of indirect effects related to the hypotheses not empirically supported (H3, H4 and H5), with the exception of hypothesis H1c, which is direct and not corroborated.

5 Conclusion

Understanding the reasons why the value predicted in the budgets deviates unfavourably from the actual value was the starting point of this study, which was guided by seven research pillars: internal and external environment factors, time/delay, process, complexity and financing, how they interrelate and what their impact will be on the performance of construction companies' projects.

In this context, two research questions presented in the introductory part of the thesis were considered, as well as the main objective of the study, enabling the construction of a guideline to identify the causes of unfavourable budget deviations, in order to allow the budget to be more credible.

However, to achieve the intended objective, as well as to obtain answers to the questions raised, quantitative research techniques and methods that use the positivist argument were adopted. From a methodological point of view, the exploratory and confirmatory factor analysis technique was used, taking into account the identification of the factors that determine the causes of unfavourable budget deviations and their impact on the performance of projects of Portuguese companies in the construction sector, subsequently supporting the viability of the proposed model.

Throughout the literature, there are several studies on non-compliance with deadlines, lack of financial control, and budget overruns in construction projects (Odeck, 2004; Couto, 2007; Remington et al., 2009; Shehu et al., 2014; Sridarran et al., 2017). Comparing the results achieved in the different studies presented in this thesis is not easy, and none of them uses confirmatory factor analysis except Zhao et al. (2019) as effectively used in the present study.

However, the assessment of the factors that determine the causes of unfavourable budget deviations is understood as a fundamental and essential instrument for measuring the relationships between the causes of unfavourable budget deviations and the performance of the organization's projects, particularly in companies in the Portuguese construction sector (Goh, 2012). In order to provide a response to the intended objective, and taking into account the literature review carried out, it can be concluded that the relationship between these constructs has not been explored in a comprehensive and integrated manner. In this context, a structural model was developed in order to test the relationships between the constructs, verifying to what extent these relationships are valid.

To conclude this article, project managers are provided with a study that can serve as a basis for their reflections, decisions, and actions, allowing them to work on preventing unfavourable budget deviations, or correcting and eliminating them if they occur. However, the theoretical and practical contributions of the study, as well as its contribution to management, study limitations, and future research, are as follows.

- **Theoretical and practical contributions**

This study contributes to the creation of mechanisms necessary to systematize knowledge about the factors that determine the causes of unfavourable budget deviations in construction companies' projects, allowing professionals and researchers to adopt ways to prevent these unfavourable budget deviations through a set of measures that can be used to prepare their forecasts. It is also assumed that this study will contribute to increasing the reliability of predicted results and, therefore, improving decision-making regarding budgeting processes.

References

Akintoye, I. R. (2008). Budget and Budgetary Control for Improved Performance: A Consideration for Selected Food and Beverages Companies in Nigeria'. *European Journal of Economics, Finance and Administrative Sciences* (12), 7-17.

Anthony, R. N. (1970). *Management accounting principles*: RD Irwin.

Anthony, R. N., Govindarajan, V., & Dearden, J. (2007). *Management control systems* (Vol. 12): McGraw-Hill New York, NY.

Arditi, D., Akan, G. T., & Gurdamar, S. (1985). Cost overruns in public projects. *International Journal of Project Management, 3*(4), 218-224.

Blumentritt, T. (2006). Integrating strategic management and budgeting. *Journal of Business strategy, 27*(6), 73-79.

Calisir, F., & Gumussoy, C. (2005). Determinants of budget overruns on IT projects. *Technovation, 25*(6), 631-636.

Couto, J. P. (2007). Incumprimento dos prazos na construção. [Dissertação de Doutoramento]. https://repositorium.sdum.uminho.pt/handle/1822/7292, accessed 2025.07.15

Creedy, G. D., Skitmore, M., & Wong, J. K. (2010). Evaluation of risk factors leading to cost overrun in delivery of highway construction projects. *Journal of construction engineering and management*.

Creswell, J. W., Hanson, W. E., Clark Plano, V. L., & Morales, A. (2007). Qualitative research designs: Selection and implementation. *The counseling psychologist, 35*(2), 236-264.

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American planning association, 68*(3), 279-295.

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2004). What causes cost overrun in transport infrastructure projects? *Transport reviews, 24*(1), 3-18.

Goh, S. C. (2012). Making Performance Measurement Systems More Effective in Public Sector Organizations. *Measuring Business Excellence, 16*(1), 3-3.

Hansen, D. R., & Mowen, M. M. (2007). *Managerial Accounting, Mason*: Thomson South-Western (8 edition).

Hansen, S. C., Otley, D. T., & Van der Stede, W. A. (2003). Practice developments in budgeting: an overview and research perspective. *Journal of management accounting research, 15*(1), 95-116.

Hansen, S. C., & Van der Stedeb, W. A. (2004). Multiple facets of budgeting: an exploratory analysis. *Management Accounting Research, 15*(4), 415-439.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*: Sage.

Jennings, W. (2012). Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic Games. *Construction Management and Economics, 30*(6), 455-462.

Johnson, H. T., & Kaplan, R. S. (1987). *Relevance lost: the rise and fall of management accounting*. Boston: Harvard Business School Press.

Kaming, P. F., Olomolaiye, P. O., Holt, G. D., & Harris, F. C. (1997). Factors influencing construction time and cost overruns on high-rise projects in Indonesia. *Construction Management & Economics, 15*(1), 83-94.

Khan, A., & Hildreth, W. B. (2002). *Budget Theory in the public sector*: Greenwood Publishing Group.

King, R., Clarkson, P. M., & Wallace, S. (2010). Budgeting practices and performance in small healthcare businesses. *Management Accounting Research, 21*(1), 40-55.

Libby, T., & Lindsay, R. M. (2010). Beyond budgeting or budgeting reconsidered? A survey of North-American budgeting pratice. Management Accounting Research, 21(1), 56-75.

Meng, F. (2006). *An examination of destination competitiveness from the tourists' perspective: the relationship between quality of tourism experience and perceived destination competitiveness.* Virginia Tech.

Nasser, M., Mah'd, O., Nimer, K., & Al-okdeh, S. (2011). The Impact of Managers' Related Variables and Department Features on Budget Characteristics: The Case of Private Jordanian Universities. *International Business Research, 4*(4), p199.

Odeck, J. (2004). Cost overruns in road construction-what are their sizes and determinants? *Transport Policy, 11*(1), 43-53.

Okpala, D. C., & Aniekwu, A. N. (1988). Causes of high costs of construction in Nigeria. *Journal of construction engineering and management, 114*(2), 233-244.

Raghunandan, M., Raghunandan, N., & Mohammed, K. R. (2012). Examining the Behavioural Aspects of Budgeting with particular emphasis on Public Sector/Service Budgets *International Journal of Business and Social Science, 3*(14), 110-117.

Remington, K., Zolin, R., & Turner, R. (2009). *A model of project complexity: distinguishing dimensions of complexity from severity.* Paper presented at the Proceedings of the 9th International Research Network of Project Management Conference.

Sharp, R. (2003). Budgeting for equity: Gender budget initiatives within a framework of performance oriented budgeting.

Shehu, Z., Endut, I. R., & Akintoye, A. (2014). Factors contributing to project time and hence cost overrun in the Malaysian construction industry. *Journal of Financial Management of Property and Construction, 19*(1), 55-75.

Skamris, M. K., & Flyvbjerg, B. (1996). Accuracy of traffic forecasts and cost estimates on large transportation projects. *Transportation Research Record: Journal of the Transportation Research Board, 1518*(1), 65-69.

Sridarran, P., Keraminiyage, K., & Herszon, L. (2017). Improving the cost estimates of complex projects in the project-based industries. *Built Environment Project and Asset Management, 7*(2), 173-184.

Welsch, G., Hilton, R., & Gordon, P. (1988). Budgeting, Profit-Planning and Control: Englewood Cliffs, NJ: Prentice-1-lall.

Welsch, G. A. (1976). Budgeting: profit planning and control-5/E: Prentice-Hall

Wildavsky, A. B. (1986). *Budgeting: a comparative theory of the budgeting process*: Transaction Publishers.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3).

Zhao, L., Wang, B., Mbachu, J., & Liu, Z. (2019). New Zealand Building Project Cost and Its Influential Factors: A Structural Equation Modelling Approach. *Advances in Civil Engineering, 2019*.

Zidane, Y. J.-T., & Andersen, B. (2018). The top 10 universal delay factors in construction projects. *International Journal of Managing Projects in Business, 11*(3), 650-672.

Proceedings of PAMDAS 2025
International conference on Physical Asset Management and Data Science

This book is available at https://pamdas.rcm2.pt

RCM^{2+} — Research Centre for Asset Management and Systems Engineering
https://www.rcm2.pt