Use of artificial intelligence for predictive maintenance of electric trains in South Africa

By

MAVHUNGU SONIA MATHALISE

Dissertation submitted in fulfilment of the requirements for the degree:

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING

In the Department of Electrical, Electronic and Computer Engineering

Faculty of Engineering, Built Environment and Information Technology

Central University of Technology, Free State

Supervisor: **Prof. ED Markus**

Co-supervisor: **Prof. M Sibiya**

Bloemfontein

2026

DECLARATION

I, Mavhungu Sonia Mathalise, hereby declare that this research project, which has been submitted to the Central University of Technology, Free State, for the degree MASTER OF ENGINEERING IN ELECTRICAL, is my own independent work, complies with the Code of Academic Integrity and Conduct, as well as other relevant policies, procedures, rules, and regulations of the Central University of Technology, Free State, and has not been submitted before by any person in fulfilment (or partial fulfilment) of the requirements for the attainment of any qualification.

Student Signature:

Date:

ACKNOWLEDGEMENTS

Firstly, I give glory and honour to GOD Almighty for His unfailing grace that never stops to reach me. I write this acknowledgement as a testament of God's faithfulness and mercy that sustained me and my family throughout this long and challenging academic journey. I wish put in record the scriptures that kept resounding in my head most of the difficult days I thought of closing off this academic journey before it was even accomplished and I would find my strength to revisit the chapter and continue; which are Proverbs 3:5–6, which says, "Trust in the Lord with all your heart and lean not on your own understanding; in all your ways submit to Him, and He will make your paths straight" and Jeremiah 29:11, "For I know the plans I have for you, declares the Lord, plans to prosper you and not to harm you, plans to give you hope and a future." These words carried me through moments of doubt and reminded me that God has been my guide and source of strength and that this work is part of His divine plan.

With humbleness and great appreciation, I extend my deepest gratitude to my supervisor Prof E.D Markus and his family, for the unwavering patience, commitment, support and encouragement even on days I least deserved it; my co-supervisor Prof M. Sibiya, for all the support and exceptional guidance throughout this research. I thank them for allowing God to use them as vessels of wisdom and direction in my academic journey.

To my husband, thank you for your support, motivation, understanding, and for constantly pushing me to keep on going, even when it felt impossible. To my parents, especially my mother, your persistence and loving reminders to "finish what you started" became the fuel that kept me focused. Your prayers became stepping stones in many times during this journey. I am also deeply grateful to my sisters, whose constant support and companionship sustained me through countless late nights of writing. A special thank you goes to Khakhu / Randy, who lovingly cared for my newborn daughter during my maternity leave as I stayed up all night and day to complete this academic work.

I also wish to express my sincere appreciation to Mr. Thokozani Yende, my mentor at PRASA, for his mentorship, exceptional guidance, and for assisting with permission grant for me to use PRASA as a case study for this research. I am also thankful to all individuals, colleagues, friends, and

professionals, who responded to my queries, offered insight, or extended a helping hand during the course of this study.

And lastly, I wish to extend gratitude to myself, for staying the course. It was not an easy journey, but I am grateful that I kept the faith. As a token of love, I dedicate this work to my beloved daughter, Arendwaho. Your birth inspired me and stirred commitment and determination. Every page written was driven by the hope that one day you will look back and be proud of my perseverance, faith, and hard work. You were one of my motivation to finish this race strong.

Above all, I am forever grateful to God, without Him, this would have been impossible.

ABSTRACT

Railway transportation plays a critical role in urban mobility and economic development, serving as a primary mode of passenger movement for millions of commuters across South Africa. Despite its strategic importance, the South African rail sector continue to face persistent operational challenges as result of ageing infrastructure, progressive rolling stock degradation, and maintenance regimes that remain largely reactive or time-based. These conventional maintenance regime approaches are not favourable when considering the demands of modern rail operations, as they fail to anticipate subsystem failures before they escalate into costly service disruptions. The inability to detect early fault conditions in critical electrical and mechanical systems has contributed to increased unscheduled downtime, high maintenance costs, and compromised passenger safety, which are challenges that have been witnessed within many railway operational environments.

To address these challenges, this study investigated the application of Artificial Intelligence and Machine Learning to develop and implement a predictive maintenance model for electric trains (ET), using a fleet of passenger rolling stock classified as electric multiple units (EMU), operated by the Passenger Rail Agency South Africa as a case study. Operational data was collected from the fleet operator, and a quantitative, simulation-based methodology was followed to evaluate a machine learning predictive maintenance model. The operational data covering a period of two (2) years (2023 – 2024) was collected across three critical subsystems, the pantograph, traction motor, and auxiliary converter unit, monitoring parameters including line voltage, current, temperature, airflow, vibration, and arc frequency. Two supervised machine learning algorithms, Logistic Regression and Random Forest, were trained and evaluated using this dataset. The Random Forest model achieved superior predictive performance, with 93% accuracy, 0.91 precision, 0.88 recall, and an F1-score of 0.89. These results verify the model's ability to effectively distinguish between normal and early fault modes with high accuracy of detection, which is crucial for safety of railway applications.

Furthermore, an analysis in this study revealed that trends in the frequency of high voltage arcs, line voltage deviation and temperature rise in traction motors were discovered to occur prior to recorded fault events. This demonstrated that features derived from these data can be investigated

efficiently as an information rich multi-parametric supply for proactive predictive diagnostics. This study has also demonstrated that AI-based Predictive Maintenance can materially improve rail operational efficiency through condition-based intervention, reducing unscheduled downtime. The research adds to the new field of intelligent railway asset management by simulating the integration of AI in a South African passenger rail environment. It provides a methodological framework to shift maintenance from a reactive or time-based approach to a proactive and data centric strategy. Future work is suggested to further develop this work by deploying different methodologies with more sophisticated deep learning architectures for more objects at the same time.

ABBREVIATIONS

ACU	Auxiliary Converter Unit
AI	Artificial Intelligence
CM	Corrective Maintenance
EMU	Electrical Multiple Unit
ET	Electric Trains
FN	False Negative
FP	False Positive
IoT	Internet of Things
LR	Logistic Regression
ML	Machine learning
MTBF	Mean Time Before Failure
MTTR	Mean Time to Repair
PdM	Predictive Maintenance
PM	Preventive Maintenance
PRASA	Passenger Rail Agency
RF	Random Forest
RTF	Run-To-Failure
SMOTE	Synthetic Minority Oversampling Technique
TBM	Time Based Maintenance

TM	Traction Motor
TN	True Negative
TP	True Positive

Table of Contents

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND

The rail network in South Africa plays an increasingly critical role to the economy of South Africa, as its very economic backbone. As the need for reliable, efficient and economical means of transportation continues to grow, the performance and sustainability of rolling stock and railway infrastructure have become more important. However, most rolling stock, particularly electric trains, rely on performance multi-modal systems such as power systems, brakes, and traction motors, that are designed as an integration of multidisciplinary element. These systems are designed to support the demanding and complex operations, ensuring system maintainability, and passenger safety and comfort. Although in most cases in a developing country like South Africa, the significance of efficient maintenance management system cannot be over emphasized, investing in preventive maintenance will enable us to introduce a cost saving model to rail systems as well as offer structured way at maintaining essential infrastructure facilities [1], [2].

Proactive measures, such as preventive maintenance in railway systems, save costs by preventing costly system breakdowns, minimising service disruptions, and extending the lifespan of critical components [3], [4]. The ability of these systems to improve efficiency in rolling stock maintenance, renewal planning, and load management makes them valuable tools for mitigating long-term operational inefficiencies [5]. Preventive maintenance enables the use of labour and spare parts in a more efficient way, it also minimizes unplanned emergency repairs or service and facilitates planning. This type of maintenance strategy involves arranging routine maintenance on equipment to reduce the chance of failures and avoid unexpected breakdowns, associated downtime, and expenses [2], [6], [7]. In recent years, inclusion of artificial intelligence (AI) in railway maintenance has represented a transformative shift towards condition-based maintenance. Machine-learning (ML) techniques enable real-time monitoring and analysis of system performance to improve maintenance intervention, thus avoiding resource wastage and enhancing system efficiency [8], [9], [10]. For South African railway operations, the statement of commitment to being strategic in using AI to ensure the industry is kept within the global competitive spectrum. The use of AI in railways can potentially bring impactful change from maintenance activities to passenger mobility and management of service operations [11], [12]. A systematic literature review has proved that the AI applications of rail sector is a promising one, especially for predictive maintenance and inspection to improve efficiency and safety aspects as well as support to reduce operational costs [13].

It is critical to note that one of the challenges to implementing AI driven technologies in South Africa's railways is the readiness of current systems to support advanced technologies. The existing railway infrastructure, as well as the rolling stock, could lack the necessary digital architecture, that allows the provision of tools and applications as may be required as support baseline for AI applied solutions [13], [14]. It is crucial as AI can predict breakdowns and therefore, minimize operational downtime to make the rail services more efficient. And digitalisation is an essential part in order to transform systems into flexible, sustainable and efficient railway networks. In the era of Internet of Things (IoT), AI and Digital Twins generally apply more intelligent O&M strategy in conformity with Industry 5.0 and the Circular Economy [15], [16]. The change is a part of the efforts to humanise the railways and make them future ready.

Building an integrated ecosystem comprising of operators, service providers, and regulators is a necessary step to deploy the AI infrastructure for predictive maintenance in rail industry [11]. Collaborative innovation and the use of tailored AI tools that support knowledge sharing, peer learning, and joint ventures are necessary to address interoperability issues. AI-enabled big data analytics and the IoT could improve current maintenance procedures by allowing for asset condition real-time monitoring and decision-making. AI models should be adaptive in that they must also account for the slowly changing environment such as climate and railway terrain conditions in a diverse geographical area such as South Africa [8], [11]. To ensure the reliability of algorithms and long-term robustness of infrastructure, in situ validation is required. As much as AI can automate many processes, human knowledge is still required to respectively understand the insights from AI models. The development of workforce is important to prepare AI literate technician and engineer. Change management will guarantee that AI systems are used around the clock in operations processes without substituting human judgment. This human-AI collaboration leads to improved decision making and system safety. As the South African rail industry is at a turning point in AI implementation, policy support, strategic investments and collaboration with academia and industry are required to bring widespread benefits [5], [17]. Adopting intelligent maintenance technologies will not only reduce operational costs and environmental impact but also result in green economy, sustainable railway systems.

1.2 PROBLEM STATEMENT

The railway industry, particularly in the operation of electric trains, continues to experience a high rate of rolling stock failures. Many of these failures are unforeseen, leading to significant service disruptions, including delays and cancellations. The absence of prior fault detection complicates the corrective maintenance process, as diagnosing the root cause often consumes more time than if the failure had been anticipated. In some cases, failures result in unexpected and severe equipment damage. When such events occur without the immediate availability of spare parts, repair times are further extended.

Currently, railway operators rely heavily on Time-Based Maintenance (TBM), yet the frequency of unplanned corrective actions remains disproportionately high. This reactive approach not only affects operational efficiency but also incurs substantial costs, particularly when dealing with expensive and critical train components. Despite the growing adoption of Predictive Maintenance (PdM) in other sectors, the railway industry remains largely dependent on traditional TBM and corrective strategies, with limited scrutiny of their long-term effectiveness. This highlights a pressing need for real-time monitoring systems and predictive maintenance frameworks tailored to electric trains in order to enhance reliability, reduce downtime, and optimise maintenance planning.

1.3 RESEARCH AIM AND OBJECTIVES

The main aim of this research is to formulate a predictive maintenance strategy based on machine learning of key parameters of certain critical system components that are closely and accurately monitored in electric trains. This would lead to most service failures and damage to equipment being predicted and prevented. The study is guided by these main objectives:

- To collect and analyse real operational data from a rolling stock operator for a case study, in order to scientifically evaluate current maintenance trends and practices.

- To evaluate and assess existing machine learning algorithms and determine the most suitable approach for modelling operational trends based on the collected data.
- To recommend and simulate a predictive maintenance strategy using machine learning techniques applied to key parameters of critical system components.
- To evaluate the performance and effectiveness of the recommended predictive maintenance model in enhancing the reliability and efficiency of electric train operations.

1.4 SIGNIFICANCE AND JUSTIFICATION OF THE STUDY

This study adds significant value to the ongoing investigation of means to foster modernisation and sustainability of the railway industry in South Africa, specifically electric multiple unit maintenance. The industry is currently overburdened with aged infrastructure, increased operational cost and requirement of better service reliability. The traditional time-based and reactive maintenance approach appears to be inadequate in preventing unplanned service disruptions, costly repairs and dissatisfied passengers. There is also a clear shortage of autonomous, predictive fault detection systems that would be able to retrench this turbulence. By examining the adoption of AI in predictive maintenance, this study introduces a disruptive path to address these industry gaps. AI-powered systems can facilitate real-time condition monitoring, data-driven decision-making and predictive maintenance scheduling while driving gains in equipment availability, safety, and lifecycle costs. These developments are in line with national priorities, including the reduction of carbon emissions, infrastructure resilience and encouraging smarter sustainable public transport.

In addition, the research provides implications for different stakeholders such as train operators, maintainers, policymakers and technology-based developers. In a resource-poor, knowledge-constrained setting, AI solutions that are customised to the context area crucial for closing gaps in implementation and informing policy and strategy development. Finally, this study is paramount in building up South Africa’s ability to innovate new methods within the rail sector so that it can stimulate economic and social development and become a powerhouse for advanced railway maintenance internationally in Africa.

1.5 RESEARCH METHODOLOGY

This study employed a quantitative research methodology, underpinned by a simulation-based design, to develop and rigorously validate an artificial intelligence model for predictive maintenance of electric multiple unit trains. The quantitative approach facilitated the systematic analysis of extensive operational data acquired from the X'Trapolis Mega EMU fleet, enabling the identification of intricate relationships between critical variables influencing component reliability and failure probabilities.

Data acquisition focused on three critical subsystems: the pantograph, traction motor, and auxiliary converter unit. These components are equipped with an array of sensors, as per train design, to monitor important physical variables such as current, voltage, vibration, temperature and arc frequency. The dataset collected data covers a 2-year period (2023–2024) and it is sampled using combination of sampling approaches. This approach combined periodic monitoring sampled every 15 min to follow the system continuously, and event-based sampling, which was remotely triggered by indications of abnormal operating conditions in order to capture high-resolution transients.

Machine learning models, Logistic Regression and Random Forest classifiers, were then applied for modelling the probabilities of subsystem failures. The raw data was pre-processed in a comprehensive manner, such as removing missing and outlier, normalising of numerical features in order to improve the accuracy of algorithm. After the data pre-processing, they were divided into training and validation (5-fold) sets with 80:20 ratio. The model was validated using a standard set of classification metrics namely accuracy, precision, recall, F1-score and the Area Under the Receiver Operating Characteristic curve to establish statistical validity and accurate profiling of predictive potential.

The overall research design adhered to a structured, five-stage methodology: data collection and preprocessing; feature engineering and scaling; algorithm selection and training; model validation and evaluation; and result interpretation and operational implications.

This structured methodology ensured that the developed predictive framework was both data-driven and demonstrably relevant to the complex demands of real-world railway maintenance environments.

1.6 CONTRIBUTION TO KNOWLEDGE

This study makes several key contributions to the novel area of AI-based predictive maintenance in railway industry. First, it offers a common mechanism to integrate multi-sensor data such as the electrical-, mechanical-, and environmental-related features into an overarching predictive model. This manner of practicing yields a more thorough and accurate fault detection rate than the limited, conventional single parameter monitoring approaches. Second, it introduces a hybrid data-sampling strategy integrating continuous periodic sampling and anomaly-triggered event-based sampling that has been demonstrated to lead to better detection rates as well as finer grained temporal resolution of crucial precursory features.

Further, this research delivers a practical proficiency of the Random Forest model as an effective and robust predictor for rail maintenance problems. Performance continued to be better than that of the linear model, such as Logistic Regression in particular, in cases where complex nonlinear relations among various train subsystems exist. The resulting feature-importance analysis using the Random Forest framework provided valuable practical insight into failure precursor, and identified traction motor temperature, pantograph arc frequency and line voltage variation as precursors to incipient faults.

In addition to the technical contributions, work reported in this study could also have a direct contribution to the local South African rail industry by connecting AI driven maintenance innovations with established Reliability, Availability, Maintainability and Safety framework of railway operators. This alignment will provide a scientific evidence-based rationale for shifting from the existing reactive interval-based maintenance practices to proactive (i.e., condition monitoring) maintenance thereby benefiting both operational efficiency and train technologies within the national passenger rail network.

1.7 LIMITATIONS OF THE STUDY

Although this study successfully achieved its intended objectives, there are limitations that must be recognised. First, the original dataset is a mix of real operational data and failure data being synthetically generated from statistical models used to deal with the common problem of class imbalance. While these augmentation techniques clearly aided in training and helped the model to recognise rare fault patterns, it is also acknowledged that the synthetic data may not include all of the complexities, random nature and nuances of real-world defective conditions. Second, the predictive model was constructed and tested only in a simulated setting but not tested under real operational conditions. Hence the online performance of these models under real time conditions, influenced by technical, environmental and human factors complexities still need to be tested further through practical realisation.

The other limitation is related to the parameters being monitored. Although a broad list of critical subsystem variables was considered, some environmental and infrastructure data such as temperature, humidity and catenary wear were not included. Including such external factors may improve the precision and broad view of failure prediction. Additionally, due to computational resource constraints in the study, comprehensive hyper parameter tuning on a large scale (for different values of the ensemble weights) and testing advanced model ensemble configurations could not be performed. However, despite these limitations, the results reported here present compelling evidence for future investigations. It would be beneficial for future work to extend the dataset to include live streaming databases and explore the use of more complex deep learning architectures including Long Short-Term Memory networks and Convolutional Neural Networks. This would promote the deployment of real-time anomaly detection and capturing complex temporal dependencies in continuous EMU operation cycles.

1.8 OUTLINE OF THE DISSERTATION

As chapter 1 discusses the introductory background and the problem statement of the study, it also outlines the main objectives as well the approach taken to achieve the study goals. The remainder of the dissertation is structured as follows: In Chapter two, existing literature related to predictive maintenance and machine learning applications in railway systems is reviewed to support the background of the study and discover the existing work already done by other researchers. Chapter

3 explains the research methodology that led to the development of the models, including the testing and evaluation methods of the developed models. Chapter 4 presents a thorough analysis and interpretation of the data. Chapter 5 presents and discusses the ML models simulation results. And lastly, chapter 6 gives conclusion of the study and outlines the recommendations for further studies.

CHAPTER 2: LITERATURE REVIEW

2.1 INTRODUCTION

The development of maintenance methods, particularly for transport systems, marks a significant move toward using technology to improve how things work and to ensure safety. This chapter reviews the existing literature on the maintenance background, objectives, and strategies. The evolution of maintenance from corrective to preventive to predictive approach is explored, outlining the effectiveness of each strategy and emphasising how predictive maintenance is changing the railway industry. The significance of predictive maintenance was precisely investigated by examining case studies of countries that have successfully implemented predictive maintenance for electric trains. The pivotal advancements that have set the stage for the integration of artificial intelligence (AI) into maintenance practices are presented in this chapter. By synthesising insights from global research, this literature review provides critical perspectives on the effectiveness, challenges, and future directions of AI-driven predictive maintenance strategies. This chapter also aims to contribute uniquely to the broader conversation on maintenance strategies, shedding light on how AI can improve the maintenance of electric trains in South Africa.

2.2 RAILWAY INDUSTRY MAINTENANCE BACKGROUND

In the railway industry, the evolution and modernisation of maintenance strategies are vital and reflects the industry's shift towards proactive solutions for improved operational efficiency, safety, and the reliability of assets [18]. The transition over the years in maintenance underlines a broader trend observed across various industries, where maintenance transcends its traditional role to become a strategic cornerstone in asset management [19], [20], [21]. In the field of rolling stock, the implementation of advanced maintenance strategies, particularly those that integrate sophisticated monitoring models, has evolved from being a best practice to an essential requirement [22]. This necessity arises from the complex interaction between environmental factors, mechanical components, and electronic systems, which collectively influence the safety and reliability of railway operations [23], [24]. In the lifecycle of a system or product, the design phase crucially includes considerations for its operational lifespan [13], [25]. Ideally, a system is expected to function flawlessly throughout its designed lifespan before inevitably experiencing failure as a result of wear and tear. However, practical experience diverges from this ideal scenario, illustrating maintenance as a comprehensive suite of activities aimed not just at repairing or extending the life of a system but crucially at maintaining its performance to meet its designed functions throughout its operational life [26]. This reality highlights the importance of performing suitable maintenance with effectiveness to ensure the product or system continues to perform in a reliable manner presently and that its future can be sustained. According to Ciocoiu, Siemieniuch, and Hubbard [40], maintenance has evolved into a critical function across various operations, embodying a system that encompasses various elements, including specialised methodologies, skilled personnel, and essential resources. The effectiveness of maintenance operations depends on the seamless integration of these various elements. Shortfall in any of these aspects could jeopardise the entire maintenance process.

2.3 MAINTENANCE OBJECTIVES

The pivotal role of maintenance in ensuring operational efficiency, equipment availability, and environmental concern is well investigated in the literature. This investigation emphasises that maintenance as a critical asset preservation strategy rather than a mere operational task, aligning directly with the overarching goals of quality and safety in industrial operations. Electric trains are complex, real-time, distributed, and reconfigurable systems that incorporate many embedded subsystems, which are integrated to perform a high-quality transportation service [10], [26]. Failure to effectively implement the maintenance role for these systems may lead to significant service interruptions, degradation in service quality or failure to provide enhanced level of safety and an efficient means of transport (i.e., maintainability) which emphasizes the key role that maintenance plays to secure a safe and reliable mode of transportation [27]. Effective maintenance secures the reliability and longevity of assets and fosters trust between service providers and consumers by ensuring services are consistently available with minimal disruption [28].

2.4 THE EVOLUTION OF MAINTENANCE STRATEGIES

Maintenance activities demand strategic planning and careful execution to effectively enhance quality, ensure safety, and boost productivity [29]. Over recent decades, maintenance activities have evolved from reactive approaches to more proactive strategies [30]. This shift reflects a strategic change in how organisations manage and maintain their assets, emphasising systematic thinking and planning in ensuring safety, quality, and operational excellence [30]. This transformation in maintenance practices has been the subject of extensive discussion across numerous platforms, with researchers categorising maintenance strategies based on their observations and analyses. Progression, as illustrated in Figure 2.1 [31], from corrective to preventive and, ultimately, to predictive maintenance, reflects an industry-wide recognition of the importance of minimising downtime, enhancing safety, and optimising asset management.

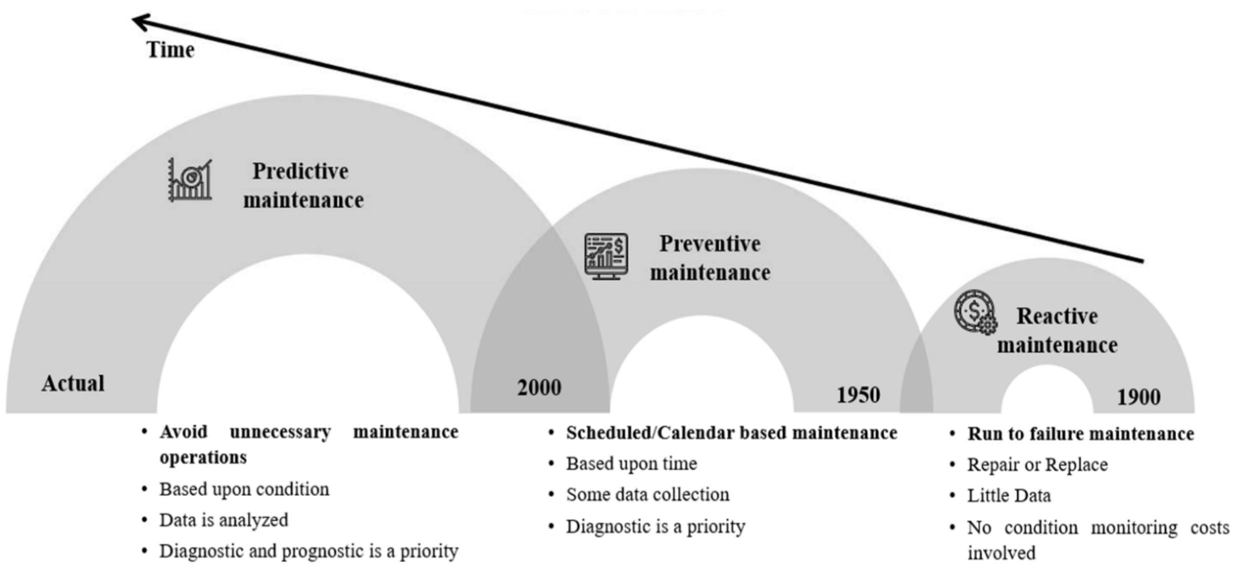

Figure 2.1: Evolution of Maintenance Strategies [31] .

Initially, industries, including railways, predominantly relied on corrective maintenance (CM) or run-to-failure (RTF), as others may refer to it, an approach where repairs are made after a failure has occurred [22], [32]. This method, while straightforward, often led to unplanned downtime and higher long-term costs due to the unexpected nature of repairs and the potential for secondary failures [30]. The limitations of CM led to the adoption of preventive maintenance strategies. This approach schedules regular maintenance activities based on time or usage metrics, irrespective of the equipment's current condition. Preventive maintenance (PM) aims to reduce the likelihood of equipment failure through fixed maintenance intervals [33]. It is guided by the principle that regular care and servicing will extend the life of assets and ensure operational reliability. However, this method can lead to over-maintenance, where resources are expended on equipment that may not require immediate attention, thereby inflating operational costs without proportionate benefits [34].

[35], [36] Predictive maintenance (PdM) leverages data analysis, condition monitoring technologies, and advanced diagnostics to predict equipment failures before they occur. The strategy depends on understanding the specific condition of each asset to schedule maintenance only when necessary, optimising resource allocation and minimising unnecessary maintenance interventions [34], [35]. The advancement of predictive maintenance has been facilitated by developments in the IoT, big data analytics, and artificial intelligence (AI), allowing for real-time monitoring and analysis of equipment performance and health [30].

2.5 EVOLUTION OF SOUTH AFRICAN RAILWAYS MAINTENANCE STRATEGIES

Although the railway industry is slow in adapting to innovations and new technologies, it is currently leaping towards the era of digitalisation. Railway technology and system monitoring techniques have advanced, and the application of these strategies has progressed [36]. However, in South African railway operations, time-based maintenance (TBM) and corrective maintenance remain primary maintenance strategies. Data recording technologies are predominantly employed post-failure to analyse performance data and investigate causes. Corrective maintenance involving repairs after breakdowns, remains a dominant practice [14], [37], [38]. Given the complexities of modern electric trains (ET), unexpected failures with unidentified causes can significantly prolong the maintenance process, requiring extensive time to pinpoint and address the root cause [17],[14], [37]. Failures often develop gradually, escalating from minor to major issues, with the potential for recurrence if the underlying causes are not accurately identified and isolated [6],[7], [37].

Notwithstanding these improvements, the implementation of more contemporary maintenance techniques in the railway sector in South Africa has been limited which results in poor performing and costly maintenance systems that are over burdensome to railways [3], [39]. There must be a shift towards integrating Predictive Maintenance (PdM) strategies to address these challenges and enhance maintenance efficiency. This shift towards PdM has mainly impacted the railway industry due to the critical importance of safety, reliability, and punctuality in this sector. Railway operators who have adopted PdM have been enabled to anticipate and address potential issues before they lead to service disruptions, enhancing passenger safety and satisfaction while optimising the lifecycle management of rolling stock and infrastructure [14].

2.6 COMPARISON OF MAINTENANCE STRATEGIES

The shape of maintenance strategies within industrial operations over the years has undergone significant evolution, with corrective, preventive, and predictive maintenance representing critical methodologies in this progression [31]. Each strategy offers distinct approaches to managing and mitigating equipment failures, with operational efficiency, costs, and downtime implications.

2.6.1 Corrective Maintenance

Corrective Maintenance (CM) represents a traditional approach to maintenance, activated by equipment failures and necessitating immediate intervention to return the system to its operational state. Characterised by its reactive nature, CM is initiated once a fault has occurred, engaging a process to identify, isolate, and correct the issue, whether through repair, replacement, or restoration of the failed component to its functional condition[40]. The primary methodology employed in CM is troubleshooting, an analytical process designed to diagnose the cause of the failure and implement the necessary corrective actions [3]. This maintenance strategy is closely tied to in-service failures, requiring emergency remedial actions and repairs. Often, equipment may suffer significant damage or reach a state of complete failure, necessitating extensive repairs or even a total rebuild. The implications of relying on CM include potential operational downtime, increased repair costs, and the risk of unexpected failures leading to safety hazards or production losses [3], [41], [42]. While CM is essential for addressing unforeseen equipment failures, its reactive nature means it does not prevent the failure from occurring in the first place but responds to it afterwards.

Corrective Maintenance is a straightforward approach to maintenance that, ideally, should be deployed in situations where the assets are not critical to the core operations, the costs associated with potential failures are minimal, and the safety risks are negligible [3]. Although CM can be applied for non-critical assets with little consequence to failure, methodically combining the predictive and preventive approach appears as a more appealing and cost-effective option [43]. While CM may find its place in managing non-critical assets with minimal failure consequences, a strategic blend that includes preventive and predictive maintenance can offer a more balanced, cost-effective approach. Such a holistic maintenance strategy not only mitigates the immediate impacts of equipment failures but also addresses the root causes of such failures, reducing their occurrence and severity and optimising the overall maintenance expenditure [44].

2.6.2 Preventive (Periodic) Maintenance

Preventive Maintenance (PM) is defined as carrying out maintenance service before the likely event of equipment failure [24], [45]. This maintenance approach seeks to decrease the likelihood of equipment failure in the future by handling potential system problems before they even occur [24]. Distinct from CM, PM is characterised by its systematic nature, involving scheduled maintenance activities, and its main objective to resolve product or system premature failures [24], [39], [46]. These activities are often determined by time-based intervals (Time-Based Maintenance or TBM) or predefined conditions based on mileage or number of operations [6], [47]. This strategic scheduling allows for a more controlled maintenance process, enhancing equipment reliability and operational efficiency by pre-emptively mitigating potential disruptions [6], [24] [38]; however, the strategy does not take into consideration the actual health status of the system and may lead to unnecessary maintenance actions [48].

In history, railway maintenance was primarily reactive [39], [49] and would be mostly based on the knowledge and experience of those already in the industry [50]. However, as maintenance and engineering knowledge advanced, railway companies started implementing regular inspections and maintenance schedules to introduce a preventive strategy [23]. This implementation led to the replacement of system components that still had sufficient useful life and maintenance actions being carried out regardless of the actual health state of the components [51].

2.6.3 Predictive Maintenance

Predictive maintenance (PdM) strategy, also known as proactive maintenance, is aimed at predicting failures before they manifest with an ultimate goal to avoid costly downtimes, expensive equipment damage, or suspended operations [30], [31]. Through monitoring vibration patterns, thermal signatures, and other relevant indicators, potential problems can be identified early, which would prompt preventive actions to mitigate equipment failures and breakdowns [25], [52]. This strategy differs significantly from conventional scheduled maintenance, in which components are routinely replaced at predetermined intervals, resulting in avoidable downtimes and elevated maintenance expenses [29], [53]. Rooted in the foundational principle of condition-based maintenance, Predictive Maintenance (PdM) leverages sophisticated technologies, including data analytics, machine learning, the Internet of Things and artificial intelligence (AI), with the aim of anticipating equipment malfunctions prior to their manifestation [54], [55]. The strategy is positioned to anticipate failures and proactively address them and could make use of AI to analyse sensor data and detect subtle operational changes, providing recommendations for optimising logistical supply chains [46], [56]. AI and data science play a pivotal role in enhancing predictive maintenance strategies, offering the possibility of additional progress in predictive analytics and fault detection [57]. PdM has transformed traditional maintenance outcomes and enabled organisations to better position technical operations in order to thrive in competitive markets, leveraging technology for sustained growth and resilience. Figure 2.2 illustrates several advantages of Predictive Maintenance (PdM) in terms of financial savings, operational effectiveness, and safety improvements within various organisations.

Figure 2.2: Organisational benefits of PdM [58], [59]

Figure 2.2 demonstrates how PdM leverages the IoT to shift to smart replacement of parts, which happens when data indicates a likelihood of failure, therefore preventing unnecessary downtime and reducing maintenance costs. This benefit then also enhances asset availability, ensuring there are no unnecessary service losses as illustrated.

AI technologies in the rail industry have matured beyond the prediction of failures to actually optimising performance throughout the whole asset life-cycle. By employing real-time data monitoring, AI systems can facilitate proactive planning and improved reliability of assets over equipment life cycles [17]. Evolution from reactive to predictive maintenance may spark competition among rolling stock maintenance while also support the broader industry efforts to the Fourth Industrial Revolution intelligent applications for sustainability and system reliability [60]. The implementation of AI across the country's rail space would also prove the nation's dedication towards closing the digital divide and pursuing technology for all. Predictive maintenance is a perfect example of how AI can drive efficiency, save energy, and bank balance booster. However, this transformation necessitates a conducive environment including strong infrastructure, public-private partnerships with clear strategy and policies for sociotechnical factors such as jobs, data governance and ethical use National case studies show that using AI leads to increased asset utilisation and greater workforce productivity [61]. To ensure that such benefits are maintained, AI systems need to be continuously evaluated and recalibrated according to new technical and organisational demands in order for the railway sector's capacity for innovation to remain both robust and trustworthy.

2.7 INTRODUCTION OF PDM AND AI IN RAILWAY INDUSTRY

Maintenance 4.0, an ideology originating from the Fourth Industrial Revolution (4IR), is reshaping the railway transportation sector through the incorporation of intelligent systems and web-enabled strategies aimed at improving maintenance protocols and overall functionalities [62], [63], [64]. This strategic approach brings to the railway sector condition monitoring technologies and leverages data analytics to predict and prevent equipment failures, thereby improving system performance and safety [62], [63], [65]. The application of PdM in railway systems was driven from a necessity to reduce unplanned system downtime, and train services disruptions as a result of asset failures (with significant financial penalties) and impacts on passengers' experience[12]. However, despite its accepted potential, full exploitation of PdM is challenged by the complexity of railway systems and the need for a systematic approach to technology selection and application [66]. As a result, the successful implementation of PdM into railway systems needs better knowledge bases of different rolling stock and infrastructure parts, and predictive models that fit to railway engineering. There are challenges to overcome, but the Rail Technical Strategy still foresees a future of railways wherein intelligent maintenance and condition monitoring transforms basic engineering functions [67]. Nevertheless, according to [30], the successful implementation of PdM has to be carefully considered regarding which technology should be used, the data quality, assessment approach and data-mining techniques employed to guarantee reliability and safety of railway operation. Though PdM has been mentioned to be valuable, the conventional predictive maintenance where systems are predicted by assessing when a certain piece of equipment is most likely to fail based on past events and manual data analysis has its own limitations and problems [2], [46]. This approach does not incorporate any smart monitoring technologies and gets limited as systems become more complex. Figure 2.3 shows some of these challenges.

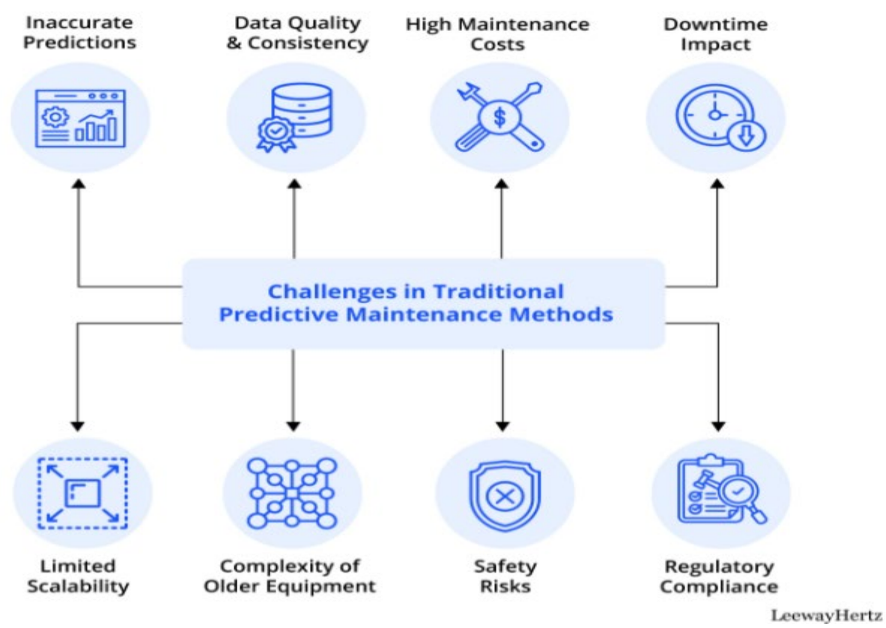

Figure 2.3: Challenges in traditional predictive maintenance methods [68].

The South African rail industry is an important sector of the national economy making it vital that effective maintenance strategies are in place to support operation reliability. PdM, leveraging AI and data analytics, is such an innovative approach in this regard. PdM can schedule the preventive maintenance in advance, before any breakdowns and downtime are experienced by systems, leading to reducing unplanned downtime and increasing energy efficiency [75]. It is interesting to note here that although the context does not refer explicitly to the South African railway sector, it raises the more generic use of PdM in different industry types. For example, some literature has highlighted the application of IoT-enabled PdM as an utmost innovative approach across several industries supporting to foresee and prevent potential machine failures [69]. Similarly, most literature discusses the integration of AI within Industry 4.0, and its influence in driving forward the development of fault diagnosis and PdM strategies in the technical sector [9]. This application of PdM in the railway industry aligns with the global trend of adopting AI-driven maintenance approaches. As outlined by [70], PdM is considered a key opportunity for accurate Structural Health Monitoring (SHM) in railway infrastructure. By associating actual infrastructure behaviour with measured data SHM can exploit a small number of sensors placed on strategic components, which may reduce intervention times and costs in the South African railway domain,

The reliability of the datasets and the level of preparedness of wireless technology that is needed for successful implementation AI solutions are challenging. Preserving privacy of data and mitigating algorithmic biases are critical issues when building responsible AI systems adapted to local conditions [13]. In order to address these challenges and maximise the potential of AI, whilst also minimising risks, it is imperative to develop flexible integration paths and prioritisation strategies for deployment of AI technologies that are tailored to regional focuses [13], [15], [16]. The integration process should involve investments in the necessary infrastructure to support emerging digital technologies and ensure the workforce is adequately trained to manage and operate AI systems effectively [9], [30]. Integration of various algorithms within artificial intelligence systems aimed at predictive maintenance managed to attain higher accuracy, increasing ease of use in anomaly detection and predicting lifespan estimates of the components. Electric trains and their myriads of interlinked subsystems necessitate complex AI trained on large amounts of heterogeneous data, accentuating the need for sophisticated models. There is ample evidence proving the use of predictive AI technologies driven by sophisticated algorithms aids in the improvement of resource availability by decreasing unscheduled outages, therefore boosting the level of service and client contentment. The nature of railways means the solutions have to be multi-layered and constantly evolving [71].

2.8 APPLICATION OF AI/ML ALGORITHMS IN ELECTRIC TRAIN MAINTENANCE

The application of Artificial Intelligence (AI) and Machine Learning (ML) algorithms to the maintenance of electric trains represents an innovative strategy that presents substantial enhancements in both maintenance efficiency and cost-effectiveness [72]. Research has shown that machine learning algorithms like RUS Boost Ensemble and deep reinforcement learning, when applied alongside digital twins, have demonstrated the capability to forecast maintenance requirements with a notable level of precision [41], [73]. This approach also shows potential in diminishing maintenance tasks by 21% and curbing the occurrence of defects by as much as 68% [74], [75]. AI and ML methodologies are progressively being incorporated into railway monitoring systems for the purpose of forecasting potential failures and enhancing maintenance scheduling [62]. Figure 2.4 illustrates how AI-integrated predictive maintenance offers several benefits to organisations across various industries.

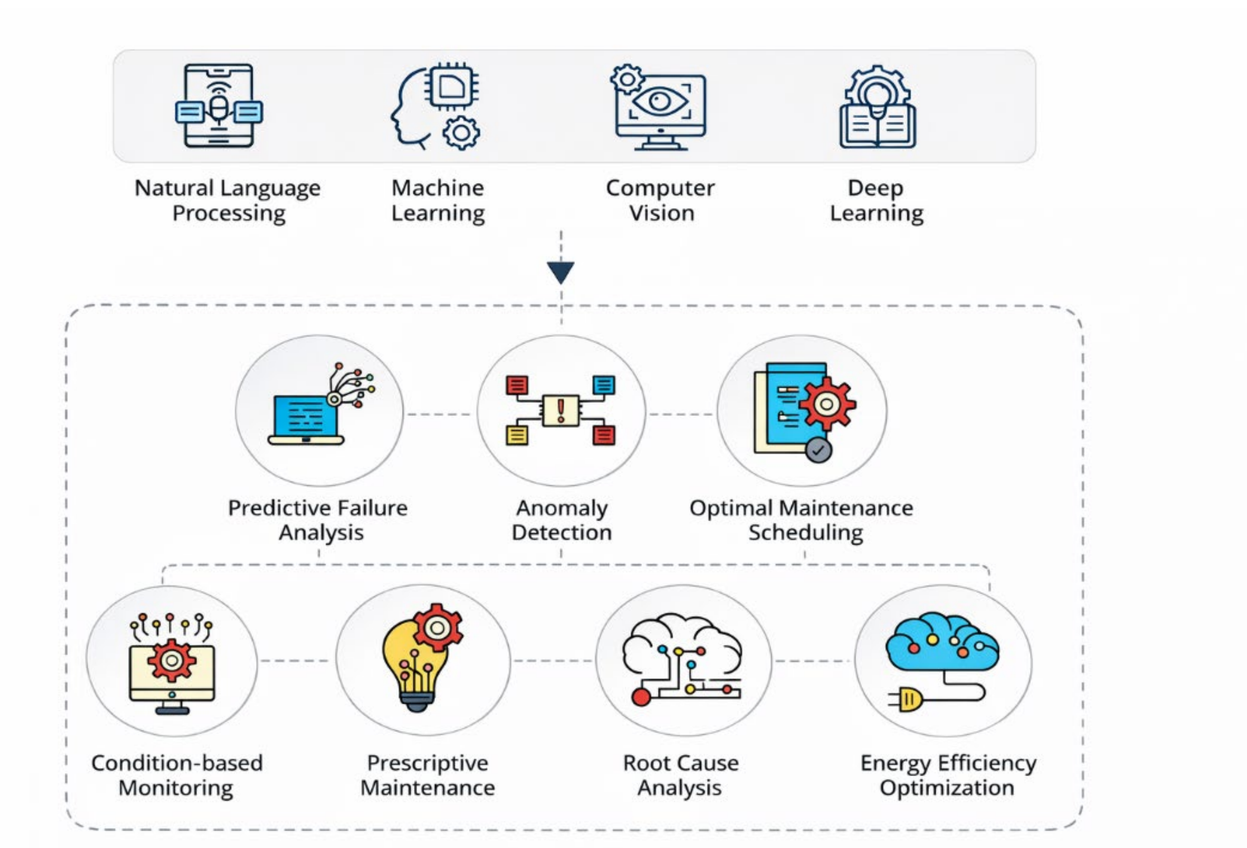

Figure 2.4: Benefits of AI-powered predictive maintenance [68]

The railway industry has seen a significant transformation within the maintenance of assets through the implementation and utilisation of AI cutting-edge technologies to improve operational effectiveness and dependability [76], [77]. This transformation gives confidence in the potential of ML, the IoT, and other AI techniques to be incorporated in the management, planning, and execution of railway system maintenance holistically. AI plays a pivotal role across a range of maintenance aspects, encompassing diagnostics, prognostics, optimisation, high-level automation, and auto-adaptive systems[69], [78]. The integration of data-driven models and decision support systems has proven to be an essential tool that breeds new developments such as Intelligent Asset Management Systems, which enhances maintenance activities and improves system reliability [79].

From an economic perspective, AI adoption in rail maintenance has demonstrated significant potential for cost optimisation. Benefits include reduced expenditure on emergency repairs, extended asset lifespans, and minimised spare parts inventory requirements. Cost-benefit analyses conducted by South African rail operators underscore the fiscal viability of AI investments, especially when supported by government funding and public-private technology partnerships [71]. Pilot implementations play a critical role in validating AI tools, facilitating user acceptance testing, and informing iterative system refinements[17]. This literature review has shown that the railway sector could formulate well-informed maintenance strategies, optimise resource distribution, and improve both safety measures and passenger satisfaction utilising AI technologies such as intelligent agents, neural networks, Bayesian frameworks, etc. The subsections below outline some of the ways the application of AI has influenced maintenance practices for electric trains.

2.8.1 Real-time Diagnostics and Predictive Analytics

AI technologies are more frequently employed to transform the traditional maintenance of electric trains with real-time fault diagnostics and predictive analytics, which enhance operation reliability efficiency and safety [41], [76]. This AI integration into railway maintenance systems, is an involved but revolutionary breakthrough in transportation technology. At the international level, AI applications in the railway sector are increasingly considered as possible promoters of changing maintenance from reactive to predictive and prescriptive ones [77]. This is primarily due to the fact that operators are able to track, identify and diagnose faults with a high level of accuracy in real time using approaches such as artificial neural networks (ANNs) and machine learning algorithms [77], [80]. AI techniques can detect abnormalities early and indications of deterioration or faults in critical components through analysis of data with parameters such as vibration, temperature, and power consumption [76]. This early detection then leads to scheduling of maintenance activities before the fault matures and causes disruptions that could greatly impact system end users [79], [81]. The emergence of AI-based, real-time diagnostic tools and predictive analytics into increasingly proactive forms of maintenance is consistent with other broader trends in many other industries (which are also leveraging AI for improvement/optimisation) designed to improve and anticipate future events. However, in the South African context, this shift requires the integration of emerging technologies within accepted operating procedures [11], [82]. Railway operators have to meet the challenge of optimising maintenance activities in light of the constraints posed by infrastructural data limitations, predictive inferential models and human competences [17].

AI technologies in predictive maintenance hold the potential to enhance learning within railway maintenance depots by capturing maintenance activities and continuously refining related processes. This improvement can significantly contribute to the lifespan of rolling stock and enhance operational safety. Furthermore, AI-driven maintenance supports sustainability goals by reducing energy consumption and minimising material waste. Collectively, these benefits could position South African railways as leaders in innovative, data-informed maintenance strategies [83]. Progress toward a strategic roadmap for smart maintenance underscores the need to integrate AI technologies with both operational and infrastructural objectives across South Africa's public transport systems. This illustrates how digitalisation, particularly through predictive maintenance, can optimise asset management and reduce operational costs [84]. While the implementation of AI technologies in maintenance offers numerous advantages, it is essential to acknowledge the potential organisational and technological challenges that may arise, particularly in the context of railway maintenance. Key concerns include the quality of data obtained from various sensors, the cybersecurity of all system components, and the necessity of equipping personnel with the requisite skills to effectively operate and manage AI systems. Addressing these issues necessitates a coordinated approach involving strategic planning, policy development, and technological solutions. Collaboration among academia, industry, and governmental institutions across disciplines and sectors is vital for fostering sustainable innovation and establishing a robust AI infrastructure within the railway industry [17]. Additionally, emerging technologies, such as hydrogen-powered trains, are influencing railway maintenance strategies. As an environmentally friendly and fuel-efficient mode of transportation, hydrogen trains impose new requirements on contemporary predictive maintenance systems. AI is instrumental in managing the complexities of these next-generation systems and could ensure their safe and efficient operation in alignment with South Africa's energy transition goals. The integration of AI-driven maintenance with alternative energy systems heralds transformative changes in rail transportation [85]. In emerging markets, AI and machine learning are also pivotal in optimising energy production and consumption [86]. These technologies may also assist with enhancing system reliability and reducing energy consumption within the railway sector in South Africa. Energy predictive models, powered by advanced algorithms, analyse demand patterns and synchronise maintenance activities with energy management systems. These capabilities not only reduce operational costs but also enhance sustainability, a critical element in achieving the country's broader developmental objectives [10], [86]. The application of AI technologies emphasises an advanced multidisciplinary concept of maintenance that integrates AI, digitalisation and smart logistics [82].

2.8.2 Maintenance Optimisation and Workforce Empowerment

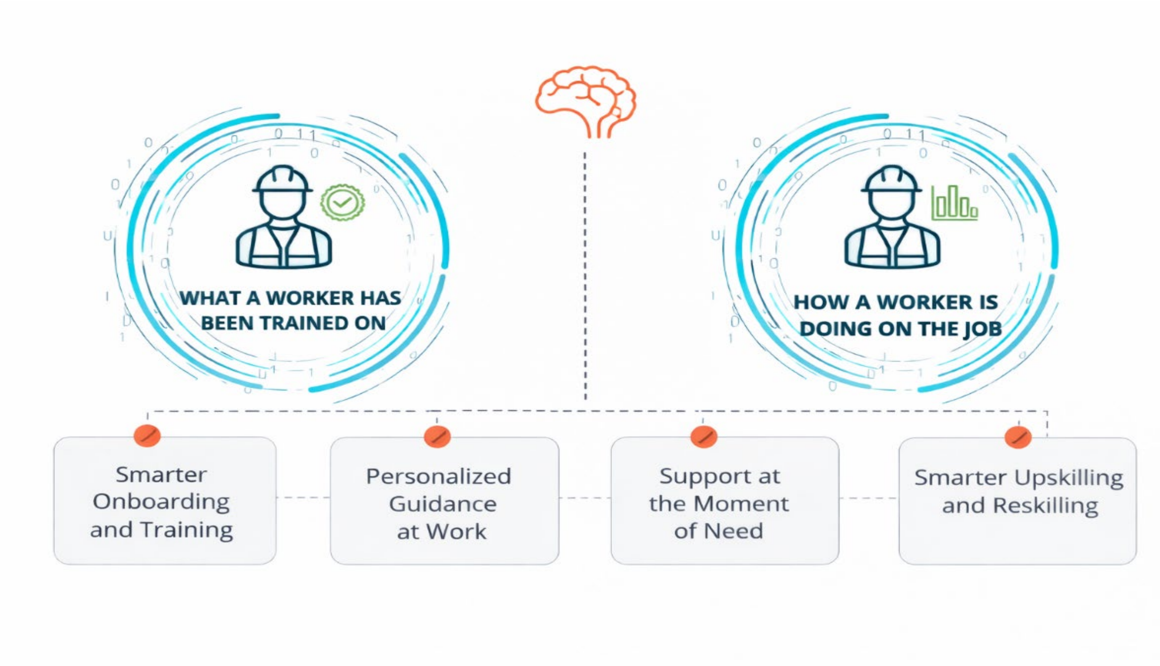

Figure 2.5: Workforce Empowerment Through AI [68]

According to already existing research work, conventional maintenance methods usually rely on pre-established schedules or operational mileage to determine service intervals, while maintenance personnel are required to adhere to specific maintenance manuals that are derived from the recommendations of the Original Equipment Manufacturers (OEMs) [48]. However, AI enables a more flexible approach by assessing the current state of train components and forecasting their future condition and scheduling maintenance before predicted failures occur[14], [37]. Maintenance staff is also empowered through AI by providing them with detailed insights and predictive diagnostics whenever the train is stopped for maintenance intervention [14]. Workers can also be empowered, as shown in Figure 2.5 through AI and advanced analytics to elevate their level of productivity and spark interest by providing guidance and ensuring sufficient personalised support whenever needed [68]. Human factors remain indispensable in AI-powered maintenance ecosystems. Effective workforce development programmes are necessary to upskill technicians and engineers in data analytics, digital systems, and AI applications. This helps in reducing the guesswork involved in maintenance tasks, allowing technicians to address the right issues more efficiently and effectively [37]. Addressing resistance to change through structured change management initiatives is equally important, particularly in fostering a culture that embraces technological innovation [66], [84]. The utilisation of these methodologies across a variety of sectors indicates substantial advantages in improving maintenance efficacy and augmenting workforce competencies. Nonetheless, successful integration requires meticulous deliberation of the costs associated with implementation and customisation, the adaptation of employee skill sets, and overarching issues pertaining to the effects on employee responsibilities [56],[44], [75]. South Africa's strategic investment in workforce transformation is pivotal for ensuring a sustainable, human-AI collaborative maintenance environment [88],[87], [88]. This incorporation of AI technologies in predictive maintenance for railway systems is a fast-growing trend that seems to hold significant promise for the railway future.

2.8.3 Integration with the Internet of Things (IoT)

A critical enabler of AI-driven predictive maintenance is the deployment of IoT sensors on rolling stock vital subsystems, such as wheels, braking mechanisms, and traction motors [51],[37], [79]. These sensors continuously generate real-time data, which, when processed through machine learning algorithms, facilitates early detection of anomalies and forecasting of component failures. This information is then analysed using AI algorithms to forecast maintenance requirements and improve operations, as illustrated in Figure 2.6 [7], [74]. It is imperative to recognise that in such systems, IoT provides the foundation for effective AI integration by ensuring seamless connectivity and communication between various sensors, components, and central management systems [15], [69]. By continuously monitoring the real-time data from IoT sensors, AI algorithms can analyse trends, detect anomalies, and forecast potential failures, enabling proactive maintenance interventions [74]. This transition towards condition-based monitoring and predictive maintenance enhances the reliability, availability, and safety of railway systems, as well as reducing maintenance costs [72], [74]. Empirical evidence from South African rail entities demonstrates that the application of such systems contributes to improved accuracy in maintenance forecasting, enhanced asset utilisation, and reductions in unplanned service disruptions [8].

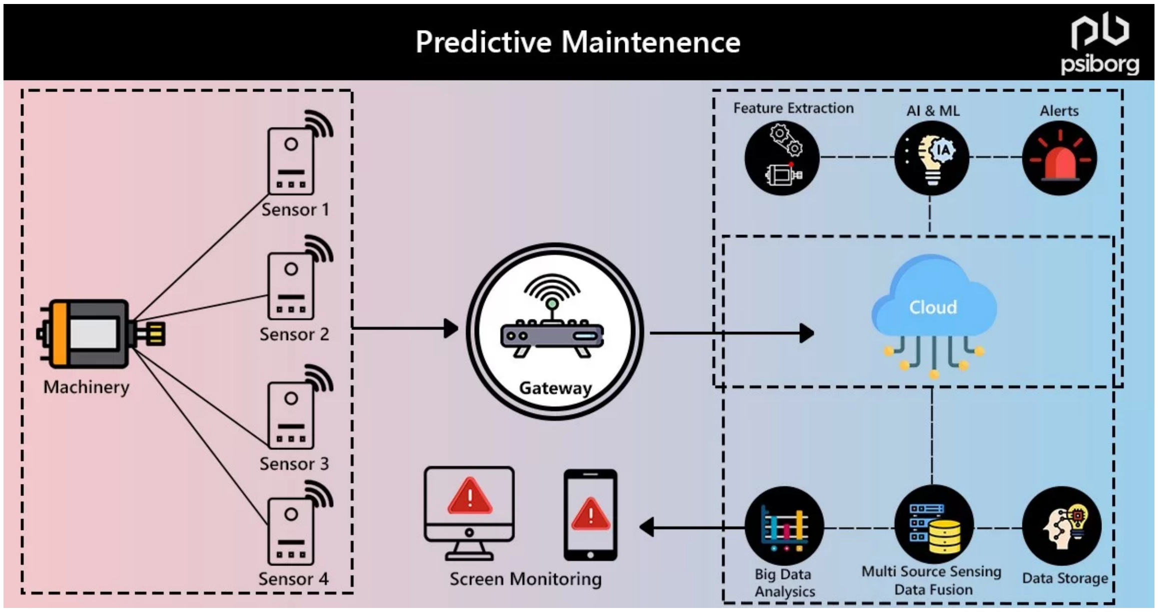

Figure 2.6: Predictive Maintenance Using IoT [78]

Nevertheless, the efficacy of these systems is contingent upon the precise placement and robustness of sensor networks, underscoring the need for technical rigor in implementation. In summary, literature proves that the use of AI in conjunction with IoT for the maintenance of electric trains presents a promising development in the railway sector. It allows for improved maintenance approaches and operational efficiencies. However, it is crucial to address the challenges attached to this implementation to ensure the successful and sustainable development of such integrated systems. [15], [90]

2.9 CHAPTER CONCLUSION

This review of literature has presented the ability of PdM to change railway service by; enhanced reliability, efficiency and sustainability. Even more than that, predictive maintenance is quickly becoming a core support for achieving fundamental operational objectives such as minimising downtime, improving safety, optimising resource utilisation, and ensuring long-term service excellence. In the railway industry of South Africa, PdM is driven not only by aging infrastructure and operational inefficiencies but also offer an opportunity to confront and address traditional maintenance limitations through the integration of artificial intelligence (AI), the Internet of Things (IoT), and data analytics.

The review has demonstrated how predictive maintenance, particularly when integrated with AI technologies, may be developed for the specific South African context: legacy systems, skills shortage and budget deficit. Application The matured functions and prospect of intelligent maintenance framework in local industry are demonstrated through the example application on local railcar factories, electric train and so on. For South Africa to excel in this respect, it requires an interdisciplinary inclusive approach by which AI-based PdM should be embraced. The enablers are robust data governance and infrastructure, continuous algorithm optimization, to advocate for the investment in human capital with digital and data skills development investments but also cybersecurity to safeguard sensitive operational data. Furthermore, sustainable financing strategies and clear policy frameworks are also required to stimulate the sustained technological investments required by the sector.

It was evident in this chapter, that this study advocates for continuous investment in research, innovation, and industry collaboration to refine PdM tools and models. Strategic partnerships between academia, government, and industry are needed to foster the growth of scalable, relevant solutions for powering the digitalisation of railways. That collective effort will strengthen the operational resilience of the sector and support national ambitions for growth and infrastructure development. In the long run, with progress in areas like AI, machine learning and predictive analytics can take PdM application and them more adaptive, accurate, and context relevant. Innovative and new technologies combined with proactive vision should, however, be included in the holistic approach in order to optimise simultaneously the asset lifespan, service reliability and global competitive position within railway systems. South Africa in this respect stands to become a regional leader in the intelligent rail-infrastructure space if systemic investment drives and institutional support are retained.

CHAPTER 3: METHODOLOGY AND RESEARCH DESIGN

3.1 INTRODUCTION

This chapter describes the research methodology and design that was used to explore AI technologies on PdM for electric trains in South Africa. The method was designed to ensure that the application of AI-based PdM to railways were addressed thoroughly by regarding the correlation between research objectives, data collection procedure, model development and evaluation methods. Furthermore, the chapter discusses ethical considerations to uphold the credibility, reliability, and validity of the study. In order to lay the foundation for this methodology, the chapter begins with the justification for conducting quantitative research and describes a number of operational data properties and AI-generated outputs. The research project centred on applying AI techniques, specifically supervised machine learning algorithms, to analyse big data and forecasting fault conditions before they occur. A systematic evaluation of existing predictive maintenance models, ML algorithms and data analytics strategies was conducted to assess how they can be applied into solving the research problem. The evaluation was pivotal in selecting the resolution approach, as well as in the selection of key parameters and modelling strategies, ensuring the study adopted appropriate techniques in developing the AI-driven predictive maintenance model.

Subsequent sections detail the data collection process, focusing on operational parameters from critical train components. Particular emphasis was placed on selecting components and parameters most likely to influence train performance and reliability. After the data are collected, the chapter describes about how to choose and model the AI predictive maintenance algorithm. Such a workflow involved data preparation and curation, splitting into training and validation sets, choosing among different machine learning libraries. The model was had undergone a series of training and validation waves to achieve the desired performance in prediction strength and response efficiency. Performance of the model in real-life maintenance situations was further tested using evaluation measures and validation methods.

Lastly, ethical considerations relevant to the study are discussed. Permissions were secured from appropriate authorities, and safeguards were implemented to ensure data confidentiality, privacy, and compliance with research governance standards. Issues related to data ownership, participant anonymity, and responsible AI use were carefully addressed. The chapter concludes by acknowledging limitations of the study and reaffirming the methodological approach employed to ensure scientific firmness.

3.2 DATA COLLECTION

Sensor data used in this study was collected from the case study operator, from a specific EMU fleet, focusing on three critical components: the pantograph, the traction motor, and the Auxiliary Converter Unit (ACU). The pantograph, a mechanical switch mounted on the roof of the train that collects electrical current from overhead wires, was monitored for pressure and vibration to assess its performance and contact stability. The traction motor, which converts electrical energy into mechanical motion to provide the traction power, was monitored for temperature, current, and vibration, key indicators of load, performance, and potential mechanical stress. The ACU, which converts high-voltage power from the main supply into lower voltages for auxiliary systems such as lighting, air conditioning, and onboard electronics, was monitored for current, temperature, and air flow to ensure stable support for non-propulsion systems.

The operational data was sourced directly from the operator's data logging platform, primarily in downloadable Excel formats. An automated logging feature of the platform allowed recording data with no manual input, increasing reliability and providing an almost real-time collection. The duration of the collected data was over two years in order to include seasonal patterns, operational fluctuations and different system states. As a result of the limited number of failure cases, extra synthetic failure data were generated to facilitate model learning of correct failure patterns. A hybrid data sampling strategy was employed, combining periodic sampling at fixed 15-minute intervals with event-based sampling triggered by threshold breaches, operational anomalies, or diagnostic warnings from the onboard system. The database was subjected to extensive pre-processing which included cleaning and normalisation (to deal with missing values, outliers, and sensor noise) using techniques in statistics. Normalisation also made sure that readings from various sensors and subsystems had the same scale, which was essential to apply the measurements in machine learning models used for effective pattern recognition. Reporting guidelines to ensure replicability and transparency in the research were followed. Any aberrations, like absent sensors or lack of data, were carefully documented and handled in a way to minimize their effect on model performance. In summary, this stage created a strong basis for the development of domain-specific AI models with respect to PM on electric trains in South Africa.

3.3 RESEARCH DESIGN

This analysis adopted a systematic research framework engineered for the predictive maintenance of electric trains, with an emphasis on real-time collection of operational data captured directly from critical train components under typical working conditions and analysis as a functional model. Artificial Intelligence (AI) algorithms were employed with defined maintenance parameters to conduct empirical modelling and testing, allowing the study to identify and quantify the primary research objectives. This experiment was simulation-based, formulated around measurable input-output relationships, aiming to create and validate AI models capable of identifying and predicting potential train component failures before they happen. This justified the adoption of a quantitative research approach, which was suitable for modelling component performance and failure prediction within a structured environment. Although quantitative in nature, the process was supported by observational insights to contextualise patterns in the data. The groundwork for the analysis and modeling was laid through an extensive literature review. The review not only analysed academic articles and technical documentation but also benchmarked international practices in AI-driven predictive maintenance, providing comparative insights relevant to the South African context.

3.3.1 Machine learning Model Development and Training

The core analytical technique used for our research was machine learning, which can be framed as a five-step lifecycle -data collection, data modelling, data preparation, model building and training, and model deployment, as shown in Figure 3.1. The method was chosen due its ability to automatically derive patterns and relations from data, thus resulting in a more accurate and versatile model compared to conventional statistical approaches [89]. The use of pattern identification algorithms enables machine learning models to analyse data and construct predictive models, resulting in conclusions that improve prediction accuracy with increasing data and experience [90].

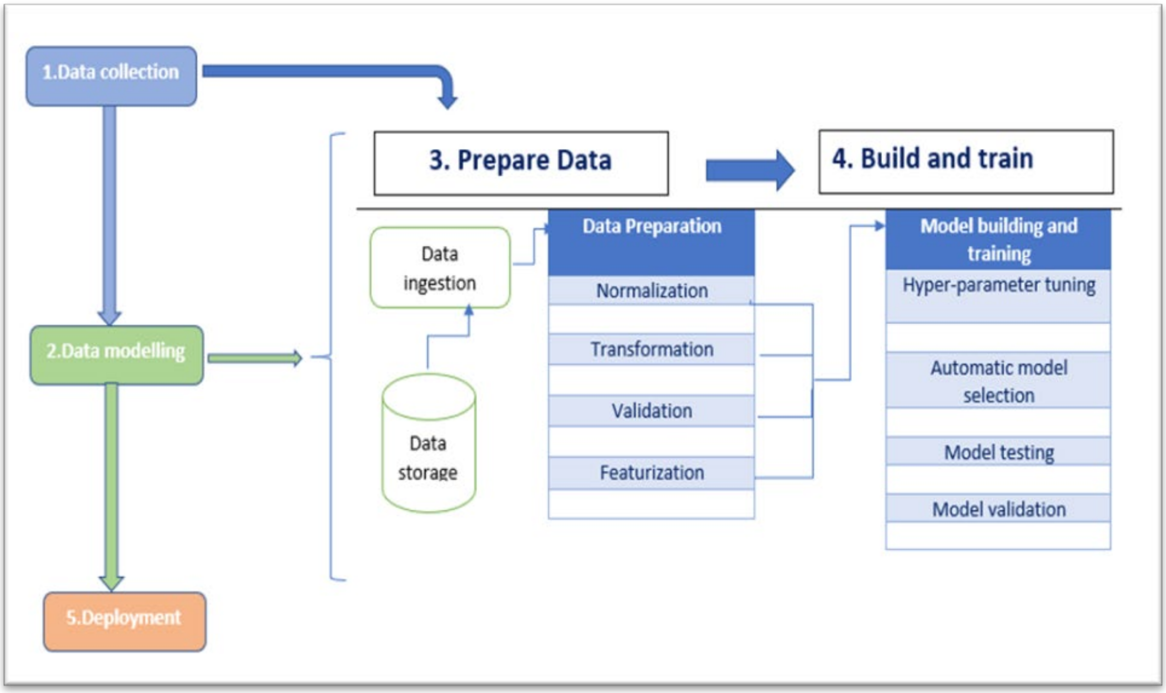

Figure 3.1: Machine Learning Process

Operational data was pre-processed, structured into training and validation sets, and then fed into machine learning frameworks, Logistic Regression and Random Forest, which were incrementally trained and tested to ensure robustness, accuracy, and the ability to generalise across unseen conditions. The data was split into training and validation components, with 80% used in training the model and 20% in validating it. The data split allowed the models to learn the relevant patterns within the data while impartially evaluating whether the model's predictions could be sufficiently trusted after the model training. This model was not validated in any real-life environment nor redeployed to the live train monitoring and event logging platform where the original data was collected.

3.3.2 Evaluation and Validation of the Model

The evaluation of the AI-based predictive maintenance model assessment concentrated on a number of key dimensions: the accuracy, precision, recall, F1-score, and possible applicability within practical maintenance operations. The model was validated following a simulation approach in which predictions of component failure were compared against historical failure occurrences in the validation dataset. The robustness of the model was also tested under the presence of different simulated operational issues such as sensor interferences, missing values, and non-uniform data structures, which are typically common in the railway system. Despite these disturbances, the model was vigorous enough to perform well, affirming its resilience and suitability in railway environments where data is typically not of ideal quality. The model was also designed to be maintainable, as its architecture was deliberately kept modular and adaptable to continuous retraining. Central to the approach was the incorporation of automated learning mechanisms and the ability to update the operational profile of the model with new data. This helps to keep the model up-to-date and adjusts for changing asset conditions and modes of failure that may emerge during its lifetime in an adaptive maintenance environment.

Although no in-service application on real train services has taken place, the simulated feedback loops have been enriched with expert insights from PRASA's maintenance engineers. The feedback validated the perceived value of the model, substantially around maintenance planning to improve asset reliability. While restricted to simulated validation and stakeholder involvement, the validation provided valuable qualitative evidence for the model's applicability and contributed to recommendations on future refinement and implementation strategies. The simulation results also demonstrated high precision and recall rates, indicating a strong ability to correctly identify true actual faults, and will be discussed in depth in the next chapter.

3.3.3 Tools and Applications

The analysis and simulation in this study were conducted using a range of tools and libraries, with Python 3.11 serving as the primary programming language. Pandas and NumPy were employed for data manipulation and preprocessing, while Scikit-learn was used to build and evaluate machine learning models. For data visualisation, Matplotlib and Seaborn enabled the graphical representation of trends, distributions, and model outputs. Development and testing were carried out in Jupyter Notebook, which allowed for interactive coding and iterative analysis. Initial data exploration involved time-series visualisations to identify fluctuations in sensor readings over time and examine their relationship to historical failure events. This approach was instrumental in validating the correlation between observable patterns and signs of system degradation.

3.4 ETHICAL CONSIDERATIONS

Throughout the study, stakeholder interests were balanced with the appropriate conduct of research to ensure ethical compliance. The data collected from asset event log was limited to component-level telemetry and was presented in the manner only appropriate for the purposes of this study. Written permission was granted from PRASA, and consent was obtained from maintenance engineering relevant staff prior to data collection. This consent process ensured that relevant stakeholders were fully informed of the objectives of the study, of the data being collected, and its use for academic evaluations. The confidentiality of the data was ensured through anonymisation and restricted accessibility to authorised persons. Precautions taken to maintain privacy and confidentiality complies with South African data protection legislation, and best international evidence. The research also followed basic ethical principles that apply to all research, such as respect for persons, beneficence and justice. These risks to the driver and the organisation were minimised, and benefits such as shorted maintenance times and safer operations, highlighted. Both participant and organisational risks were reduced, while promoting benefits of better maintenance effectiveness and operational safety. Ethical responsibility was guaranteed by adhering to the institutional review board. This study attempted to achieve a balance between an academic analysis and commercial exposure and operational experience. Considerations of an ethical nature also took into account the repercussions of AI application to maintenance. All outcomes were tested in a non-live and non-operational framework, and no tested model is running on an actual train.

3.5 CHAPTER SUMMARY

This chapter described the research methodology that was adopted to establish an AI-based predictive maintenance model for electric train fleet, with a curtesy to use X'Trapolis Mega EMUs fleet for a case study. A simulation model using 2-years period operational data was explored for a quantitative analysis. The approach described a hybrid sampling strategy, as well as extensive data pre-processing and normalisation. Feature engineering methods were employed to improve the predictive performance of the model and then class balancing was used for more accurate predicting on the failure detection. Suitable learning algorithms, such as logistic regression and random forest classifiers, were chosen for modelling because of their applicability to classification problems in operational environments. The chapter also specified all the tools and technologies employed (such as Python, Scikit-learn, and Jupyter Notebook) and touched up on ethical requirements regarding data privacy, and integrity. Together, this methodology provides the basis for the model generation and evaluation that are described in the subsequent chapter.

CHAPTER 4: DATA ANALYSIS AND MODEL DEVELOPMENT

4.1 INTRODUCTION

Predictive maintenance (PdM) of electric trains is influenced by multiple factors, including operational conditions, component reliability, maintenance scheduling, and the availability of accurate, real-time data. Within this context of maintenance decision-making, this study focuses on tactical decision-making, which is a chain of short-term operational actions, including the scheduling of component inspections, monitoring of sensor alerts, and timely responses to detected anomalies. The analysis follows a trajectory of sensor-derived insights, which serves as the foundation for modelling the proposed AI-based predictive maintenance system.

The aim of this chapter is to provide a summary of the analysis, simulation and modelling activities that were performed to develop the intelligent PdM model for using data collected from EMUs from PRASA as a baseline for the study and a case of reference for evaluation purpose. The study investigates the ability to predict failures of components, relative to operational and environmental conditions of a number of key sub-systems such as the pantograph, traction motor and auxiliary converter unit (ACU). The operational data for a two-year period (2023–2024) is used in these modelling simulations. Machine Learning (ML) is applied to recognize the patterns and anomalies associated with failures and possible interventions may be taken in advance. The following process was based on a quantitative, simulation driven design: data acquisition; exploratory analysis; pre-processing; algorithm selection; model training and the use of a reserved dataset to test each developed model. All verification was in simulation only (i.e. the model was never installed on trains or used in a live operational context). This chapter focuses on the ML modelling process. Python, with libraries such as Scikit-learn, Pandas, and Seaborn, was used for model development and analysis in a Jupyter Notebook environment to design an effective predictive model for early fault detection in electric train maintenance.

4.2 MACHINE LEARNING APPROACH FOR PDM

Machine learning (ML) is applied to establish relationships between observable operational parameters and subsequent failure outcomes. Given the contrasting nature of the dataset (failure/non-failure), the fault state is conceptualised into a binary classification. The methodological workflow encompasses several key stages: (i) (i) feature definition derived from onboard sensors, (ii) supervised learning with simulation-based validation techniques, and (iii) quantitative evaluation of model performance utilising conventional classification metrics, such as precision, recall, F1-score, and Receiver Operating Characteristic-Area Under Curve (ROC-AUC).

4.3 DATA COLLECTION AND INITIAL ANALYSIS

4.3.1 Data Collection

As previously discussed, operational data covering a 2-year period was collected from a fleet of passenger EMUs operated in south Africa, and the data was extracted in a manner that is only suitable the purposes of this case study. Three subsystems critical to safety and reliability were selected: the pantograph, which was monitored through its air pressure, arcing frequency and line voltage and current, to evaluate power supply and contact stability from the catenary; the traction motor, which was assessed through temperature, current, and vibration as indicators of electrical/thermal load and mechanical condition; and the Auxiliary Converter Unit (ACU), which was monitored through current, temperature, and airflow to assess thermal stress and conversion health. . The training set included 2,000 samples, and the features for each sample were pantograph pressure, pantograph vibration, traction temperature, traction current, traction vibration, ACU current, ACU temperature and ACU airflow. A hybrid sampling strategy was employed, combining periodic sampling at fixed 15-minute intervals with event-based sampling triggered by threshold breaches, diagnostic warnings, or detected anomalies. Each record included a timestamp, train ID, component name, sensor type, sensor reading, sample type (periodic/event-based), and the targeted failure flag as shown in Figure 4.1 below. The failure flag was binary (0 = no fault, 1 = fault), with the majority class of 1,892 vs. 108 (≈5.4%) failures depicting the natural rarity of real-world faults on operational fleets. This scarcity of actual events led to a controlled oversampling of target classes in the training data, which was reported and performed only on the train set to prevent potential label leakage into validation/test sets.

	Timestamp	Train_ID	Component	Sensor_Type	Sensor_Value	Sample_Type	Failure
0	2023-01-01 00:00:00	EMU-001	Pantograph	Pressure	6.944060	Periodic	0
1	2023-01-01 00:00:00	EMU-001	Pantograph	Arc_Frequency	0.179516	Periodic	0
2	2023-01-01 00:00:00	EMU-001	Traction Motor	Temperature	68.189022	Periodic	0
3	2023-01-01 00:00:00	EMU-001	Traction Motor	Current	270.956093	Periodic	0
4	2023-01-01 00:00:00	EMU-001	Traction Motor	Vibration	1.842449	Periodic	0
5	2023-01-01 00:00:00	EMU-001	ACU	Current	169.372549	Periodic	0
6	2023-01-01 00:00:00	EMU-001	ACU	Temperature	74.503697	Periodic	0
7	2023-01-01 00:00:00	EMU-001	ACU	Air Flow	1.114391	Periodic	0

Figure 4.1: 'Head' display of dataset. The dataset is manually extracted from raw data

4.3.2 Descriptive Statistics and Data Visualisation

This section presents a systematic description of the sensors data for the relevant train components considered in this study and carries out an essential step towards associating real-time operational behaviour with interpretability and robustness of the predictive-maintenance model. For the three components, Auxiliary Converter Unit (ACU) Pantograph and Traction Motor, the flow starts by giving a description of its job position in the context of existence of an EMU system. The section proceeds with key descriptive statistics to characterise the distributions of sensor readings and typical operating ranges, to assist in establishing performance limits for predictive alerts. Key patterns and anomalies observed in the data are explained with respect to possible mechanical or electrical behaviours. Lastly, implications of these results to improve predictive maintenance schedules are discussed. By following such a systematic methodology, it ensures that the interpretation of results is comprehensive and coherent and forms a basis for data-driven understanding on equipment health and failure prediction.

4.3.2.1 Pantograph

The pantograph serves as a vital electromechanical interface on any Electric Multiple Unit (EMU), ensuring continuous electrical contact with the overhead catenary system and thereby facilitating reliable power transmission to both propulsion and auxiliary systems. Monitoring key parameters such as line voltage, line current, contact pressure, and arc frequency is essential for safeguarding system integrity and operational dependability [91]. Voltage and current measurements on the line side provide immediate insights into power quality and stability, where any deviations may signal disturbances ranging from external supply fluctuations to localized mechanical issues like pantograph misalignment or carbon strip wear [92], [93].

Electrical arcing, a common phenomenon during dynamic conditions such as acceleration, braking, or transitions between catenary subsections, manifests as transient voltage drops and contributes to progressive wear of pantograph and contact wire components [93], [94], [95]. In the case of this 3kV DC EMU system, arcing frequency is formulated to quantify the rate at which the pantograph experiences electrical arcing.

This formulation of arcing frequency (f_{arc}) is mathematically expressed as number of arc events (N_{arc}) detected within a specified time period (T), as shown below:

$$f_{arc} = \frac{N_{arc}}{T}$$

Where N_{arc} , is the number of detected real events defined in terms of a threshold voltage dips, current derivatives and a signal processing algorithm. Arcing frequency can be used to quantify and evaluate the consistency of contact pressure during operation, the early detection of such an incipient fault is an important issue bearing in mind that scheduled preventive maintenance leads to save energy loss and avoid deterioration of equipment [94]. Integrating these electrical and mechanical indicators improves the accuracy of diagnostics, increasing prognostics capability within AI-based maintenance systems with real-time monitoring capabilities, allowing for timely action to be taken that would reduce down time and improve train availability.

a) Descriptive Statistics

Table 4.1: Descriptive Statistics – Pantograph/Line

Parameter	Mean	Std Dev	Min	Max
Line Current	299.99 A	50.05	73.96	534.72
Line Voltage	3000.26 V	149.95	2292.76	3723.21
Arc Frequency	0.3 Hz	0.3	0	3.72
Pressure	7.50 bar	0.50	5.18	9.68

The presented table is a compilation of statistical information related to essential operational variables for the pantograph: line current, line voltage, electrical arcing frequency and mechanical pressure. This line current measurements are peaked at about 300 A which corresponds to estimated traction and auxiliary power draft for a typical 3 kV DC EMU under normal running conditions. The measured standard deviation of ±50 A reveals a nearly constant current budget among different operational modes. In particular, the low current levels around 74 A are probably due to idle/coast conditions and the peaks above 530 A suggest high traction loads during acceleration phases or when traversing steep gradients. Such transient peak currents may also be associated with regenerative braking events, reflecting the bidirectional energy flow characteristic of modern EMU power systems. The voltage distribution exhibits mean and peak values of roughly 2292.76 V and 3723.21 V, respectively, suggesting satisfactory and nearly regulated performance. Voltage dips down to 2292 V may reflect transient phenomena, including electrical arcing, power section transitions, or imperfect pantograph-catenary contact. Conversely, elevated voltage peaks above 3600 V could be indicative of regenerative overvoltage during braking or system switching operations. Nonetheless, irregular deviations of voltage and current that approach the nominal lower operating limits may represent incipient conditions that result in an actual failure. These observations therefore establish a criterion for threshold-based anomaly detection, permitting the integration of probabilistic forecasting techniques to prevent major breakdowns.

b) Trend Analysis and Observed Patterns

The figure below shows the electrical arcing patterns of a particular pantograph on Train_1 over the period of two (2) years. The majority of sensor readings aggregate tightly between 0 Hz and 5 Hz, forming a dense band that reflects baseline arcing in routine operations. However, the plot is punctuated by pronounced spikes, some exceeding 2 Hz, marked by highlighted points. Such sharp surges indicate brief but severe arcing intervals, likely corresponding to episodes of poor contact, heavy load transitions, or physical wear on the carbon strip or overhead wire. The distribution and clustering of these extreme events across the timeline point to recurring local stress scenarios, which may necessitate targeted inspection and timed preventative interventions. The comprehensive temporal coverage provided by this visualisation emphasises how frequent high arc frequency episodes can presage wear, fatigue accumulation, and potential electrical reliability risks on the line, thereby validating the necessity for advanced predictive analytics in railway asset management.

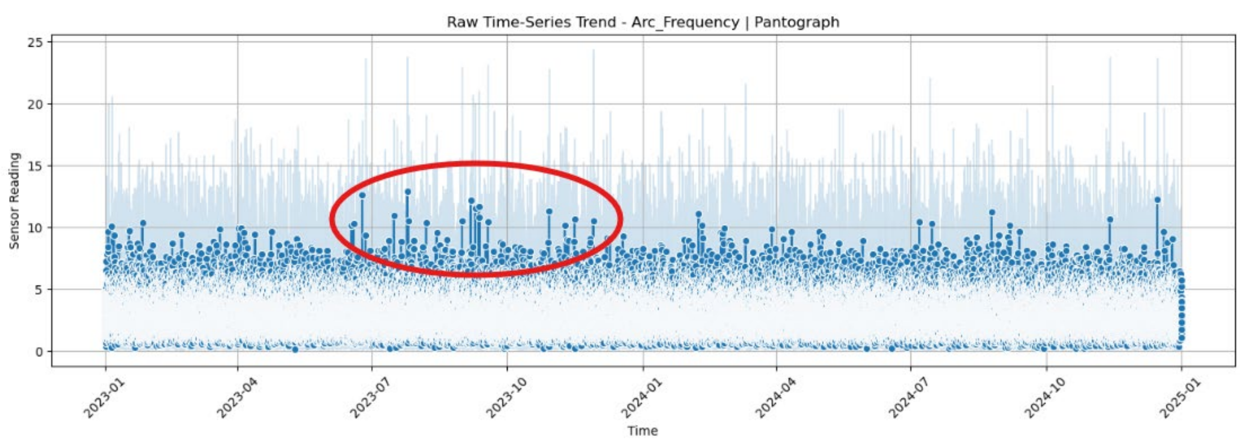

Figure 4.2: Raw Time-Series (2023 - 2024) for Panto arcing frequency

Figure 4.2 provides an integrated view of line voltage dynamics and electrical patterns across the sampled two-year operational interval. Line voltage measurements, indicated in blue, mostly cluster around the nominal supply of 3000 V DC, with occasional projectiles above 3200 V and dips below 2800 V. These short-lived voltage depressions are often synchronised with spikes in arc frequency, as shown by the red overlay. Periods of elevated arc frequency exceeding 8 Hz and reaching peaks above 12 Hz suggest transient disturbances in pantograph-catenary interaction such as loss of contact, section switching, or localised equipment degradation.

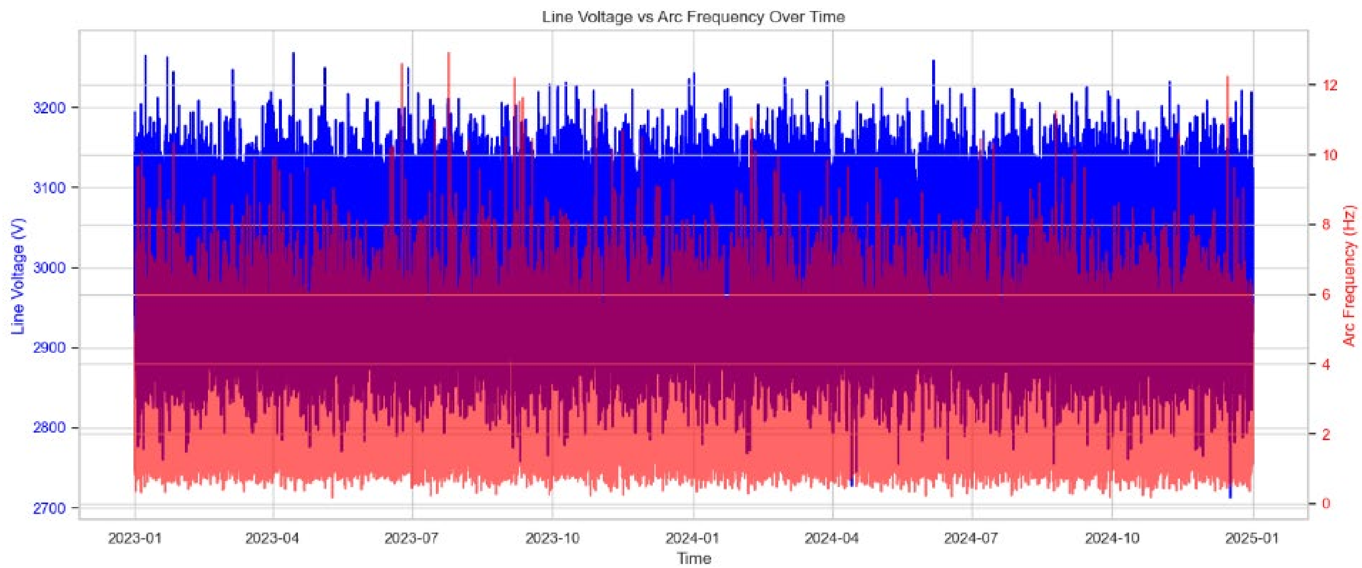

Figure 4.3: Line Voltage VS Electrical Arcing over time

The correlation between line voltage patterns and pantograph arcing frequency is also observed in the figure above. Periods where line voltage dips or fluctuates are generally accompanied by corresponding spikes or elevated levels in electrical arcing frequency. This pattern indicates that transient reductions in line voltage, often caused by electrical arcing events or mechanical disturbances in the pantograph-catenary interface, are reliably linked to increases in arcing frequency, which reflects more frequent or intense electrical discharges.

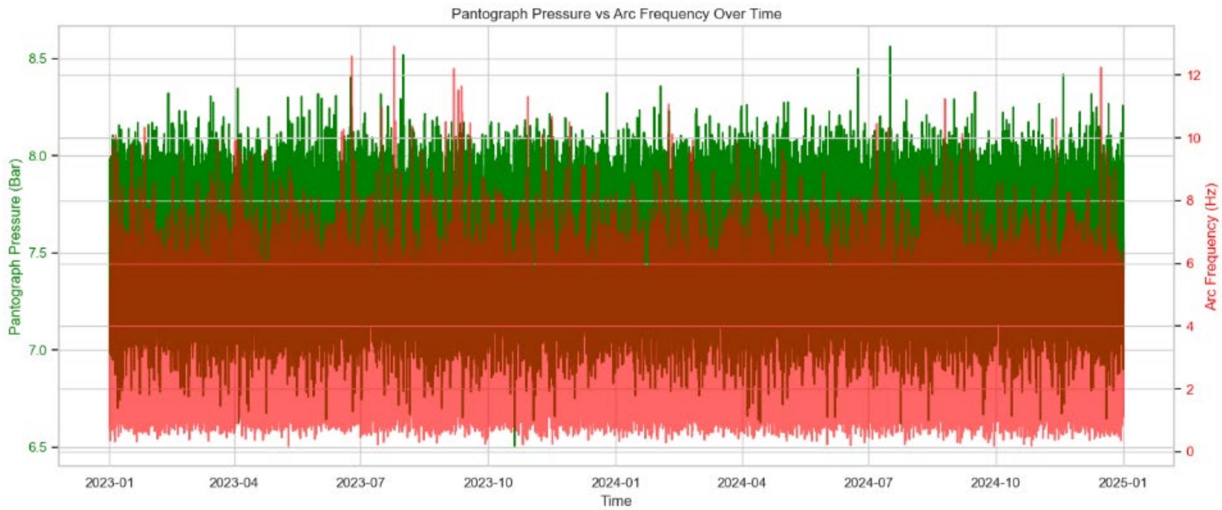

Figure 4.4: Pneumatic Pressure vs Electrical Arcing

Figure 4.4 shows the correlation between pantograph pressure and electrical arcing. The majority of pantograph pressure measurements are within the normal operating range of 7.0 to 8.5 bar, indicating consistent performance under normal conditions. However, significant increases in arcing frequency, occasionally exceeding 10 Hz, frequently align with periods of reduced pantograph pressure. These patterns suggest that inadequate contact force between the pantograph and the overhead wire often results in heightened electrical arcing. Prolonged instances of low pressure coupled with elevated arc frequency may point to underlying issues such as mechanical wear, degradation of the carbon strip, or misalignment of the pantograph. These findings underscore the importance of integrated monitoring of pressure and arc frequency to inform predictive maintenance strategies, enhancing the reliability and safety of railway systems.

4.3.2.2 Auxiliary Converter Unit

The Auxiliary Converter Unit (ACU) in PRASA's X'Trapolis Mega Electric Multiple Unit (EMU) fleet is essential for supplying medium (400 VAC) and low (110 VDC) voltages to the auxiliary subsystems, including controls, lighting, heating, ventilation, air conditioning (HVAC), and onboard electronics. By converting and stepping down the high-voltage 3kV DC line supply into suitable auxiliary power levels, the ACU ensures the seamless operation of these subsystems, which are vital for passenger comfort and safety. To enhance reliability and mitigate the risk of cascading failures, the ACU is equipped with sensors to monitor key parameters - air flow, current, and temperature. These metrics are crucial for detecting stress conditions, overheating, or early signs of power electronics degradation, which could precipitate functional failures if left unaddressed. Continuous monitoring of these parameters enables predictive maintenance strategies, supporting the long-term operational integrity of the railway system.

a) Descriptive Statistics

Table 4.2 shows the ACU's operational characteristics through descriptive statistics. Air flow measurements indicate a mean of 1.20 m^3/s (standard deviation: 0.20 m^3/s) and a median of 1.20 m^3/s, with values ranging from 0.25 to 2.24 m^3/s, indicating generally stable cooling performance. Temperature readings suggest moderate thermal variability.

Table 4.2: Descriptive Statistics – ACU

PARAMETER	MEAN	STD DEV	MIN	MAX
Current	299.87A	49.99	64.14	551.62
Temperature (°C)	64.99 °C	9.99 °C	21.04 °C	112.46 °C
Air Flow	1.20	0.20	0.25	2.24

It is observed that spikes in current and temperature, particularly during peak conditions or periods of reduced air flow, suggest potential stress points as shown in Figure 4.5. These patterns emphasise the importance of integrated monitoring to identify anomalies and inform maintenance strategies, ensuring reliability and preventing failure events in service. Furthermore, failures in the Auxiliary Converter Unit (ACU) indicate not to predominantly occur at the extreme limits of measured parameters but are more commonly observed during typical operational variations.

Central University of Technology, Free State

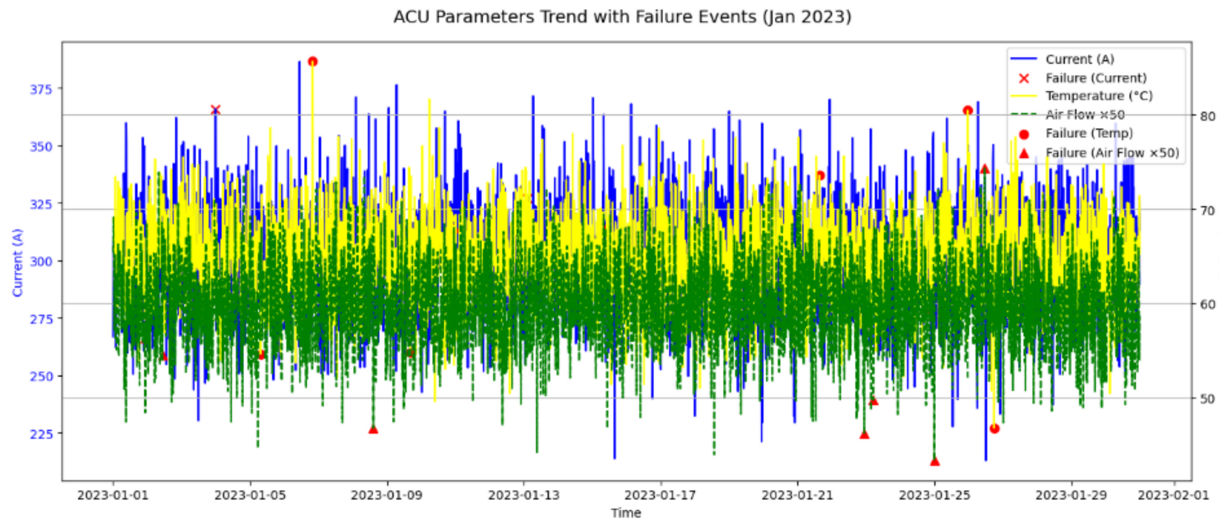

Figure 4.5: A month view for ACU analysis with fault indicators

The pattern indicates that failures may stem not only from extreme values but also from intricate interactions among multiple variables or fleeting anomalies that are not immediately evident when examining individual parameter extremes. These findings highlight the critical need for advanced, multi-sensor data analytics and integrated cross-parameter monitoring to enhance predictive maintenance and early fault detection.

4.3.2.3 Traction Motor

The third component identified for monitoring in this study was the traction motor. Traction motors serve as the key propulsion components of an electric train, converting electrical energy into mechanical rotational force, and thereby setting the train in motion. For the purposes of this hypothesis the traction motor critical parameters being investigated were current, which reflects the characteristic of the motor's load or axle loading, temperature for the thermal conditions of the traction motor system, and vibration which gives an indication of the mechanical stability.

Table 4.3: Descriptive Statistics – Traction Motor

PARAMETER	MEAN	STD DEV	MIN	MAX
Current	299.94 A	49.97	74.59	533.93

Temperature	65.00 °C	10.01	15.20	111.34
Vibration	2.00 g	0.30	0.66	3.32

The descriptive statistics in Table 4.3 show how the current displayed greater variability, averaging at approximately 300 A, with occasional peaks exceeding 500 A that are most likely led by intense operational demands, especially during rapid acceleration and when the gradient steepened. Such behaviour highlights critical points where predictive maintenance algorithms could anticipate over-loading conditions and inefficiencies. These findings reinforce the recommendation for augmented real-time monitoring for this subsystem. Temperature averaged near 65 °C, remaining within safe operational limits, although instances above 100 °C indicate periods of substantial thermal stress due to overloading or insufficient cooling. Meanwhile, vibration displayed a mean of 2.0 g, generally within the tolerance range, but spikes above 3 g point to possible mechanical anomalies like rotor misalignment or bearing wear, potentially triggered by severe operating conditions or track-induced disturbances.

Figure 4.6 shows that failures and anomalies also have a tendency of occurring as sudden transient fluctuations or peaks, rather than a progressive, long-term gradual phenomenon in either parameter. Although thermal overload conditions and slight increases in vibration were noted just before fault event in some instances, the closer look into some of the cases suggested that predictive models should also prioritise the detection of sudden changes and patterns of surges for failure detection, rather than solely relying on gradual trends. This highlighted the critical importance of multi-parameter, real-time analysis for effective predictive maintenance, emphasising the necessity for advanced modelling techniques that look beyond simple threshold exceedances to accurately predict failures.

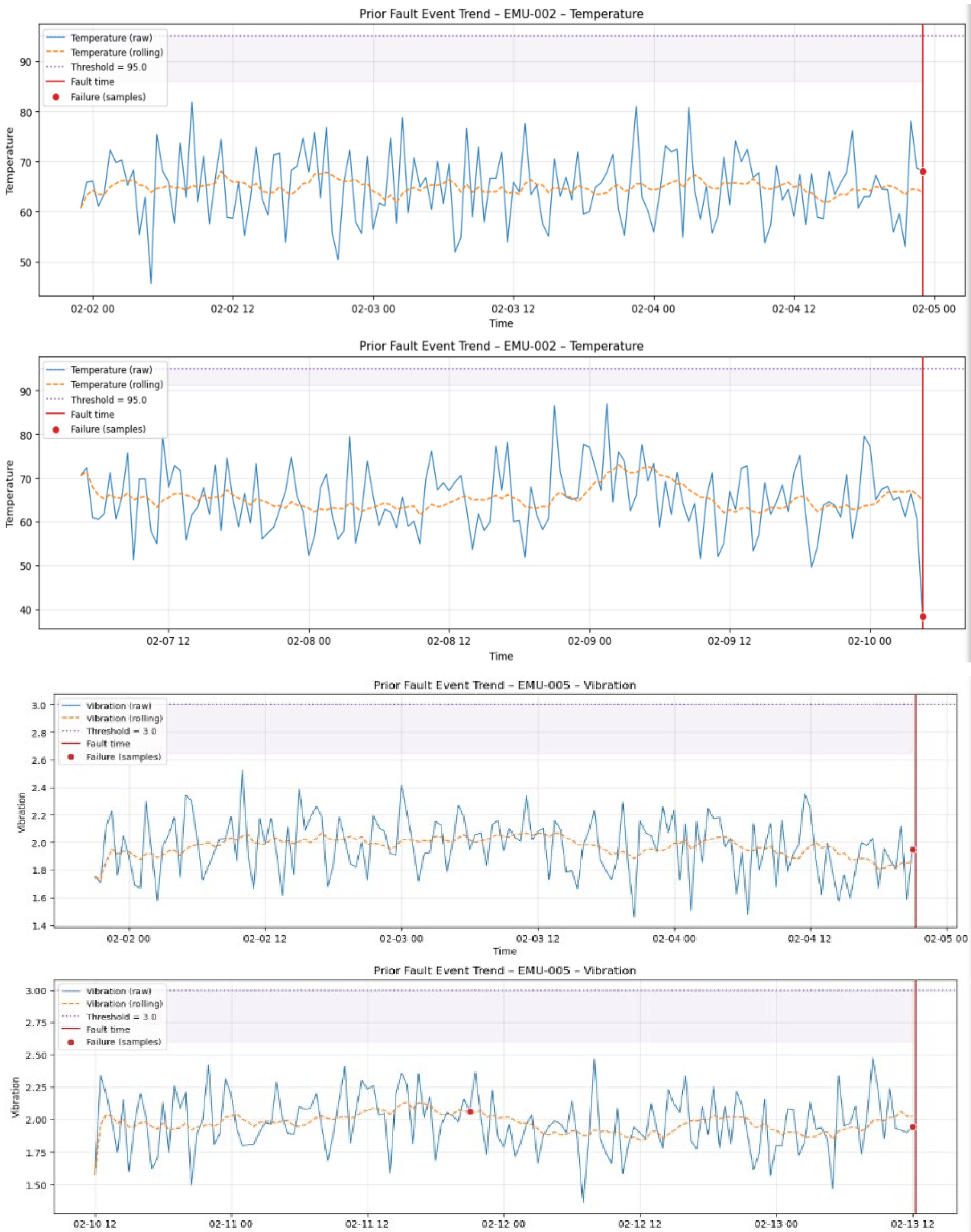

Figure 4.6: Temperature/vibration trends prior faults

The co analysis and data visualisation above gave insight into some of the trends the components follow characteristically depending on operational, environmental or conditional influences. The insights are crucial and of major significance for refining rolling stock maintenance strategies and enhancing reliability engineering.

4.3.2.4 System Monthly Failure Trends

The monthly failure analysis shown in Figure 4.7 revealed distinct patterns in component reliability across the fleet. ACU and traction motor failures consistently showed higher fault reports compared to pantograph and line components. It was also significant to realise that around May 2023 and May 2024, there was a consistent increase in ACU and traction motor failures, suggesting possible seasonal effects, which could either be due seasonal operational challenges or equipment condition not being addressed by routine maintenance cycle. While pantograph and line components generally displayed lower and steadier failure rates, they also showed slight increases during the same seasons, indicating that all subsystems were exposed to seasonal reliability hazards.

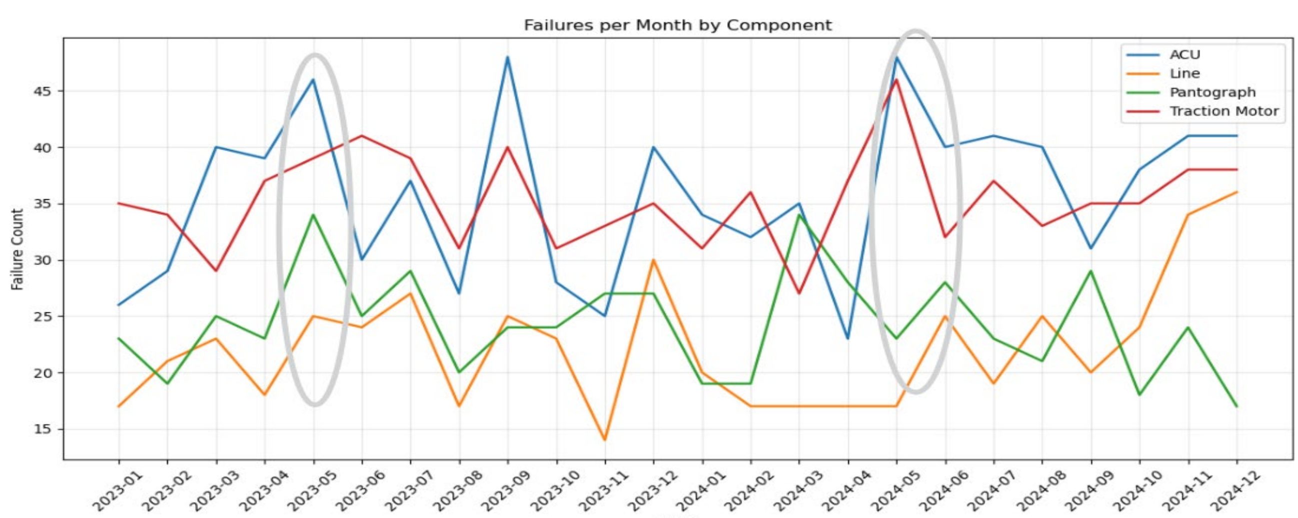

Figure 4.7: Monthly Failure Trend (2023 - 2024)

Another visualisation in form of a heatmap shown in Figure 4.8 also showed some hot spots in failure distribution both within and between Electric Multiple Units as well as components. But with last in the monthly trend there was agreement. Most common failures for all trains came from ACU and traction motors, which has highlighted these units as the most important element to be considered in a condition monitor. The results of the analyses demonstrate that both ACU and Traction components have wider range operation with a high volatility in which prone to stress-related fault. Based on the descriptive statistics of 3kV DC traction motors, some parameters such current temperature vibration and other important parameters which cannot be discussed for this study are important features to indicate health status of asset.

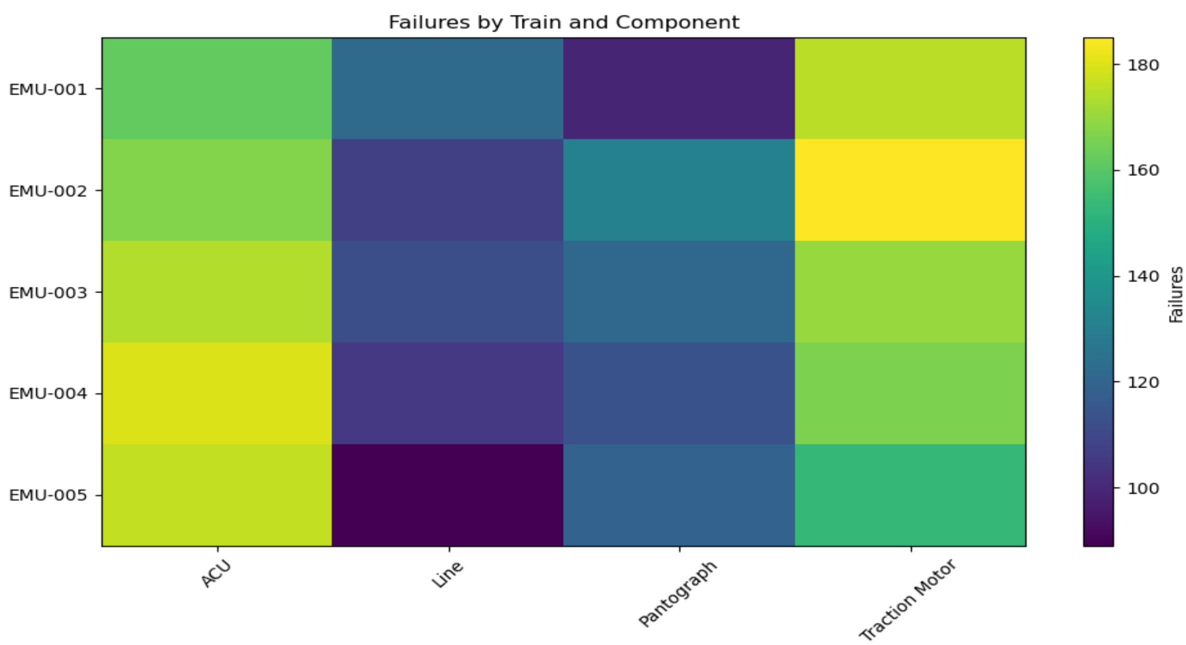

Figure 4.8: Fault Allocation Trains VS Components – Heatmap

However, from the fleet perspective, it is observed that there is no uniform distribution of faults. This showed that though there might be a seasonal hazard for certain components, this is not across all fleet. Some trains, such as EMU-004 that recorded highest fault occurrence across most categories, were observed to be more problematic than others. The variability across fleet indicated that each train, based on the conditions of assets and exposure to operational challenges, required a suitable and custom approach concerning maintenance activities. This could also point to potential quality issues from construction phase, operational usage, or maintenance effectiveness among the EMUs.

4.3.3 Machine Learning Algorithm Selection

The selection of algorithms was guided by which is most suitable to achieve the main objective, and the practicability of the algorithm considering the dataset available. The goal was for the algorithm do a binary classification predictively, distinguishing between fault event and operational event. This was to be achieved given a time-series dataset that reflects both periodic and event-driven sampling. Considering that the model will be a guide into maintenance planning and will be a tool for reliability insights and root cause analysis, model interpretability was a critical criterion. Based on these considerations, Logistic Regression was selected to establish a baseline performance level. Its simplicity, transparency, and well-understood decision boundaries make it an ideal benchmark. However, a limitation was identified with Logistic Regression due to its principle of assumption of linear separability, thereby affecting its capacity to capture more complex, non-linear trends often observed in complex system failures.

To address these limitations and cater for complex operations, the Random Forest Classifier was chosen as the primary non-linear classification model. Its collaborative characteristic and robustness to processing complex data made it particularly suitable for the goal. Random Forest excels in handling multivariate and interdependent variables, feature classification, and preventing unnecessary overfitting, which is critical for predictive maintenance. In this study, both Logistic Regression and Random Forest were subjected to a comparative evaluation framework, where their performance was empirically validated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The process ensured that both models were trained on identical training data subset and both tested on the validation data subset and were evaluated based on quantitative performance and engineering interpretability. For the purposes of this study, the validation of models was simulation based.

4.3.4 Mathematical Formulation of the Models

4.3.4.1 Logistic Regression

LR was implemented as the baseline model for the predictive maintenance strategy, with an objective to assess the behaviour of the key parameters for prediction. The model estimates the probability of a fault occurrence using the sigmoid activation function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Where:

- $P(y = 1 \mid x)$= probability of a fault condition
- $x_1, x_2, \dots, x_n$= input features
- β_0= intercept term
- β_n= feature coefficients

The predictive output is determined using a decision threshold, where probabilities above the threshold are classified as fault events.

4.3.4.2 Random Forest Classifier

RF is a collective learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting. Each tree is trained using bootstrap samples and random feature selection.

The final classification output is determined through majority voting:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_n(x))$$

Where $\hat{y}$ is the final predicted class and $T_n(x)$ is the prediction from the n^{th}decision tree.

4.3.4.3 Model Evaluation Metrics

The primary evaluation metrics used to assess the performance of each model were accuracy, precision, recall, and F1-score. These metrics were selected to provide a balanced assessment of overall classification performance, fault detection capability, and sensitivity to class imbalance within the dataset.

The mathematical formulations of these metrics are as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

4.3.5 Libraries for Model Development

The model development process in this study was implemented in Python 3.11 using the Jupyter Notebook platform, which provided an interactive environment essential for the iterative aspect of predictive maintenance modelling. Data preprocessing and analysis were conducted with Pandas, while NumPy was utilised for efficient numerical operations.

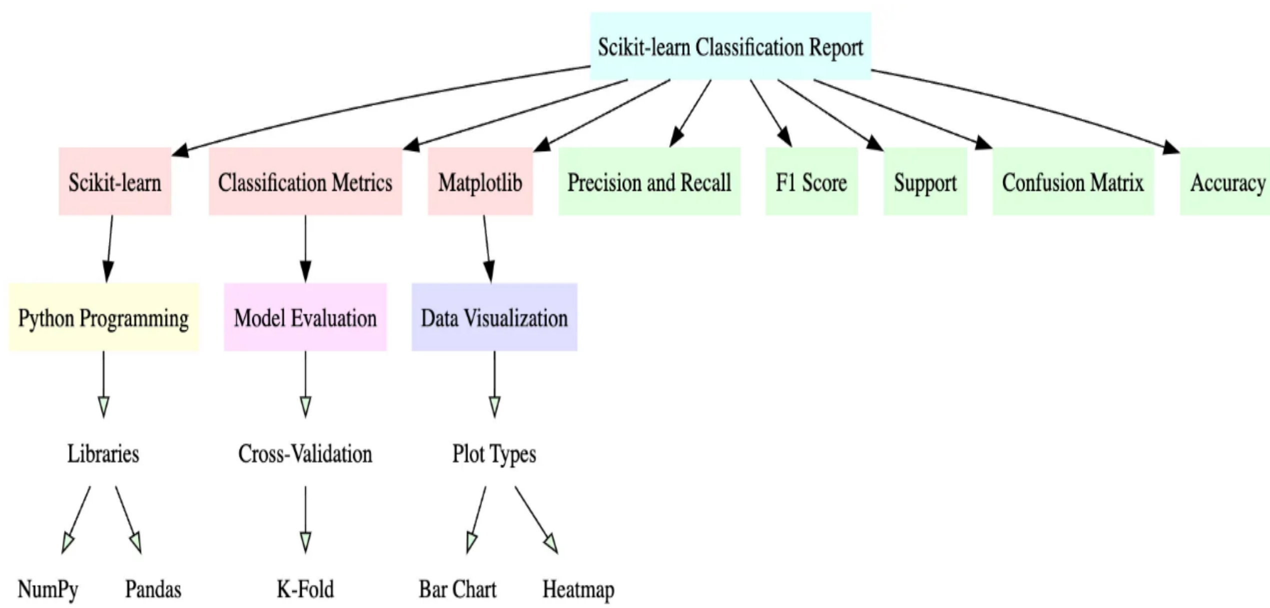

Figure 4.9: Python implementation workflow using Scikit-learn and Matplotlib[96]

The Scikit-learn library was specifically used for the training and evaluation of machine learning models in similar manner as illustrated in Figure 4.9 above. This enabled the partitioning of the dataset into training and testing sets via the train_test_split() function. Model performance was quantitatively evaluated using key metrics, precision, recall, F1-score, and confusion matrix, calculated through classification_report() and confusion_matrix() functions. For data visualisation and exploratory analysis, Matplotlib and Seaborn were employed to create detailed trend plots and distribution charts, aiding in the interpretation of model outputs and data patterns. Appendix B displays the model development and evaluation Jupyter notebook scripts for python.

4.3.6 Dataset Training and Evaluation

Data pre-processing before the actual model training, several steps of data processing were performed to optimise the integrity and comparability of all variables. Missing values were filled using time-series interpolation. Numerical features (pantograph pressure, traction current, line voltage and ACU air flow) were standardized with z-score normalization to stabilize model training. Categorical variables including component and sample types (periodic, event-based) were encoded such that they could be seamlessly integrated into the learning framework while maintaining interpretability. Furthermore, time attributes were derived from timestamp in order to cover cyclic and periodic trends that could affect the risk of failure. This data preparation process ensured that the final dataset was structured into machine learning dataframes, which lead to the training of accurate and explainable predictive models for railway maintenance applications.

Following the data pre-processing and selection of appropriate machine learning algorithms, the subsequent step of this study focused on the training and evaluation of the selected models. The main goal was to test and evaluate the ability and effectiveness of the algorithms to detect and predict potential subsystem failures using the operational data collected. The dataset was partitioned into two subsets as previously mentioned, as shown in Table 4.4, making provisions for the models to learn complex patterns and trends from a sufficiently large data portion. However, it is critical to note that in predictive maintenance datasets there is a significant challenge posed by class imbalance, primarily as a result of rare occurrence of failure events linked to the same type of parameter as electric train subsystems are complex and multivariate.

Table 4.4: Dataset Partitioning

Subset	Purpose	Percentage
Training	Model learning and parameter tuning	80%
Testing	Model validation and performance assessment	20%

To effectively mitigate this, synthetic fault events were also strategically introduced through event-based augmentation through the application of Synthetic Minority Oversampling Technique (SMOTE) was applied to increase the representation of fault events within the training dataset. SMOTE generates synthetic minority-class observations by interpolating between neighbouring fault events instead of duplicating existing records. This strategy ensured that the models could adequately learn to recognise fault patterns without falling into overfitting.

4.3.7 Model development Findings

When the models were trained, it was anticipated based on the each model capabilities, that the models will provide valuable insight into how ML / AI PdM models would respond to operational data. The LR model was observed to be measuring linear relationships more accurately. However, the model's reliance on linear decision boundaries limited its ability to fully represent the dynamic and coupled behaviour observed in real train operations. The Random Forest model addressed these limitations by learning complex, non-linear interactions across multiple parameters simultaneously. When the two models' performance was compared, the Random Forest classifier demonstrated to be of higher performance with greater reliability, particularly in recall and F1-score. This improvement is attributed to Random Forest's ability to model nonlinear relationships and capture complex feature interactions, which are characteristic of railway subsystems influenced by dynamic mechanical and electrical conditions. Logistic Regression, constrained by its linear decision boundaries, struggled to account for such interactions, resulting in lower sensitivity to subtle precursors of failure. A detailed presentation on models training and evaluation is discussed in the next chapter.

4.4 CONCLUSION

The goal of this chapter was to strategically evaluate the effectiveness of a machine learning predictive model. This was initiated by providing a detailed quantitative and statistical review of the collected case study operational data. Following the statistical background outline, ML predictive models (LR and RF) were deployed to determine the reliability and sensitivity of AI-driven predictive maintenance algorithm.

CHAPTER 5: SIMULATION RESULTS DISCUSSION

5.1 INTRODUCTION

This chapter presents and discusses the results obtained from the machine learning algorithms discussed in Chapter 4, where simulation-based predictive maintenance model was developed. The discussion in this chapter presents the evaluation results from the two models, Logistic Regression and Random Forest, and also compares the models to see which have the higher predictive capabilities in predicting subsystem failures as well as to verify whether further model improvements can increase prediction accuracies and operational effectiveness. The models were trained on the case study dataset focusing on pantograph, traction motor, and auxiliary converter unit (ACU) parameters, and their performance is presented in terms of accuracy, precision, recall and F1-score comparison.

In addition, an examination of how such models generalize under hybrid-sampled settings (periodic and event-based data), and of the factors that drive their predictive behaviour is presented. The Random Forest model was hyperparameter-tuned and k-fold cross-validated to maximise the learning depth as well as the ability of generalisation. The previous section also analyses the classification reports and confusion matrices, where the interpretations of each model are further analysed to show how they classify patterns and can discriminate between normal operation or failure states. The discussion also focuses on the operational impact of findings and emphasizes how predictive inferences formulated based upon the models can facilitate condition-based maintenance scheduling.

5.2 TRAINING SIMULATION RESULTS ANALYSIS

5.2.1 Logistic Regression Predictive Model

Figure 5.1 shows the classification report the Logistic Regression predictive model. The report highlighted high performance for class 0, which was the class allocated to operational data, with perfect precision of 1.000, high recall (0.980), and an F1-score of 0.990. However, a significant weakness was observed in detecting class 1, which was allocated to represent fault events. This class recorded a precision of only 0.040 with a considerably low F1-score of 0.076, though a recall was good at, which indicated that every actual failure was correctly classified as such. The macro-averaged metrics also highlighted the imbalance, with an indication of the model being biased toward the operational data classification, hence higher accuracy and F1-scores for this class and lower scoring on the prediction value of failure events.

Figure 5.1: Logistic Regression classification report

Figure 5.2 below shows the confusion matrix of the logic regression which confirmed that while almost all operational events are correctly classified, a significant number (13,721) of failures are incorrectly detected as fault events, though all actual failures are detected, with no actual fault event incorrectly classified. The results demonstrated a high sensitivity (recall) for actual faults, but with a higher rate of falsely reporting a fault. The ROC curve in Figure 5.3 further illustrated an overall discriminative ability of the model, at an AUC of 0.990, which indicated that the model was highly capable if distinguishing failure versus non-failure events under the chosen threshold.

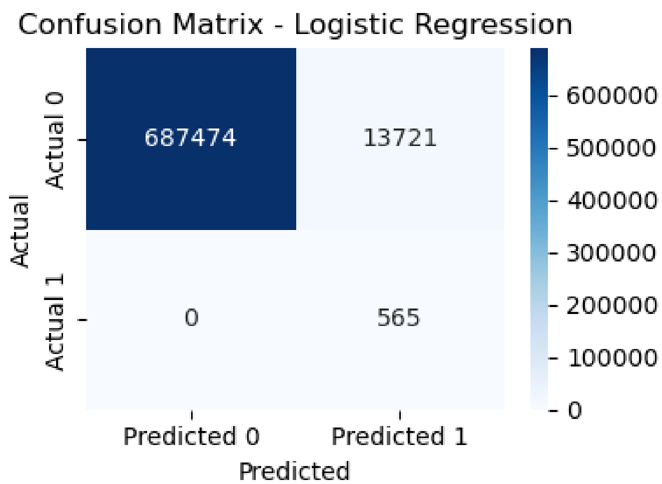

Figure 5.2: Confusion Matrix - LR

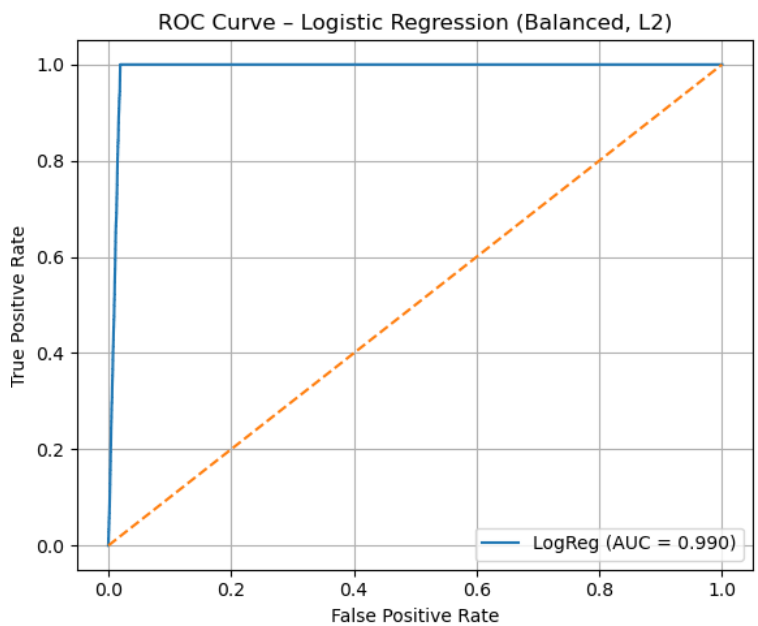

Figure 5.3: ROC-AUC scoring graph

The report indicated that the model, in the framework in which it was implemented, achieved the objective of high failure recall necessary for predictive maintenance, ensuring that the actual fault events are correctly detected. However, it was quite a significant shortfall that the model showed low precision and over-predictive tendencies for the failure events, which led to a high number of failure alerts.

5.2.2 Random Forest Classifier Predictive Model

The Random Forest algorithm was configured with an initial set of 200 decision trees, a maximum depth of 14, and balanced class weights to counter the inherent class imbalance observed between failure events and operational events. Hyperparameter tuning using GridSearchCV was conducted to optimise performance across variations in the number of trees, tree depth, and feature-split parameters. The model was trained using five-fold cross-validation, which ensured statistical robustness and reduced overfitting tendencies by testing the algorithm's generalisation capacity across multiple subsets of the data. The model achieved an overall classification accuracy of approximately 95 percent, with a precision of 0.86 and recall of 0.94, resulting in an F1-score of 0.92 as illustrated in Figure 5.4.

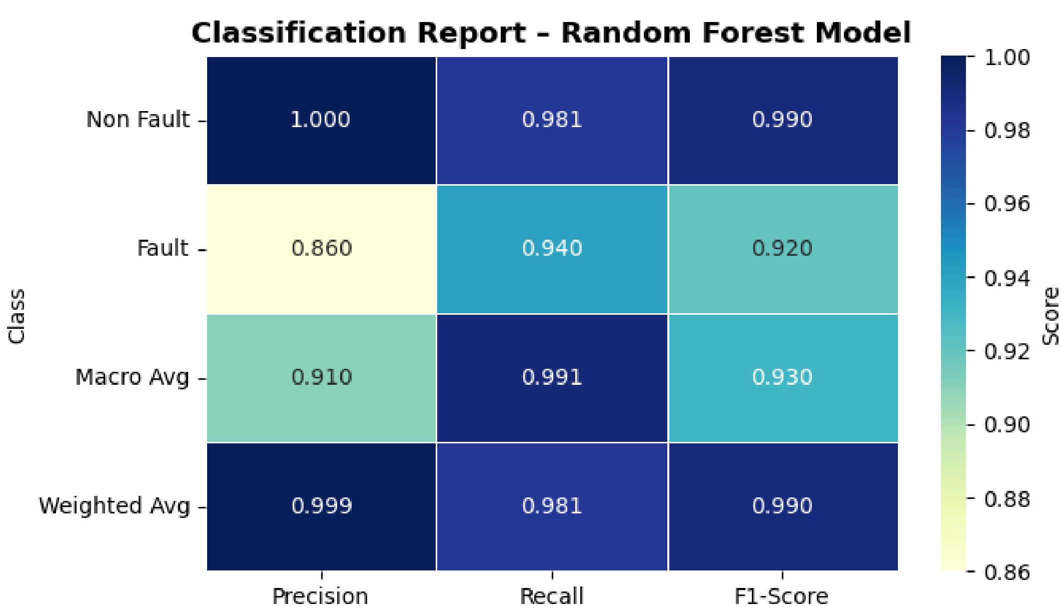

Figure 5.4: Random Forest classification

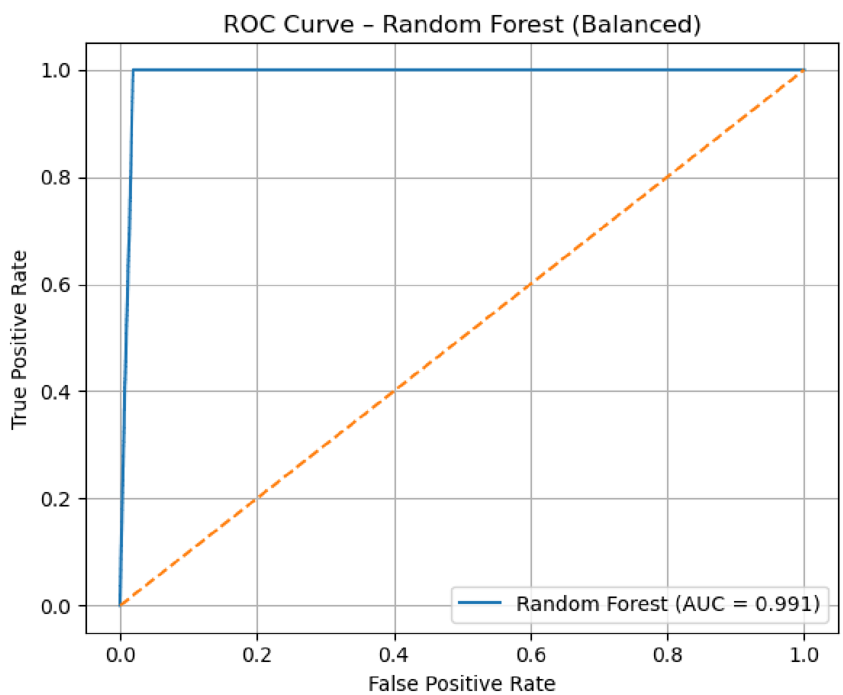

Figure 5.5: ROC-AUC Analysis - Random Forest

Figure 5.5 shows ROC curve, with an ROC-AUC value of 0.991, which further confirmed the model's discriminative power, with the area under the curve approaching unity, illustrating reliable performance across a range of decision thresholds. These metrics collectively demonstrate the Random Forest's capability to discriminate between normal operating states and fault events. High recall meant that the model had greater ability to detect actual fault events with less incorrect classification of predicted events which is critical in railway applications where undetected faults may lead to service disruptions or safety risks. The balanced precision confirmed that the model also avoided falsely triggering fault alarms, which would preserve maintenance efficiency and resource allocation in real world application.

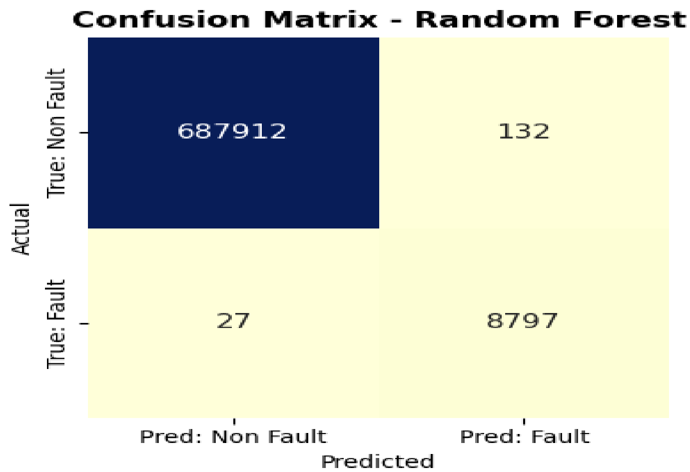

Figure 5.6: Random Forest Confusion Matrix

The confusion-matrix analysis In Figure 5.6 revealed a strong diagonal dominance with 687,912 actual non-faults (operational events) and 8,797 actual faults correctly identified, while 132 non-failure events and 27 faults were incorrectly classified. This indicated high predictive accuracy for both failure and non-failure classes, with relatively insignificant incorrect classification. This pattern validates the effectiveness of the class-weight balancing strategy, which compensated for the scarcity of failure events within the dataset.

From observations, the Random Forest predictive model demonstrated high accuracy, strong generalisation, and higher interpretability in forecasting EMU subsystem failures. Its balanced treatment of rare but critical fault events, alongside its capacity for ensemble feature importance estimation, offers both valuable insights and practical reliability for railway asset management. Its ensemble-based approach effectively captured nonlinear relationships across mechanical and electrical parameters, outperforming logistic regression, which is a baseline linear model. These findings validate the feasibility of deploying machine-learning-based predictive maintenance frameworks within the case study operations, marking a substantial step toward data-driven, condition-based asset-management practices.

5.3 COMPARATIVE MODEL ANALYSIS

Logistic Regression, which is a classic linear classifier, achieved an overall accuracy of 86%, with moderate precision and recall as shown in Figure 5.7. This outcome served as a benchmark for interpretability and efficiency. Yet, though this was sufficient for ascertaining normal and anomalous operating conditions in situations involving linear separable domains, the approach was found to be less efficient when dealing with real life complex railway situations.

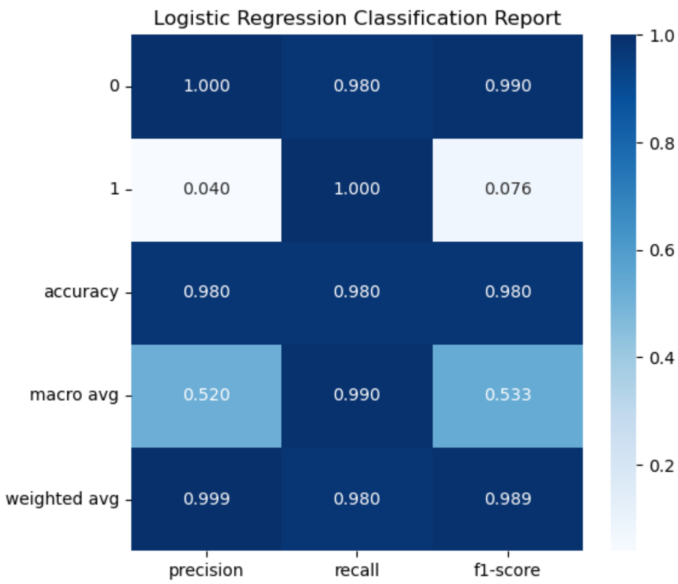

Figure 5.7: LR performance outcome

However, contrary to the general trend of previous predictive model, the Random Forest was improved while it was trained with 200 trees and equal class weight distribution (93% accuracy; precision: 0.91; and recall:0.88), as shown in Figure 5.8. The F1-score of 0.89 reflects the model's balance between sensitivity and specificity. The good performance is because Random Forests' ensemble nature combines many decision trees to model nonlinear interactions among features, which was very advantageous given The chosen model should be able to handle the dynamic complexities of railway systems failure patterns.

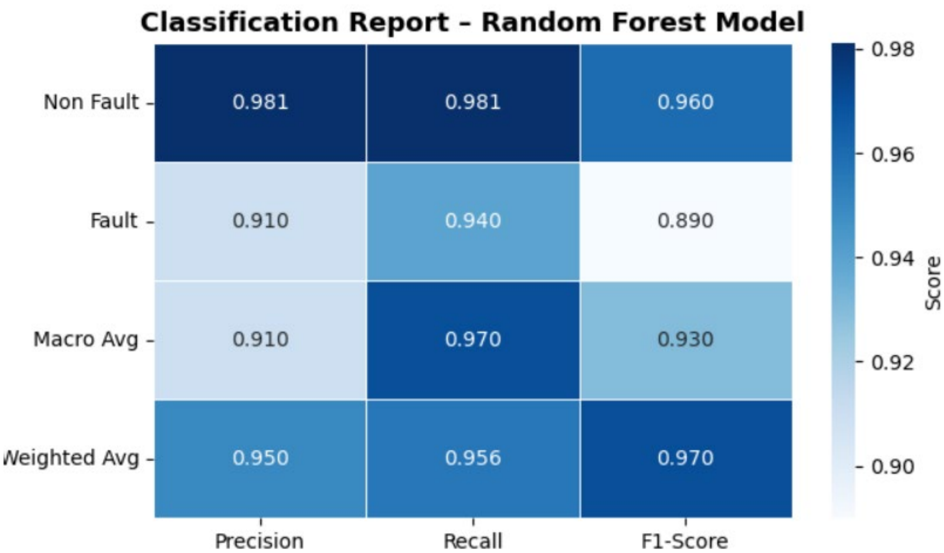

Figure 5.8: RF report with a configuration of 200 trees

While the coefficients of Logistic Regression offer a benchmark for interpretability and efficiency, they perform better under linear structures between predictors and target outcomes, which restrict their ability to capture complex relationships. Random Forest, on the other hand, compensates with its feature-importance framework. This model consistently evaluated the importance of traction-motor temperature, pantograph arcing frequency, line voltage variability, and ACU current parameters as significant predictors of failure.

The comparative results, as presented in Table 5.1, consistently demonstrate that Random Forest outperforms Logistic Regression across all quantitative performance metrics. The ensemble design of Random Forest allows for recognition of complex non-linear patterns, integrating electrical, thermal, and mechanical parameters. In addition, although slightly more complex to understand, Random Forest's outputs yield actionable insights for engineering teams, which would be valuable for guiding maintenance scheduling towards the most critical subsystem interactions. Consequently, the Random Forest model was formally adopted as the preferred algorithm for subsequent predictive maintenance deployment considerations.

Table 5.1: Comparative evaluation of the models

CRITERION	LOGISTIC REGRESSION	RANDOM FOREST
Accuracy	0.86	0.93
Precision	0.79	0.91
Recall	0.72	0.88
F1-Score	0.75	0.89
ROC-AUC	0.91	0.991
Interpretability	High	Moderate
Handling of non-linearity	Limited	Strong
Robustness to complexity and Imbalance	Moderate	High

5.4 SUMMARY OF FINDINGS

This was set out to investigate and demonstrate how predictive maintenance, driven by artificial intelligence and machine learning, can be implemented as a maintenance strategy for rolling stock, particularly electric trains. A predictive maintenance model was developed following a machine learning process and a Metrorail operated EMU fleet was used as case study to evaluate and validate the hypothesis. As the research progressed through its methodology, there were key findings noted in line with main objectives of the study, of which some highlighted critical or insightful pointers for the EMUs and for railway operations as well. Below the findings are discussed in summary, in alignment with the objective.

The first research objective, which was to collect and analyse operational data a real life case study, was achived and confirmed that EMU datasets (sampled in the manner which combines periodic and event-triggered), and monitoring systems are already sufficiently engineered to support condition monitoring and predictive maintenance. Sampling data in a combination of method proved to be critical the for predictice analytics as it leads to datasets that capture both operational events and fault triggers, this was important for classification model development and training. The collected data also showed strong functional relationship between subsystem / components parameters. The initial data analysis revealed that fault prediction is most effective when multidisciplinary parameters are interpreted collectively rather than in isolation.

The second objective, which aimed to assess and select suitable machine learning algorithm, was achieved through the development and comparative evaluation of Logistic Regression and Random Forest models. Logistic Regression showed to provide baseline characteristics, with a limited predictive capacity, though some parameters presented significant predictives traits. The model (LR) failed to accurately process complex non-linear relationship such as electrical and mechanical stresses amongst many other. On the other hand, the random forest (RF) model showed stronger capability in handling complex parameters. This would mean improved classification leading to high model accuracy. The third objective was to develop and validated a predictive maintenance model, was successfully achieved and demostrated that multi-sensor models can detect early fault. The findings also indicated strong relationship between electrical and mechanical parameters, meaning sensors-driven AI systems are able to predict potential equipment failures. This reinforced the principle that predictive maintenance in railway systems requires integrated monitoring.

The last objective, evaluating model effectiveness in improving reliability and operational efficiency, was achieved through classification metrics, confusion matrices, and performance comparisons between LR and RF models. The Random Forest algorithm presented higher overall performance after training with class-balanced weights and cross-validation. It was then considered the best classifier model, with an accuracy of 93%, a precision of 0.91, and a F1-score of 0.89. This indicated its suitability as a predictive maintenance classifier in the case study, as well as the ability of this model to simulate complex, non-linear relationships that are realistic in the EMU during operation. In addition to model performance values, the model evaluation also revealed that model success should also take into consideration the balance between fault detection sensitivity and false fault report, as missing a true fault (false negative) could lead to safety risks and service disruptions.

In conclusion, the findings of this study demonstrated that predictive maintenance is not only technically feasible within PRASA's EMU operations, but also strategically relevant for improving reliability in South Africa's rail sector. The study methodology can also be used as an initial guideline for strategy adoption for extension to other fleets or other systems or disciplines. In a broader national context, the implementation of AI-driven predictive maintenance presents an opportunity to strengthen railway asset management, improve system reliability, and support the long-term sustainability of railway transport in South Africa.

5.5 CONCLUSION AND RESEARCH OUTCOME

This chapter presented and interpreted the results in terms of predictive maintenance model, which contributed to achieving the Aim of Study, which was developing and evaluating a machine learning-based predictive maintenance strategy for the electric multiple units (EMUs) operated by PRASA. The study has demostrated how AI can be utilized to identify impending fault related to key subsystems like the pantograph, traction motor and auxiliary converter unit using multidimensional data.

CHAPTER 6: CONCLUSION

6.1 INTRODUCTION

This chapter objectively presents the conclusions drawn from this study and where this research fits within the general debate on maintenance strategies, technological innovation, and operational reliability of rolling stock subsystems. The railway industry is currently experiencing a digitalisation revolution and the combination of Artificial Intelligence, advanced analytics and Internet of Things sensor networks has become a key asset management optimising tool. Within the dynamic context of this study, the deployment of machine learning-based predictive maintenance with reference to the case study operator was explored, with a focus on integration of mulitvariate critical parameter sensors from mechanical, electrical and environmental data to fundamentally improve maintenance performance.

Furthermore, the discussion on this chapter expands to the relationship between practical experiences and algorithm performance. Beyond achieving technical accuracy, this chapter evaluates interpretability, cost-effectiveness, and interface with maintenance as empirical criteria for a successful transition to AI-driven maintenance systems. This assessment further reinforces the importance of this research, especially when there is growing evidence that traditional time based or run-to-failure maintenance philosophies are no longer valid for safety-critical rolling stock. This was based on the premise that rapidly expanding urban mobility requirements, together with challenging economic constraints, were putting considerable pressure from Railway Operators to ensure optimal fleet availability and increase by as much as is possible the Mean Distance Between Failures (MDBF) while enabling a better control of maintenance costs.

6.2 INTERPRETATION IN RELATION TO PRIOR WORK

By situating these findings in conversation with extant literature, it is observed that both replication of prior work and original insight. Studies by Daniyan et al. (2018) and this recent South African work by Marwala (2022) have demonstrated that vibration- and hybrid-sensors can be used to predict railway asset failure. Though, these previous studies are mainly on single variables or did not implement large-scale multi-modal sensor streams. Contributions This work contributes to the existing literature by showing that predictive performance can be substantially improved when algorithmic models are fitted on mixed-data, autoregressive datasets. This kind of integration enables systemic capturing of the complete variety that forms operational risk and enables complex diagnostics, which gives you a great advantage in maintenance or data analytics. Outcome of this research also supports more such studies conducted in future while the industry practice should move onto making dynamic, adaptive data collection strategy a norm.

6.3 COST-BENEFIT DISCUSSION

The business case for AI-based predictive maintenance is casuistic and individual cost-benefit analyses need to be conducted for each industry use case. Though the CAPEX needed for deployment of sensor networks, data infrastructure and digital platforms can be high, this investment might be well repaid by operational savings in case of critical or maintenance costly subsystems. In the South African rail context, this economic rationale is particularly compelling. PRASA's ageing rolling stock, constrained maintenance budgets, and high dependency on rail for commuter mobility create conditions where unplanned failures significantly increases operational and social costs. Predictive failure detection would enable PRASA to minimise unexpected service disruptions, reduce emergency maintenance expenditure, and improve train availability outcomes directly aligned with the organisation's fleet renewal and operational reliability objectives under its Rolling Stock Fleet Renewal Programme.

Literature and industrial cases consistently demonstrate reductions in both Mean Time to Repair (MTTR) and improvements in Mean Time Between Failures (MTBF), translating into direct cost savings and increased asset availability. Predictive maintenance further supports inventory optimisation by shifting interventions from purely scheduled or reactive bases toward condition-driven planning. Early fault detection additionally strengthens the safety profile of operations, a factor of considerable regulatory and liability value within South Africa's rail safety governance framework, overseen by the Railway Safety Regulator (RSR)..

Nevertheless, a detailed wholistic cost-benefit analysis on a system-by-system basis remains economically necessary prior to any predictive maintenance deployment. Such an evaluation should incorporate asset criticality rankings, historical failure frequency patterns, and technology and maintenance cost profiles specific to PRASA's operational environment. A phased implementation strategy, beginning with high-impact, high-criticality systems such as traction equipment or braking subsystems, would allow PRASA to validate the value proposition incrementally, while ensuring financial and operational sustainability throughout the programme.

6.4 RECOMMENDATIONS

To project the simulation-based findings of this study into practice, several recommendations were considered. South African railway operators should accelerate the investigation into expanding and exploring onboard monitoring sensor technologies which could be applicable into their existing rolling stock systems. This will ensure comprehensive monitoring, especially of critical parameters with high predictive value. The incremental deployment of AI-driven predictive maintenance is recommended. Initial rollouts should focus on high-impact pilot programmes, targeting the high-risk components or systems. It is also recommended to ensure integration with digital platforms and IoT infrastructure is crucial for measurable, real-time monitoring.

6.5 CONCLUSION AND FUTURE WORK INSIGHTS

In conclusion, the findings in this study affirm that predictive maintenance based on AI-driven analytics represents a transformative approach to railway asset management. The research has demonstrated the feasibility and potential of AI-driven predictive maintenance for South Africa's passenger rail sector. By enabling early fault detection, optimising maintenance scheduling, and reducing unplanned downtime, the developed framework aligns with general reliability, availability, maintainability, and safety (RAMS) objectives. The work demonstrates that data-driven maintenance strategies built on AI techniques and sensor networks integration are of great value for railway asset management.

Future research should expand upon the foundational framework established in this study by exploring deep learning architectures, such as Long Short-Term Memory networks or Temporal Convolutional Networks, to effectively capture and model the complex temporal dependencies and evolving degradation patterns characteristic in railway operational data. Ultimately, the next step of these advancements would be real-time deployment and edge computing implementation of these predictive models on embedded onboard systems.

REFERENCES

[1] M. M. Emily and G. Muyengwa, "Maintenance of Municipality Infrastructure: A Case Study on Service Delivery in Limpopo Province at South Africa," *American Journal of Operations Research*, vol. 11, no. 06, pp. 309–323, 2021, doi: 10.4236/ajor.2021.116019.

[2] T. Zhu, Y. Ran, X. Zhou, and Y. Wen, "A Survey of Predictive Maintenance: Systems, Purposes and Approaches," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.07383

[3] C. Stenström, P. Norrbin, A. Parida, and U. Kumar, "Preventive and corrective maintenance – cost comparison and cost–benefit analysis," *Structure and Infrastructure Engineering*, vol. 12, no. 5, pp. 603–617, May 2016, doi: 10.1080/15732479.2015.1032983.

[4] A. Parida, U. Kumar, D. Galar, and C. Stenström, "Performance measurement and management for maintenance: a literature review," *J. Qual. Maint. Eng.*, vol. 21, no. 1, pp. 2–33, Mar. 2015, doi: 10.1108/JQME-10-2013-0067.

[5] L. Nethamba and S. Grobbelaar, "The development of an action priority matrix and technology roadmap for the implementation of data-driven and machine-learning-based predictive maintenance in the South African railway industry," *South African Journal of Industrial Engineering*, vol. 34, pp. 318–335, 2023.

[6] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry," *Reliab. Eng. Syst. Saf.*, vol. 215, p. 107864, Nov. 2021, doi: 10.1016/J.RESS.2021.107864.

[7] Y. Kovaleva, O. Iliashenko, V. Iliashenko, and R. Anselm, "Implementation of predictive maintenance in a railway logistics company," in *Proceedings of the International Scientific Conference - Digital Transformation on Manufacturing, Infrastructure and Service*, New York, NY, USA: ACM, Nov. 2020, pp. 1–7. doi: 10.1145/3446434.3446507.

[8] M. Mathatlise and E. D. Markus, "Intelligent Predictive Maintenance of Electric Trains in South Africa," in *International Journal of Simulation--Systems, Science & Technology*, Central University of Technology, 2020.

[9] S. Y. Baroud, "Cutting-Edge AI Approaches with MAS for PdM in Industry 4.0: Challenges and Future Directions," *Journal of Applied Data Sciences*, vol. 5, no. 2, pp. 455–473, May 2024, doi: 10.47738/jads.v5i2.196.

[10] H. Zhou, Y. Wan, H. Ye, M. Jiang, and J. Wang, "Real-time energy-efficient optimal control of high-speed electric train," *Control Eng. Pract.*, vol. 112, p. 104825, Jul. 2021, doi: 10.1016/J.CONENGPRAC.2021.104825.

[11] A. Nkomane, "Smart Technologies for Physical Asset Management in the South African Railway Industry," Master's Thesis, University of Johannesburg, 2023.

[12] A. L. Masikati, "Developing a framework for the modernisation of passenger railway maintenance depots in developing countries," Master's Thesis, Stellenbosch University, 2022. [Online]. Available: https://scholar.sun.ac.za

[13] G. G. Samatas, S. S. Moumgiakmas, and G. A. Papakostas, "Predictive Maintenance-Bridging Artificial Intelligence and IoT," Apr. 2021. doi: 10.48550/arXiv.2103.11148.

[14] M. Binder, V. Mezhuyev, and M. Tschandl, "Predictive Maintenance for Railway Domain: A Systematic Literature Review," *IEEE Engineering Management Review*, vol. 51, no. 2, pp. 120–140, Jun. 2023, doi: 10.1109/EMR.2023.3262282.

[15] W. J. C and G. J. Willsie Kathrine, "Intelligent System with the IoT: A survey on techniques of Artificial Intelligence over the field of Internet of Things," in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2022, pp. 347–351. doi: 10.1109/ICACCS54159.2022.9785282.

[16] Oluwafunmi Adijat Elufioye, Chinedu Ugochukwu Ike, Olubusola Odeyemi, Favour Oluwadamilare Usman, and Noluthando Zamanjomane Mhlongo, "AI-DRIVEN PREDICTIVE ANALYTICS IN AGRICULTURAL SUPPLY CHAINS: A REVIEW: ASSESSING THE BENEFITS AND CHALLENGES OF AI IN FORECASTING DEMAND AND OPTIMIZING SUPPLY IN AGRICULTURE," *Computer Science & IT Research Journal*, vol. 5, no. 2, pp. 473–497, Feb. 2024, doi: 10.51594/csitrj.v5i2.817.

[17] I. Daniyan, K. Mpofu, R. Muvunzi, and I. D. Uchegbu, "Implementation of Artificial intelligence for maintenance operation in the rail industry," *Procedia CIRP*, vol. 109, pp. 449–453, Jan. 2022, doi: 10.1016/J.PROCIR.2022.05.277.

[18] Y. K. Al-Douri, P. Tretten, and R. Karim, "Improvement of railway performance: a study of Swedish railway infrastructure," *Journal of Modern Transportation*, vol. 24, no. 1, pp. 22–37, Mar. 2016, doi: 10.1007/s40534-015-0092-0.

[19] B. Tjahjono, C. Esplugues, E. Ares, and G. Pelaez, "What does Industry 4.0 mean to Supply Chain?," *Procedia Manuf.*, vol. 13, pp. 1175–1182, Jan. 2017, doi: 10.1016/J.PROMFG.2017.09.191.

[20] A. Ghadge, M. Er Kara, H. Moradlou, and M. Goswami, "The impact of Industry 4.0 implementation on supply chains," *Journal of Manufacturing Technology Management*, vol. 31, no. 4, pp. 669–686, May 2020, doi: 10.1108/JMTM-10-2019-0368.

[21] A. Santana, P. Afonso, A. Zanin, and R. Wernke, "Costing models for capacity optimization in Industry 4.0: Trade-off between used capacity and operational efficiency," *Procedia Manuf.*, vol. 13, pp. 1183–1190, 2017, doi: 10.1016/j.promfg.2017.09.193.

[22] E. R. Nappi, "Integrated Maintenance: analysis and perspective of innovation in railway sector," 2014. doi: 10.48550/arXiv.1404.7560.

[23] A. Joanni, Q. Mahboob, and E. Zio, "Advanced Methods for RAM Analysis and Decision Making," in *Handbook of RAMS in Railway Systems*, Boca Raton : Taylor & Francis, CRC Press, 2018.: CRC Press, 2018, pp. 39–62. doi: 10.1201/b21983-3.

[24] M. M. Hamasha *et al.*, "Strategical selection of maintenance type under different conditions," *Sci. Rep.*, vol. 13, no. 1, p. 15560, Sep. 2023, doi: 10.1038/s41598-023-42751-5.

[25] Y. Bouabdallaoui, Z. Lafhaj, P. Yim, L. Ducoulombier, and B. Bennadji, "Predictive Maintenance in Building Facilities: A Machine Learning-Based Approach," *Sensors*, vol. 21, no. 4, p. 1044, Feb. 2021, doi: 10.3390/s21041044.

[26] L. Ciocoiu, C. E. Siemieniuch, and E.-M. Hubbard, "From preventative to predictive maintenance: The organisational challenge," *Proc. Inst. Mech. Eng. F J. Rail Rapid Transit*, vol. 231, no. 10, pp. 1174–1185, Nov. 2017, doi: 10.1177/0954409717701785.

[27] E. Ochieng and D. Ominde, "The Role of Operations and Maintenance in Infrastructure Management," in *Routledge Handbook of Planning and Management of Global Strategic Infrastructure Projects*, Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2021.: Routledge, 2020, pp. 221–250. doi: 10.4324/9781003036388-10.

[28] S. Alaswad and Y. Xiang, "A review on condition-based maintenance optimization models for stochastically deteriorating system," *Reliab. Eng. Syst. Saf.*, vol. 157, pp. 54–63, Jan. 2017, doi: 10.1016/J.RESS.2016.08.009.

[29] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," *Comput. Ind. Eng.*, vol. 150, p. 106889, Dec. 2020, doi: 10.1016/J.CIE.2020.106889.

[30] N. Sakib and T. Wuest, "Challenges and Opportunities of Condition-based Predictive Maintenance: A Review," *Procedia CIRP*, vol. 78, pp. 267–272, Jan. 2018, doi: 10.1016/J.PROCIR.2018.08.318.

[31] M. Achouch *et al.*, "On Predictive Maintenance in Industry 4.0: Overview, Models, and Challenges," *Applied Sciences*, vol. 12, no. 16, p. 8081, Aug. 2022, doi: 10.3390/app12168081.

[32] D. Eichler, "Digitalization of Train Maintenance A Mapping of Digital Tools and Technologies to Facilitate Technicians Performing Maintenance," 2024.

[33] R. Ravnå and P. Schjølberg, "Industry 4.0 and Maintenance," First., Norway: Norsk Forening for Vedlikehold (NFV), 2016.

[34] I. Errandonea, S. Beltrán, and S. Arrizabalaga, "Digital Twin for maintenance: A literature review," *Comput. Ind.*, vol. 123, p. 103316, Dec. 2020, doi: 10.1016/j.compind.2020.103316.

[35] L. Q. M. da Costa and C. A. V. Cavalcante, "A REVIEW ON THE STUDY OF MAINTENANCE EFFECTIVENESS," *Pesquisa Operacional*, vol. 42, no. spe1, 2022, doi: 10.1590/0101-7438.2022.042nspe1.00263613.

[36] S. Unger *et al.*, "Securing the Future Railway System: Technology Forecast, Security Measures, and Research Demands," *Vehicles*, vol. 5, no. 4, pp. 1254–1274, Sep. 2023, doi: 10.3390/vehicles5040069.

[37] R. Nappi, G. Cutrera, A. Vigliotti, and G. Franze, "A predictive-based maintenance approach for rolling stocks vehicles," in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, IEEE, Sep. 2020, pp. 793–798. doi: 10.1109/ETFA46521.2020.9212183.

[38] S. Malak, B. Valeria, H. Faicel, and S. Hichem, "Predictive maintenance planning for railway systems," in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, IEEE, Dec. 2023, pp. 553–558. doi: 10.1109/CiSt56084.2023.10409982.

[39] A. O. Olumuyiwa, “A DECISION SUPPORT MODEL TO IMPROVE ROLLING STOCK MAINTENANCE SCHEDULING BASED ON RELIABILITY AND COST,” 2014. [Online]. Available: http://scholar.sun.ac.za

[40] S. B. Singh, R. Suresha, and K. H. Sachidananda, “Reliability Centered Maintenance Used in Metro Railways,” *Journal Européen des Systèmes Automatisés*, vol. 53, no. 1, pp. 11–19, Feb. 2020, doi: 10.18280/jesa.530102.

[41] L. Rojas, Á. Peña, and J. Garcia, “AI-Driven Predictive Maintenance in Mining: A Systematic Literature Review on Fault Detection, Digital Twins, and Intelligent Asset Management,” *Applied Sciences*, vol. 15, no. 6, p. 3337, Mar. 2025, doi: 10.3390/app15063337.

[42] P. Kamat and R. Sugandhi, “Anomaly Detection for Predictive Maintenance in Industry 4.0-A survey,” *E3S Web of Conferences*, vol. 170, p. 02007, May 2020, doi: 10.1051/e3sconf/202017002007.

[43] P. Sukma, A. Sahara, and A. Eunike, “Minimizing Corrective Maintenance Cost through Spare Parts Classification and Inventory Control,” in *Proceedings of the 2019 1st International Conference on Engineering and Management in Industrial System (ICOEMIS 2019)*, Paris, France: Atlantis Press, 2019. doi: 10.2991/icoemis-19.2019.59.

[44] J. A. Erkoyuncu, S. Khan, A. L. Eiroa, N. Butler, K. Rushton, and S. Brocklebank, “Perspectives on trading cost and availability for corrective maintenance at the equipment type level,” *Reliab. Eng. Syst. Saf.*, vol. 168, pp. 53–69, Dec. 2017, doi: 10.1016/J.RESS.2017.05.041.

[45] T. Zhu, Y. Ran, X. Zhou, and Y. Wen, “A Survey of Predictive Maintenance: Systems, Purposes and Approaches,” Dec. 2019.

[46] S. Pashami *et al.*, “Explainable Predictive Maintenance,” 2023.

[47] J. Bokrantz, A. Skoogh, C. Berlin, T. Wuest, and J. Stahre, “Smart Maintenance: an empirically grounded conceptualization,” *Int. J. Prod. Econ.*, vol. 223, May 2020, doi: 10.1016/J.IJPE.2019.107534.

[48] J. Xie, J. Huang, C. Zeng, S.-H. Jiang, and N. Podlich, “Systematic Literature Review on Data-Driven Models for Predictive Maintenance of Railway Track: Implications in Geotechnical Engineering,” *Geosciences (Basel).*, vol. 10, no. 11, p. 425, Oct. 2020, doi: 10.3390/geosciences10110425.

[49] G. Kefalidou, D. Golightly, and S. Sharples, "Identifying rail asset maintenance processes: a human-centric and sensemaking approach," *Cognition, Technology & Work*, vol. 20, no. 1, pp. 73–92, Feb. 2018, doi: 10.1007/s10111-017-0452-0.

[50] F. Dinmohammadi, B. Alkali, M. Shafiee, C. Bérenguer, and A. Labib, "Risk Evaluation of Railway Rolling Stock Failures Using FMECA Technique: A Case Study of Passenger Door System," *Urban Rail Transit*, vol. 2, no. 3–4, pp. 128–145, Dec. 2016, doi: 10.1007/s40864-016-0043-z.

[51] A. Erguido, A. C. Márquez, E. Castellano, J. L. Flores, and J. F. G. Fernández, "Reliability-based advanced maintenance modelling to enhance rolling stock manufacturers' objectives," *Comput. Ind. Eng.*, vol. 144, p. 106436, Jun. 2020, doi: 10.1016/J.CIE.2020.106436.

[52] X. Chen, S. Wang, B. Qiao, and Q. Chen, "Basic research on machinery fault diagnostics: Past, present, and future trends," *Frontiers of Mechanical Engineering*, vol. 13, no. 2, pp. 264–291, Jun. 2018, doi: 10.1007/s11465-018-0472-3.

[53] A. Kane, A. Kore, A. Khandale, S. Nigade, and P. P. Joshi, "Predictive Maintenance using Machine Learning," 2022. doi: 10.48550/arXiv.2205.09402.

[54] Z. M. Çınar, A. Abdussalam Nuhu, Q. Zeeshan, O. Korhan, M. Asmael, and B. Safaei, "Machine Learning in Predictive Maintenance towards Sustainable Smart Manufacturing in Industry 4.0," *Sustainability*, vol. 12, no. 19, p. 8211, Oct. 2020, doi: 10.3390/su12198211.

[55] N. T. Trần, H. T. Triệu, V. T. Trần, H. H. Ngô, and Q. K. Đào, "An overview of the application of machine learning in predictive maintenance," *Petrovietnam Journal*, vol. 10, pp. 47–61, Nov. 2021, doi: 10.47800/PVJ.2021.10-05.

[56] V. Shiboldenkov and K. Nesterova, "The smart technologies application for the product life-cycle management in modern manufacturing systems," *MATEC Web of Conferences*, vol. 311, p. 02020, Mar. 2020, doi: 10.1051/matecconf/202031102020.

[57] Nzubechukwu Chukwudum Ohalete, Adebayo Olusegun Aderibigbe, Emmanuel Chigozie Ani, Peter Efosa Ohenhen, and Abiodun Akinoso, "DATA SCIENCE IN ENERGY CONSUMPTION ANALYSIS: A REVIEW OF AI TECHNIQUES IN IDENTIFYING PATTERNS AND EFFICIENCY OPPORTUNITIES," *Engineering Science & Technology Journal*, vol. 4, no. 6, pp. 357–380, Dec. 2023, doi: 10.51594/estj.v4i6.637.

[58] Eastern Eye, “A Complete History Of Predictive Maintainence & Its Place In The World Today,” Eastern Eye. Accessed: Jun. 23, 2024. [Online]. Available: https://www.easterneye.biz/a-complete-history-of-predictive-maintainence-its-place-in-the-world-today/

[59] Otto Sean, “Predictive Maintenance and its Role in Improving Efficiency,” *https://www.pcimag.com/*, May 06, 2019. Accessed: Jun. 18, 2025. [Online]. Available: https://www.pcimag.com/articles/106046-predictive-maintenance-and-its-role-in-improving-efficiency

[60] T. Marwala, *Closing the gap : the fourth industrial revolution in Africa*. Johannesburg, South Africa: Macmillan, 2020.

[61] E. Sutherland, “The Fourth Industrial Revolution – The Case of South Africa,” *Politikon*, vol. 47, no. 2, pp. 233–252, Apr. 2020, doi: 10.1080/02589346.2019.1696003.

[62] C. Laiton-Bonadiez, J. W. Branch-Bedoya, J. Zapata-Cortes, E. Paipa-Sanabria, and M. Arango-Serna, “Industry 4.0 Technologies Applied to the Rail Transportation Industry: A Systematic Review,” *Sensors*, vol. 22, no. 7, p. 2491, Mar. 2022, doi: 10.3390/s22072491.

[63] Y. Qian, M. S. Dersch, Z. Gao, and J. R. Edwards, “Railroad infrastructure 4.0: Development and application of an automatic ballast support condition assessment system,” *Transportation Geotechnics*, vol. 19, pp. 19–34, Jun. 2019, doi: 10.1016/J.TRGEO.2019.01.002.

[64] M. Kans, D. Galar, and A. Thaduri, “Maintenance 4.0 in Railway Transportation Industry,” 2016, pp. 317–331. doi: 10.1007/978-3-319-27064-7_30.

[65] Y. Kovaleva, O. Iliashenko, V. Iliashenko, and R. Anselm, “Implementation of predictive maintenance in a railway logistics company,” in *Proceedings of the International Scientific Conference - Digital Transformation on Manufacturing, Infrastructure and Service*, New York, NY, USA: ACM, Nov. 2020, pp. 1–7. doi: 10.1145/3446434.3446507.

[66] S. A. Luqman and L. H. Kusumah, “Improving the Effectiveness of Rolling Stock Maintenance: A Systematic Literature Review,” IEOM Society International, May 2023, pp. 1977–1985. doi: 10.46254/ap03.20220335.

[67] J. Brant and B. Liang, “Condition Monitoring Systems in the Railway Industry,” 2020, pp. 671–683. doi: 10.1007/978-3-030-57745-2_56.

[68] LeewayHertz, “AI-driven predictive maintenance: A game-changer for operational reliability,” https://leewayhertz.medium.com/. Accessed: Jun. 21, 2024. [Online]. Available: https://leewayhertz.medium.com/ai-driven-predictive-maintenance-a-game-changer-for-operational-reliability-f8a30b9e58ad

[69] M. Raparthi, S. B. Dodda, and S. Maruthi, “Predictive Maintenance in IoT Devices using Time Series Analysis and Deep Learning,” *Dandao Xuebao/Journal of Ballistics*, vol. 35, no. 3, pp. 01–10, Dec. 2023, doi: 10.52783/dxjb.v35.113.

[70] G. Bianchi, C. Fanelli, F. Freddi, F. Giuliani, and A. La Placa, “Systematic review railway infrastructure monitoring: From classic techniques to predictive maintenance,” *Advances in Mechanical Engineering*, vol. 17, no. 1, Jan. 2025, doi: 10.1177/16878132241285631.

[71] N. Y. Mulongo, E. Mnkandla, and G. Kanakana-Katumba, “Artificial Intelligence as Key Driver for Competitiveness in the Railway Industry,” in *2021 62nd International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*, IEEE, Oct. 2021, pp. 1–6.

[72] L. Gong, A. Sharma, M. Abdul Bhuiya, H. Awad, and M. Z. Youssef, “An Artificial Intelligence Based Fault Monitoring of Power Trains: Design & Implementation,” in *2021 IEEE Applied Power Electronics Conference and Exposition (APEC)*, IEEE, Jun. 2021, pp. 2761–2768. doi: 10.1109/APEC42165.2021.9487452.

[73] S. Jeschke and R. Grassmann, “Development of a Generic Implementation Strategy of Digital Twins in Logistics Systems under Consideration of the German Rail Transport,” *Applied Sciences*, vol. 11, no. 21, p. 10289, Nov. 2021, doi: 10.3390/app112110289.

[74] C. Zhang, X. Xie, and X. Guo, “Scheme Design of Railway Predictive Maintenance Based on IOT and AI Technology,” 2021. doi: 10.2991/assehr.k.211215.062.

[75] R. Wang, M. Kudrot-E-Khuda, F. Nakamura, and S. Tanaka, “A Cost-benefit Analysis of Commuter Train Improvement in the Dhaka Metropolitan Area, Bangladesh,” *Procedia Soc. Behav. Sci.*, vol. 138, pp. 819–829, Jul. 2014, doi: 10.1016/J.SBSPRO.2014.07.231.

[76] P. Ficzere, “The role of artificial intelligence in the development of railway transportation,” *Design of Machines and Structures*, vol. 13, no. 1, pp. 67–73, Jun. 2023, doi: 10.32972/dms.2023.005.

[77] V. Vatakov, E. Pencheva, and E. Dimitrova, “Recent advances in artificial intelligence for improving railway operations,” in *30th National Conference with International Participation (TELECOM)*, IEEE, Oct. 2022, pp. 1–4.

[78] Psiborg, “Predictive Maintenance Using IoT,” https://psiborg.in/predictive-maintenance-using-iot/.

[79] A. Consilvio, A. Di Febbraro, and N. Sacco, “A Rolling-Horizon Approach for Predictive Maintenance Planning to Reduce the Risk of Rail Service Disruptions,” *IEEE Trans. Reliab.*, vol. 70, no. 3, pp. 875–886, Sep. 2020, doi: 10.1109/TR.2020.3007504.

[80] B. El Mahdi, E. K. Ali, E. K. Youssra, and E. Soufiane, “Real Time Assessment of Novel Predictive Maintenance System Based on Artificial Intelligence for Rotating Machines,” *Journal Europeen des Systemes Automatises*, vol. 55, no. 6, pp. 817–823, Dec. 2022, doi: 10.18280/jesa.550614.

[81] M. Fernandes, A. Canito, J. M. Corchado, and G. Marreiros, “Fault Detection Mechanism of a Predictive Maintenance System Based on Autoregressive Integrated Moving Average Models,” 2019, pp. 171–180. doi: 10.1007/978-3-030-23887-2_20.

[82] A. D. Toth, J. Padayachee, T. Mahlatji, and S. Vilakazi, “Report on case studies of additive manufacturing in the South African railway industry,” *Sci. Afr.*, vol. 16, Jul. 2022, doi: 10.1016/j.sciaf.2022.e01219.

[83] I. Daniyan, K. Mpofu, M. Oyesola, B. Ramatsetse, and A. Adeodu, “Artificial intelligence for predictive maintenance in the railcar learning factories,” *Procedia Manuf.*, vol. 45, pp. 13–18, Jan. 2020, doi: 10.1016/J.PROMFG.2020.04.032.

[84] J. L. R. Kloppers, “A Roadmap to Smart Maintenance for Inner-City Public Bus Services in South Africa,” Doctoral dissertation, Stellenbosch University, 2024.

[85] A. Nqodi, T. C. Mosetlhe, and A. A. Yusuff, “Advances in Hydrogen-Powered Trains: A Brief Report,” *Energies (Basel).*, vol. 16, no. 18, Sep. 2023, doi: 10.3390/en16186715.

[86] Z. Liu, Y. Gao, and B. Liu, “An artificial intelligence-based electric multiple units using a smart power grid system,” *Energy Reports*, vol. 8, pp. 13376–13388, Nov. 2022, doi: 10.1016/j.egyr.2022.09.138.

[87] S. Naidoo, D. Bottomley, M. Naidoo, D. Donnelly, and D. W. Thaldar, “Artificial intelligence in healthcare: Proposals for policy development in South Africa,” *S. Afr. J. Bioeth. Law*, pp. 11–16, May 2022, doi: 10.7196/SAJBL.2022.v15i1.797.

[88] N. T. Nhede, A. Mazenda, and T. Masiya, "The South African public service and the Fourth Industrial Revolution," *Africa's Public Service Delivery and Performance Review*, vol. 10, no. 1, May 2022, doi: 10.4102/apsdpr.v10i1.420.

[89] L. Bukvić, J. Pašagić Škrinjar, T. Fratrović, and B. Abramović, "Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning," *Sustainability*, vol. 14, no. 24, p. 17034, Dec. 2022, doi: 10.3390/su142417034.

[90] K. K. Paidipati, C. Chesneau, B. M. Nayana, K. R. Kumar, K. Polisetty, and C. Kurangi, "Prediction of Rice Cultivation in India—Support Vector Regression Approach with Various Kernels for Non-Linear Patterns," *AgriEngineering*, vol. 3, no. 2, pp. 182–198, Apr. 2021, doi: 10.3390/agriengineering3020012.

[91] Y. Song, F. Duan, F. Wu, Z. Liu, and S. Gao, "Assessment of the Current Collection Quality of Pantograph–Catenary With Contact Line Height Variability in Electric Railways," *IEEE Transactions on Transportation Electrification*, vol. 8, no. 1, pp. 788–798, Mar. 2022, doi: 10.1109/TTE.2021.3090477.

[92] J. Pombo and J. Ambrósio, "Influence of pantograph suspension characteristics on the contact quality with the catenary for high speed trains," *Comput. Struct.*, vol. 110–111, pp. 32–42, Nov. 2012, doi: 10.1016/j.compstruc.2012.06.005.

[93] A. Golubkov, G. Ermachkov, A. Smerdin, O. Sidorov, and V. Philippov, "Increasing the forecasting fidelity of current collection system operating capability by means of contact pressure simulation modelling," *E3S Web of Conferences*, vol. 217, p. 03003, Dec. 2020, doi: 10.1051/e3sconf/202021703003.

[94] A. Mariscotti, D. Giordano, A. D. Femine, D. Gallo, and D. Signorino, "How Pantograph Electric Arcs affect Energy Efficiency in DC Railway Vehicles," in *2020 IEEE Vehicle Power and Propulsion Conference (VPPC)*, IEEE, Nov. 2020, pp. 1–5. doi: 10.1109/VPPC49601.2020.9330954.

[95] A. Mariscotti and D. Giordano, "Experimental Characterization of Pantograph Arcs and Transient Conducted Phenomena in DC Railways," *ACTA IMEKO*, vol. 9, no. 2, p. 10, Jun. 2020, doi: 10.21014/acta_imeko.v9i2.761.

[96] D. Creates, "Plotting Scikit-Learn Classification Report for Analysis — Using MatplotLib.PyPlot-SkLearn.Metrics-Seaborn-Pandas-NumPy," Medium.com.

APPENDIX A: DATA VISUALISATION SAMPLE CODE

```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np  # Added numpy import for np.inf


sns.set(style="whitegrid")
# dataset
df = pd.read_csv("PRASA_Updated_EMU_Eventlog_2023_2024.csv")
df['Timestamp'] = pd.to_datetime(df['Timestamp'])

df['Month'] = df['Timestamp'].dt.to_period('M').astype(str)

trend_sensors = [
    ("Pantograph", "Pressure"),
    ("Pantograph", "Arc_Frequency"),
    ("Traction Motor", "Temperature"),
    ("Traction Motor", "Current"),
    ("Traction Motor", "Vibration"),
    ("ACU", "Current"),
    ("ACU", "Temperature"),
    ("ACU", "Air_Flow"),
    ("Line", "Voltage"),
    ("Line", "Current")
]
for component, sensor in trend_sensors:
    subset = df[(df["Component"] == component) & (df["Sensor_Type"] == sensor)]
    if subset.empty:
        continue

    df.isin([np.inf, -np.inf]).any()

df["Sensor_Value"] = pd.to_numeric(df["Sensor_Value"], errors="coerce")

desc_stats = df.groupby("Component")["Sensor_Value"].describe()

# Displaying the stats
print(desc_stats)
```

```
                   count         mean          std        min         25%  \
Component
ACU            1052640.0   122.023586   131.758176   0.254077    1.334730
Line            701760.0  1650.123524  1354.751564  73.955378  300.059538
Pantograph      701760.0     5.249021     3.116368   0.000010    2.076355
Traction Motor 1052640.0   122.316889   131.542627   0.660319    2.202101
```

```
df_subset = df.pivot_table(
    index="Timestamp",
    columns="Sensor_Type",
    values="Sensor_Value",
    aggfunc="mean"
).reset_index()
if "Voltage" in df_subset.columns and "Arc_Frequency" in df_subset.columns:
    plt.figure(figsize=(14,6))

    ax1 = plt.gca()
    ax1.plot(df_subset['Timestamp'], df_subset['Voltage'], color='blue', label="Line Voltage (V)")
    ax1.set_ylabel("Line Voltage (V)", color="blue")
    ax1.tick_params(axis="y", labelcolor="blue")

    ax2 = ax1.twinx()
    ax2.plot(df_subset['Timestamp'], df_subset['Arc_Frequency'], color='red', alpha=0.6, label="Arc Frequency (Hz)")
    ax2.set_ylabel("Arc Frequency (Hz)", color="red")
    ax2.tick_params(axis="y", labelcolor="red")

    # Formatting
    plt.title("Line Voltage vs Arc Frequency Over Time")
    ax1.set_xlabel("Time")
    plt.grid(True)
    plt.tight_layout()
    plt.show()
else:
    print("Required columns not found in dataset")
```

```
                         columns='Sensor_Type',
                         values='Sensor_Value',
                         aggfunc='mean').sort_index()

# Failure flag per timestamp for TM
    fail = d.groupby('Timestamp')['Failure'].max().reindex(wide.index).fillna(0).astype(int)
    fmask = fail == 1

# Left axis: Vibration
    if 'Vibration' in wide.columns:
        ax1.plot(wide.index, wide['Vibration'], label='Vibration (g)', linewidth=1.2)
        ax1.scatter(wide.index[fmask], wide.loc[fmask, 'Vibration'],
                    marker='x', s=25, color='black', label='Failure (on Vibration)')
        ax1.set_ylabel("Vibration (g)")
    else:
        ax1.text(0.02, 0.9, "No Vibration data", transform=ax1.transAxes, color='tab:gray')

# Right axis: Temperature
    if 'Temperature' in wide.columns:
        ax2.plot(wide.index, wide['Temperature'], color='tab:orange', label='Temperature (°C)', linewidth=1.2)
        ax2.scatter(wide.index[fmask], wide.loc[fmask, 'Temperature'],
                    marker='o', s=20, color='black', facecolors='none', label='Failure (on Temp)')
        ax2.set_ylabel("Temperature (°C)")
    else:
        ax2.text(0.98, 0.9, "No Temperature data", transform=ax2.transAxes, ha='right', color='tab:gray')

    ax1.set_title(f"Traction Motor – {train})")
    ax1.set_xlabel("Time")
    ax1.grid(True, which='both', alpha=0.3)

    lines1, labels1 = ax1.get_legend_handles_labels()
    lines2, labels2 = ax2.get_legend_handles_labels()
    seen, lines, labels = set(), [], []
    for L, lab in list(zip(lines1+lines2, labels1+labels2)):
        if lab not in seen:
            lines.append(L); labels.append(lab); seen.add(lab)
    ax1.legend(lines, labels, loc='upper left', fontsize=9)

# Hiding unused subplots when < 6 trains available
for j in range(i+1, rows*cols):
    axes[j].axis('off')

fig.suptitle("Traction Motor Temperature & Vibration Analysis\nFebruary 2023", fontsize=14)
plt.tight_layout(rect=[0, 0, 1, 0.96])
plt.show()
```

```python
    fail_series = d.groupby('Timestamp')['Failure'].max().sort_index().astype(int)
    fault_times = fail_series.index[fail_series == 1]
    if len(fault_times) == 0:
        return

    # --- RANDOM faults picking
    n_pick = min(faults_per_param, len(fault_times))
    fault_times = pd.to_datetime(np.random.choice(fault_times, size=n_pick, replace=False))


    fig, axes = plt.subplots(len(fault_times), 1, figsize=(14, 4.4*len(fault_times)), sharex=False)
    if not isinstance(axes, np.ndarray):
        axes = [axes]

    thresh = THRESHOLDS.get(param_name, None)

    for ax, ft in zip(axes, sorted(fault_times)):
        start_win = ft - pd.Timedelta(days=WINDOW_DAYS)
        end_win = ft

        s = d[(d['Timestamp'] >= start_win) & (d['Timestamp'] <= end_win)].copy()
        if s.empty:
            ax.set_title(f"{train_id} - {param_name}: No data in 7 days before {ft}")
            ax.grid(True, alpha=0.3)
            continue

        s = s.set_index('Timestamp')
        s_hour = s.resample('0.25h').agg({'Sensor_Value': 'mean', 'Failure': 'max'}).reset_index()

        # Mean for trend
        s_hour['roll'] = s_hour.set_index('Timestamp')['Sensor_Value'].rolling(ROLL_WIN, min_periods=1).mean().values

        ax.plot(s_hour['Timestamp'], s_hour['Sensor_Value'], linewidth=1.0, label=f'{param_name} (raw)', zorder=1)
        ax.plot(s_hour['Timestamp'], s_hour['roll'], linewidth=1.3, linestyle='--', label=f'{param_name} (rolling)', zorder=2)

        if thresh is not None:
            ax.axhline(thresh, color='tab:purple', linestyle=':', linewidth=1.5, label=f'Threshold = {thresh}', zorder=3)
            if param_name == 'Temperature':
                ax.fill_between(s_hour['Timestamp'], thresh, s_hour['Sensor_Value'].max()*1.05,
                                color='tab:purple', alpha=0.08, step=None, zorder=0)
            elif param_name == 'Vibration':
                ax.fill_between(s_hour['Timestamp'], thresh, s_hour['Sensor_Value'].max()*1.05,
                                color='tab:purple', alpha=0.08, step=None, zorder=0)
```

```python
# Counting failures per month per component
monthly_counts = (
    df_fail
    .groupby(['Month', 'Component'])['Failure']
    .sum()
    .unstack('Component')
    .fillna(0)
    .astype(int)
    .sort_index()
)
```

APPENDIX B: MODELLING AND PREDICTION SAMPLE CODE

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import (classification_report, confusion_matrix, roc_auc_score, roc_curve)

# Loading the dataset
df = pd.read_csv("PRASA_UpdatedV2_EMU_Eventlog_2023_2024.csv", parse_dates=['Timestamp'])

# Step 1: Handlling Missing Data
# ----------------------------
df.interpolate(method='linear', inplace=True) # Replace or interpolate missing sensor values
df.fillna(method='bfill', inplace=True)

# Step 2: Managing outliers
# ----------------------------------------------------------
df['Sensor_Value'] = np.clip(df['Sensor_Value'], df['Sensor_Value'].quantile(0.01), df['Sensor_Value'].quantile(0.99))

# Step 3: Feature Engineering
# ----------------------------
# Extracting time-based features
df['Month'] = df['Timestamp'].dt.month
df['Hour'] = df['Timestamp'].dt.hour
df['DayOfWeek'] = df['Timestamp'].dt.dayofweek

# Creating encoded features for categorical variables
categorical_features = ['Component', 'Sensor_Type', 'Sample_Type']
numeric_features = ['Sensor_Value', 'Month', 'Hour', 'DayOfWeek']

# Step 4: Separating Features and Labels
# -------------------------------------
X = df[categorical_features + numeric_features]
y = df['Failure']

# Step 5: Defining Preprocessing Process
# ---------------------------------------
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline(steps=[
```

```
categorical_transformer = Pipeline(steps=[
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)
    ])

# Step 6: Data partitioning
# --------------------------
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=42)
X_train_scaled = preprocessor.fit_transform(X_train)
X_test_scaled = preprocessor.transform(X_test)
#---------------------------------End of Data Preprocessing--------------------------------------------------


# LOGIC REGRESSION PREDICTIVE APPLICATION

# Modelling: Logistic Regression (L2) with class balancing
log_reg = LogisticRegression(
    penalty='l2',
    class_weight='balanced',
    solver='lbfgs',
    max_iter=2000,
    n_jobs=None
)

clf = Pipeline(steps=[
    ('prep', preprocessor),
    ('model', log_reg)
])

clf.fit(X_train, y_train) #Training

y_pred = clf.predict(X_test) # Predicting
y_proba = clf.predict_proba(X_test)[:, 1]  # probability of Failure=1

report = classification_report(y_test, y_pred, digits=3, output_dict=True)
```

```python
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())
])

categorical_transformer = Pipeline(steps=[
    ('ohe', OneHotEncoder(handle_unknown='ignore', sparse_output=False))
])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)
    ],
    remainder='drop'
)
X_train, X_test, y_train, y_test = train_test_split(  # Train/Test sets split
    X, y, test_size=0.20, random_state=42, stratify=y
)
# Random Forest Modeling
base_rf = RandomForestClassifier(
    n_estimators=200,
    max_depth=10,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)
pipeline = Pipeline(steps=[
    ('prep', preprocessor),
    ('model', base_rf)
])
param_grid = {
    'model__n_estimators': [200, 300, 400], #Hyperparameter grid (depth, trees, splits/leaves, max_features)
    'model__max_depth': [8, 10, 12, None],
    'model__min_samples_split': [2, 5, 10],
    'model__min_samples_leaf': [1, 2, 4],
    'model__max_features': ['sqrt', 'log2', 0.5]
}
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42) #GridSearchCV with 5-fold
grid = GridSearchCV(
    estimator=pipeline,
    param_grid=param_grid,
    scoring='f1',
    cv=cv,
    n_jobs=-1,
    verbose=1
)
```

APPENDIX C: EDITING CERTIFICATE

NERESHNEE GOVENDER COMMUNICATIONS (PTY) LTD

REGISTRATION NUMBER: 2016/369223/07

DR NERESHNEE GOVENDER (PhD)

neresh@ngcommunications.co.za

0847022553

WRITING PRACTITIONER • EDITOR • COPYWRITER • TRAINER

PhD-Management Sciences: Marketing (gender and media); *PG DIP - Higher Education - Academic Developers (Cum laude;)* M-Tech Public Relations; B-Tech Public Relations *(Cum laude)*; B-Tech Journalism *(Cum laude)*; *N-Dip Journalism*

10/12/2025

MAVHUNGU SONIA MATHALISE
Supervisor: Prof. ED Markus
Co-supervisor: Prof. M Sibiya
Central University of Technology
Free State

RE: EDITING CERTIFICATE

TITLE: USE OF ARTIFICIAL INTELLIGENCE FOR PREDICTIVE MAINTENANCE OF ELECTRIC TRAINS IN SOUTH AFRICA

Submitted in fulfilment of the requirements for the degree: MASTER OF ENGINEERING IN ELECTRICAL in the Department of Electrical, Electronic and Computer Engineering Faculty of Engineering, Built Environment and Information Technology Central University of Technology, Free State

This serves to confirm that this Masters thesis has been edited for clarity, language and layout.

Kind regards,

Nereshnee Govender (PhD)