Mälardalen University
School of Innovation Design and Engineering
Västerås, Sweden

Thesis for the Degree of Bachelor of Science in Computer Science
15.0 credits

PREDICTIVE MAINTENANCE FOR FREIGHT TRAINS: A MACHINE LEARNING APPROACH TO OPTIMIZING WHEEL REPLACEMENT SCHEDULES

Max Strang
MaxStraang@outlook.com

Examiner: Mobyen Uddin Ahmed
Mälardalen University, Västerås, Sweden

Supervisor: Shaibal Barua
Mälardalen University, Västerås, Sweden

Company Supervisor: Tim Janke
Green Cargo, Stockholm, Sweden

02/07/2025

Abstract

Efficient maintenance planning is crucial for railway operators to minimize costs and operational disruptions. Green Cargo, a major rail freight provider, currently replaces train wheels based on scheduled maintenance or reactive repairs, leading to inefficiencies. This thesis explores the application of machine learning to predictive maintenance by analyzing sensor data from DPC detectors mounted along railway tracks. These detectors measure force impacts as the train moves, providing key indicators of wheel wear.

Using supervised learning techniques and grid search optimization, predictive models for left and right wheel damage were developed and evaluated. The left wheel model achieved a test R_2 of 1.0000 with an RMSE of 0.08 kilonewton, while the right wheel model achieved a test R_2 of 0.9873 with an RMSE of 0.36 kilonewton. Forecasting trends over a three-year horizon identified one axle requiring replacement, with left wheels typically failing first. Maintenance scheduling logic was implemented to replace both wheels on an axle when either exceeded predefined thresholds.

This research demonstrates the potential of predictive maintenance to enhance efficiency and reduce costs in railway operations by leveraging advanced machine learning methods and historical data from DPC detectors. Future work will focus on integrating real-time monitoring and refining predictive capabilities for broader industrial applications.

Contents

Page 6

List of Figures

Page 7

List of Tables

1. Introduction

Efficient railway maintenance plays a pivotal role in ensuring the reliability and safety of freight transport systems. In particular, optimizing the maintenance of critical components, such as train wheels, is essential for minimizing operational disruptions, reducing costs, and enhancing safety. However, one of the key challenges in railway operations is determining the optimal time to replace train wheels, a decision that has traditionally relied on fixed schedules or reactive measures. This often leads to either early replacements, which increase operational costs, or delayed replacements, which could lead to potential failures and safety risks [1].

Green Cargo, one of Sweden's leading rail freight companies, operates a large and diverse fleet of trains across various routes with different operational conditions and braking systems. These variations lead to differences in wear patterns, which complicate the decision-making process regarding wheel replacement. The traditional approach of using fixed maintenance schedules does not account for these variations, presenting an opportunity for more adaptive and data-driven maintenance strategies.

Currently, Green Cargo employs a threshold-based monitoring system for wheel maintenance using sensor data from Dynamic Pressure Check (DPC) detectors installed by Trafikverket [2]. This system operates by setting predefined thresholds for specific values generated by the DPC system. When a value exceeds a certain limit, an alarm is triggered, prompting the maintenance team to schedule maintenance. While this system successfully prevents catastrophic failures, it operates purely reactively, providing alerts only after predefined thresholds have been exceeded. This reactive nature means that Green Cargo has no advance warning of impending maintenance needs, limiting their ability to optimize maintenance scheduling and resource allocation.

The limitations of threshold-based approaches become particularly apparent in operational planning and cost management. The inability to predict when wheels will require maintenance means that all maintenance actions are triggered reactively, often resulting in expensive emergency interventions. These unplanned maintenance events are significantly more costly than scheduled activities, as they involve potential service delays, overtime labor, and urgent resource allocation. Furthermore, the current approach provides no insights into wheel degradation patterns or trends that might help predict future maintenance needs, representing a missed opportunity for proactive maintenance planning.

This thesis addresses these limitations by exploring the application of machine learning techniques to predict wheel replacement needs before they become critical. By analyzing historical sensor data from DPC detectors, this study aims to develop predictive models that can forecast when wheel maintenance will be required well in advance of threshold exceedance. Such an approach could enable Green Cargo to transition from reactive to truly predictive maintenance [3], allowing for better resource planning, cost optimization, and operational efficiency.

This study focuses specifically on three Transmontana series locomotives (M4004, M4008, M4016) operating on the Borlänge-Boden route. This route was selected due to its unique operational characteristics, particularly the use of metal-against-metal brake systems that create more pronounced wheel wear patterns compared to modern brake technologies used on other routes. This increased wear provides richer data signals for machine learning analysis, making this route an ideal test case for developing predictive maintenance models. Although the study examines only three locomotives, these units are representative of a broader population of 20 similar locomotives operating on the same route with identical brake systems and operational characteristics. The insights gained from this analysis can potentially be generalized to the entire fleet of similar locomotives and their associated wagons, providing a foundation for broader predictive maintenance implementation across Green Cargo's operations.

Previous research has demonstrated the effectiveness of predictive maintenance in various industries [4], including specific applications to railway systems [5], [6]. However, most existing studies focus on generalized predictive models without considering the specific challenges posed by the unique operational conditions of rail freight, such as varying braking systems and loading conditions. Additionally, while threshold-based monitoring is common in the industry, the transition to truly predictive approaches using advanced machine learning techniques remains underexplored in operational freight railway applications.

The primary objective of this thesis is to develop and evaluate machine learning models capable

of predicting optimal wheel replacement timing using sensor data from DPC detectors. Specifically, this study investigates the application of XGBoost (eXtreme Gradient Boosting) [7] and linear regression [8] for predictive modeling as well as Isolation Forest [9] for anomaly detection to identify patterns in force measurement data that can indicate the need for wheel maintenance. While real–time validation of predictions has not been conducted due to time constraints, the developed models was evaluated on historical data, allowing for a robust assessment of their predictive performance.

This work is of significant relevance as it could contribute to improving maintenance practices in rail freight operations, potentially reducing costs, minimizing unplanned downtime, and enhancing the safety and efficiency of rail transport. By providing a detailed evaluation of predictive maintenance models tailored to the specific needs of Green Cargo, this research offers valuable insights that can be applied to other operators facing similar challenges in maintaining critical train components. The methodology developed in this study could also serve as a foundation for expanding predictive maintenance applications to other railway components beyond wheels.

The remainder of this thesis is structured as follows: Chapter 2 provides background information on predictive maintenance, locomotive systems, and DPC detector technologies. Chapter 3 reviews related work in railway predictive maintenance and machine learning applications. Chapter 4 formulates the research problem and specific objectives. Chapter 5 describes the methodology used for data collection, preprocessing, and model development. Chapter 6 discusses ethical and societal considerations. Chapter 7 presents the experimental setup and implementation details. Chapter 8 reports the results of the predictive models' performance. Chapter 9 discusses the implications and limitations of the findings. Finally, Chapter 10 provides conclusions and suggestions for future research directions.

2. Background

2.1 Predictive Maintenance

Predictive Maintenance (PdM), described in Mobley [3], is a maintenance strategy that uses data-driven techniques to anticipate equipment failures, allowing maintenance to be scheduled proactively before breakdowns occur. Unlike traditional corrective maintenance (performed after a failure) or preventive maintenance (performed on fixed schedules), PdM uses real-time condition monitoring to guide decisions, minimizing unnecessary interventions and costly disruptions.

2.2 Locomotive Overview

The locomotives studied in this project are the Transmontana series, manufactured by Softronic and referred to as "MB" in the Green Cargo fleet (see Figure 1). These locomotives are six-axle electric locomotives designed for heavy freight operations.
Transmontana locomotives are equipped with modern electric traction systems that provide efficient power delivery and robust hauling capacity. However, their braking systems use a "metal-against-metal" mechanism, which causes significant wear and tear on the wheels. This characteristic makes these locomotives particularly suitable for studying predictive maintenance as it generates substantial wear and tear data, offering a rich dataset for modeling and analysis. The pronounced wear provides clear signals of degradation, making it easier to identify patterns and evaluate the effectiveness of predictive maintenance strategies.

Figure 1: Mb locomotive operated by Green Cargo. Image credited to Green Cargo, sourced from Järnväg.net [10]

2.3 Dynamic Pressure Check (DPC) Detectors

Trafikverket, the Swedish Transport Administration, has deployed various types of Dynamic Pressure Check (DPC) detectors across the railway network to monitor the condition of train components, including wheels (see Figure 2 and 3).

Figure 2 shows a map of the detector locations across Sweden. Each detector type is marked

with a unique color, making it possible to distinguish between them. The PHOENIX MDS WILD detectors, which are marked in blue and the Schenck detectors marked in green, are the types investigated in this study. While the map covers the entire country, it is possible to visually identify the detectors located along the Borlänge – Boden route, the railway section chosen for this study. The route was selected for two primary reasons. Firstly, it offers access to data from both PHOENIX and Schenck detectors, enabling a comprehensive analysis. Secondly, the locomotives operating on this route use older brake systems, as described in Section 2.2. These brake systems involve significant metal-on-metal contact during braking, contributing to increased wear and tear, which makes this route particularly relevant for studying predictive maintenance.

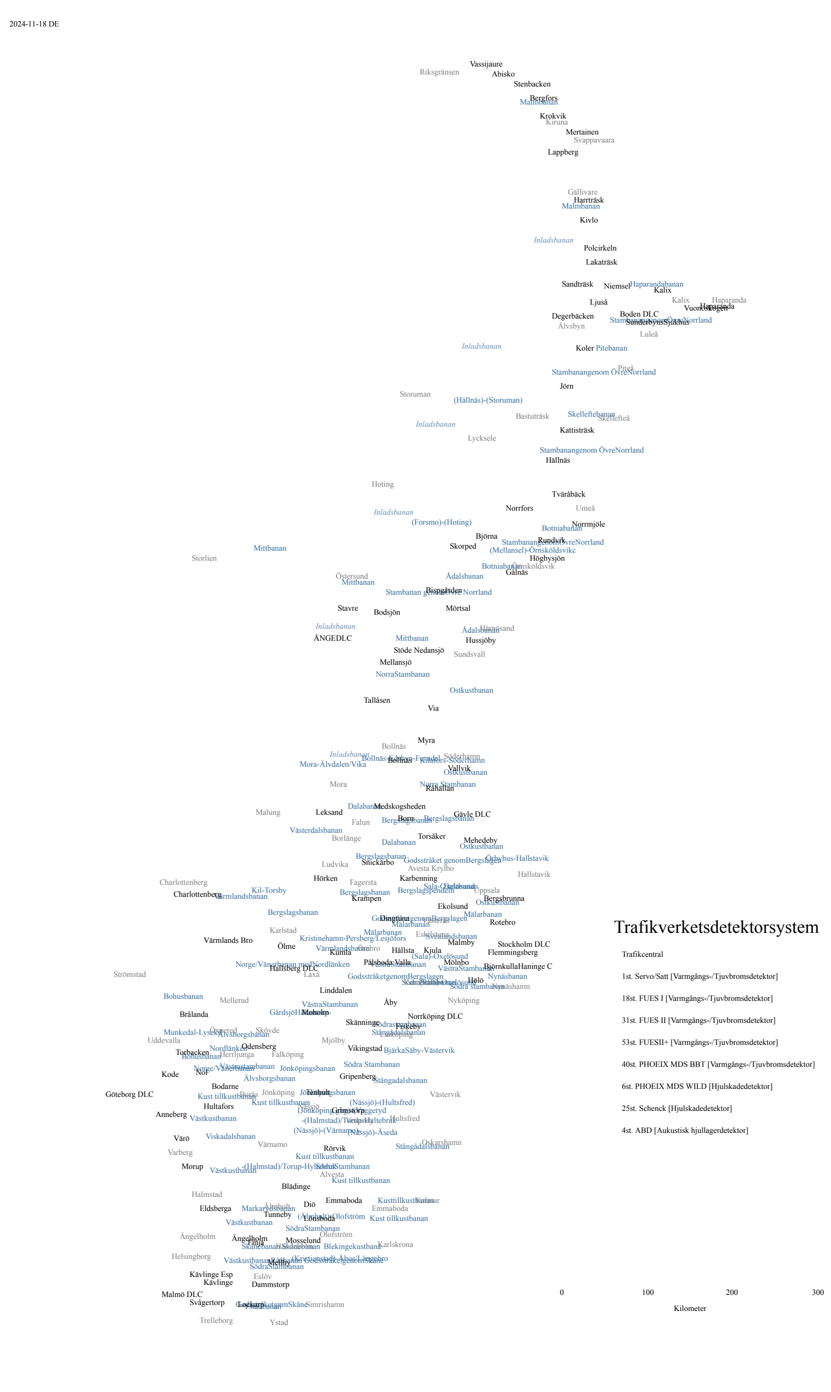
Trafikverketsdetektorsystem
Trafikcentral
1st. Servo/Satt [Varmgångs-/Tjuvbromsdetektor]
18st. FUES I [Varmgångs-/Tjuvbromsdetektor]
31st. FUES II [Varmgångs-/Tjuvbromsdetektor]
53st. FUESII+ [Varmgångs-/Tjuvbromsdetektor]
40st. PHOEIX MDS BBT [Varmgångs-/Tjuvbromsdetektor]
6st. PHOEIX MDS WILD [Hjulskadedetektor]
25st. Schenck [Hjulskadedetektor]
4st. ABD [Aukustisk hjullagerdetektor]
0
100
200
300
Kilometer

Figure 2: Map of Trafikverket's DPC Detector Locations. Source: Trafikverket's official documentation [11].

Figure 3 provides a close-up view of the actual DPC detectors embedded in the railway tracks.

These detectors are strategically placed to measure the dynamic forces exerted by train wheels as

they pass over them. The image highlights their physical integration into the track infrastructure, illustrating the robust design required to endure constant exposure to heavy loads and environmental conditions.

- Wheel-Damage Detectors These detectors are crucial in identifying potential damage to the wheels as trains pass over them. There are two main types of wheel-damage detectors used by Trafikverket to monitor the condition of the wheels and those are the ones investigated in this study:
  - PHOENIX MDS WILD Detectors: Six of these detectors are installed at specific locations along the tracks. These sensors measure the dynamic forces exerted by the wheels on the track.
  - Schenck Detectors: Twenty-five Schenck detectors are spread across the railway system. Like the PHOENIX detectors, these sensors capture the dynamic pressure data from train wheels.

Figure 3: Image of a DPC detector. Source: Trafikverket's official documentation [12].

These detectors are responsible for measuring and collecting data, but they do not process or analyze it. The data collected by the detectors is provided to the customers (such as Green Cargo), who use their own software tools to analyze the data and identify potential issues, such as irregular wear or damage.

Table 1: Data collected from DPC detectors for each wheel during passage

Collected Information
Date & Time
Detector Location
Mean Load
Peak Load
Train Speed

For each wheel passage over a detector, the DPC system records comprehensive data for both wheels on the axle, as shown in Table 1. This information provides the foundation for analyzing wheel condition and predicting maintenance needs [13].

2.3.1 Wheel Shape and Force Measurement

However, there are limitations in the detection system. Since the wheels are conically shaped and oscillate as the trains move along the track, a damaged portion of the wheel might not always come into direct contact with the detector. This means that potential damage may go undetected if the wheel's damaged area is not aligned with the sensor at the time the train passes over it.

Figure 4: Conical train wheels showing contact points with rails. Only small areas of each wheel make contact with the track surface.

This conical shape significantly affects how DPC detectors collect sensor data. As shown in Figure 4, only a narrow section of each wheel's surface contacts the rail during operation. This has several important implications for data collection:

1. Limited Contact Area: Since only part of the wheel touches the detector, localized damage or wear may not be detected if it occurs on portions of the wheel that don't make regular contact with the rail.
2. Force Measurement Patterns: The concentrated contact creates specific force patterns in the sensor readings. A well-maintained wheel produces consistent force values, while damaged wheels create irregular patterns as different surface conditions pass over the detectors.
3. Multiple Measurements Needed: Because wheel damage might not always align with the sensor during a single pass, analyzing patterns across multiple detector encounters becomes essential for accurate condition assessment.

This geometric constraint explains why the machine learning models must analyze force variations over time rather than relying on individual measurements to predict wheel replacement needs effectively.

2.4 Current State of Predictive Maintenance at Green Cargo

In rail freight operations, PdM has significant potential to optimize performance, safety, and cost-efficieny. Green Cargo, a leading Swedish logistics company specializing in rail freight, operates a diverse fleet of cargo trains where wheels being critical components, undergo significant wear due to continuous loading, friction, and track impacts. Currently, decisions regarding wheel replacements are drawn from three main strategies: Run to Failure (R2F), Scheduled Maintenance, and a treshold-based system referred to as Predictive Maintenance.

- Run to Failure (R2F): is a maintenance strategy where components like wheels are allowed to fail before replacement occurs. While this may save on upfront maintenance costs, it is highly risky and costly, as it leads to unscheduled downtimes and requires emergency maintenance when a failure occurs, often in the middle of a track operation, as described in section 1.1.1 of Mobley [3].
- Scheduled Maintenance/Preventive Maintenance: on the other hand, involves replacing components at fixed intervals. While it helps avoid the extreme costs of R2F, it also

carries the risk of premature replacement, which results in unnecessary costs. Additionally, scheduled maintenance can be delayed, increasing the risk of encountering an R2F scenario, as described in section 1.1.2 of Mobley [3].

- Threshold-Based Predictive Maintenance leverages data from the DPC detectors, where predefined thresholds are set for specific sensor values. When these thresholds are exceeded, an alarm is triggered, prompting the scheduling of maintenance. As established by Trafikverket and detailed in wheel-rail impact studies [13], specific threshold limits have been set for locomotive wheel peak force measurements, as shown in Table 2. These peak force values represent the maximum impact force recorded when a wheel passes over the DPC detector. When wheels develop damage such as flat spots, cracks, or irregular wear, these damaged areas create significantly higher impact forces upon contact with the rail compared to smooth, undamaged wheel surfaces.

Table 2: Alarm limits for locomotives set by Trafikverket

Alarm Limit	Measurement Type	Threshold
High	Peak Load	425 kN
Warning (Low)	Peak Load	350 kN

The "high" alarm is triggered at 425 kilonewton, requiring immediate attention, while a "warning" alarm occurs at 350 kilonewton, allowing continued operation with speed restrictions to the final destination where certified staff should inspect the locomotive. These thresholds are universally applied across all Swedish railway operations regardless of train type, track segment, or operating company. They are designed to detect wheel conditions that create abnormally high peak force impacts during wheel passage, such as wheel flats, cracks, or surface irregularities that generate force spikes significantly above the average contact force.

While these regulatory limits apply to all operators, individual companies like Green Cargo may implement additional internal thresholds and decision criteria within their maintenance planning systems.

This threshold-based system has been and still is, a valuable tool for Green Cargo in planning maintenance activities. However, it does not fully utilize the potential of predictive analytics. It only responds to exceedances of predefined limits, rather than forecasting when failures are likely to occur. This limitation means that the current system is more reactive than predictive.

2.5 Machine Learning in Predictive Maintenance

Machine learning (ML), a subfield of artificial intelligence, offers powerful tools to process and learn from large datasets [14]. In the field of predictive maintenance (PdM), ML models uncover patterns in sensor data, enabling organizations to predict equipment failures with greater accuracy and optimize maintenance schedules [15], [16]. This capability minimizes unplanned downtime, reduces operational costs and extends equipment lifespan.

Predictive maintenance relies on the continuous monitoring of equipment through sensors that capture operational data such as temperature, pressure, vibration, and usage metrics. This data forms the foundation for machine learning models, which can identify patterns that signal wear, degradation or failure. Rather than relying on traditional scheduled or reactive maintenance, PdM powered by ML can anticipate failures before they occur, resulting in more efficient and cost-effective operations.

Machine learning approaches for PdM generally fall into two main categories: Supervised learning and Unsupervised learning. These methods differ fundamentally in their interaction with data and the types of maintenance challenges they address.

Page 15

2.6 Categories of Machine Learning

- Supervised Learning: In supervised learning, models are trained on historical datasets where input features (e.g., sensor readings) are labeled with corresponding outcomes (e.g., failure events, time-to-failure, or continuous damage levels). This makes it suitable for tasks such as predicting the remaining useful life (RUL) of components, forecasting continuous degradation progression, or classifying fault types. By learning direct relationships between inputs and outputs, supervised models excel in scenarios with abundant labeled data. For example, a supervised learning algorithm might predict the RUL of a pump based on vibration, temperature and pressure readings from previous instances [17].
- Unsupervised Learning: Unsupervised learning operates on unlabeled data, identifying hidden structures or patterns within it. This makes it valuable for detecting anomalies that may signal emerging issues or clustering similar failure modes for deeper analysis. Unsupervised learning can be useful when labeled data is scarce or when equipment behavior deviates from historical norms. For instance, anomaly detection might reveal unexpected behavior in a motor's performance that precedes a failure, even if no previous failures have been documented for that motor [18].

The choice of ML approach often depends on factors such as the availability of labeled data, the specific objectives of the maintenance strategy and the nature of the equipment being monitored. This study explores both categories by utilizing linear regression [8] for trend analysis, XGBoost [19] for regression-based damage prediction, and Isolation Forest [9] for unsupervised anomaly detection. These methods exemplify the breadth and versatility of ML in predictive maintenance.

2.6.1 Advantages of Machine Learning in Predictive Maintenance

Machine learning offers several advantages in PdM:

- Adaptability: Models can be tailored to a wide variety of equipment types and operating conditions.
- Scalability: Machine learning models can process vast amounts of data from multiple sensors across different systems, making them highly scalable in large industrial settings.
- Efficiency: By predicting failures early, ML models enable more targeted maintenance, which helps avoid unnecessary downtime and reduces operational costs.
- Cost-Effectiveness: Accurate failure predictions reduce the need for expensive, unnecessary repairs and replacements, ensuring maintenance is only conducted when truly needed.

Additionally, advancements in cloud computing and IoT (Internet of Things) technologies have made it easier to deploy and scale machine learning models for predictive maintenance, further enhancing their impact.

2.7 Machine Learning Models

2.7.1 Supervised Learning Model

Linear Regression Linear regression represents a supervised learning approach for modeling relationships between input features and continuous target variables. The method estimates the parameters of a linear equation that best fits the relationship between predictor variables and the response variable.

For simple linear regression with a single predictor:

$$y = \beta_0 + \beta_1 x + \epsilon$$

For multiple linear regression with multiple predictors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Page 16

Where y represents the dependent variable, x_i are the independent variables, β_0 is the intercept, β_i are the slope coefficients representing the rate of change, and ϵ is the error term.

Linear regression uses the ordinary least squares method to minimize the sum of squared residuals, providing optimal parameter estimates under certain assumptions. The method excels in applications requiring interpretable models and performs particularly well when the underlying relationship between variables is approximately linear [8].

Linear regression was selected for long–term forecasting because wheel damage follows approximately linear degradation trends over time, providing stable 3–year predictions. Kang et al. found linear regression effective for capturing degradation trends in their predictive maintenance study, noting it showed regular degradation patterns suitable for trend fitting [20]. It also serves as a baseline for validating XGBoost predictions.

Gradient Boosting Gradient Boosting is an ensemble learning technique that builds models sequentially, combining multiple weak learners (typically decision trees) to improve prediction accuracy. Each tree in the sequence is trained to correct the errors of its predecessors and the process iteratively minimizes a loss function to optimize predictions [21].

The general objective function combines a loss term with regularization term to balance fit and complexity:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, y_i) + \Omega(f)$$

Where:

- $L(y_i, y_i)$: Loss function (e.g., mean squared error for regression or log loss for classification).
- $\Omega(f)$: Regularization term that penalizes model complexity (e.g., tree depth or number of leaves).

XGBoost (eXtreme Gradient Boosting)
XGBoost refines Gradient Boosting with several computational and mathematical improvements [19]:

- Loss Function: Includes both the gradient and second-order derivative (Hessian) for more precise updates:

$$\text{Objective} = \sum_{i=1}^{n} \left[L(y_i, y_i) + \frac{1}{2} h_i w_2 \right] + \Omega(f)$$

  Where h_i is the Hessian (second derivative of the loss function) and w represents leaf weights.

- Regularization: Adds an L_1 or L_2-norm penalty to control tree complexity:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

  Where T is the number of leaves, and λ controls the regularization strength.

XGBoost was chosen because it excels with mixed tabular data, handles non-linear relationships between wheel damage factors, provides interpretable feature importance scores essential for Research Question 2, and includes built-in regularization to prevent overfitting. XGBoost has proven effective for regression–based damage prediction in predictive maintenance, successfully preventing 42% of production line failures [22].

2.7.2 Unsupervised Learning Models

Isolation Forest
Isolation Forest is an unsupervised algorithm specifically designed for anomaly detection. It isolates anomalies by constructing random decision trees that partition the data. Key characteristics include:

- Random Partitioning: Decision trees are created by recursively splitting the data based on randomly chosen features and thresholds.
- Path Length: Anomalies require fewer splits to isolate due to their sparsity in the feature space.
- Anomaly Score: An anomaly score is computed for each data point based on the average path length across trees:

$$\text{Anomaly Score} = 2^{-\frac{\bar{h}(x)}{c(n)}}$$

  Where:

  - $\bar{h}(x)$: Average path length for point x.
  - c(n): Normalizing constant based on sample size.

Isolation Forest excels in detecting rare or unexpected faults in complex, high-dimensional datasets and does not require labeled data, making it ideal for scenarios where anomalies are infrequent or previously unknown [9].

2.8 Model Optimization and Hyperparameter Tuning

Hyperparameter optimization is a critical component of machine learning model development that involves systematically searching for the optimal configuration of model parameters that are not learned during training. Unlike model parameters (such as weights in neural networks), hyperparameters control the learning process and model complexity [23].

Grid search is a comprehensive optimization technique that evaluates possible combinations of hyperparameters within predefined ranges. Combined with cross-validation, grid search provides robust hyperparameter selection by evaluating each configuration's performance across multiple data partitions, ensuring optimal model performance while preventing overfitting to specific data subsets [24].

2.9 Challenges and Opportunities

Working with sensor data introduces challenges such as missing values, outliers, and noisy signals. Pre–processing techniques, such as normalization, are essential for ensuring data quality. Feature engineering, where raw data is transformed into meaningful input features, also plays a crucial role in improving model performance. This process involves creating new variables from existing measurements, such as calculating damage indicators from peak and mean force values, implementing rolling statistics to capture temporal patterns, and encoding categorical variables for machine learning algorithms [25]. Effective feature engineering can significantly enhance model performance by providing algorithms with more informative representations of the underlying physical processes.

Another significant challenge in predictive maintenance is the limited availability of actual failure data for model validation. Unlike controlled laboratory experiments, real–world industrial systems rarely run to complete failure, making it difficult to validate predictions against confirmed failure events. This challenge is particularly pronounced in safety–critical applications where equipment is replaced preemptively based on threshold exceedance rather than actual failure occurrence.

The temporal nature of degradation presents additional complexity, as damage progression patterns may vary significantly between individual assets due to operational differences, environmental conditions, and maintenance history. Long–term forecasting compounds these challenges, as model accuracy typically decreases with extended prediction horizons.

Evaluating the effectiveness of machine learning models in regression–based predictive maintenance requires metrics that assess continuous prediction accuracy rather than discrete classification performance. Metrics such as coefficient of determination (R_2), Root Mean Square Error (RMSE) [26], and Mean Absolute Error (MAE) [26] are particularly valuable for understanding model performance in predicting continuous damage progression and threshold exceedance timing.

2.10 Evaluation Metrics

The effectiveness of machine learning models in predictive maintenance tasks is evaluated using regression metrics that assess the accuracy of continuous damage level predictions. In this study, the following evaluation metrics are employed to assess the models' ability to predict wheel damage progression:

- Coefficient of Determination (R_2): Measures the proportion of variance in the target variable that is predictable from the input features, providing a scale-independent assessment of model performance [27]. R_2 is defined as:

$$R_2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i)_2}{\sum_{i=1}^{n}(y_i - \bar{y})_2}$$

  Where y_i represents actual values, y_i represents predicted values, and $\bar{y}$ is the mean of actual values. R_2 values closer to 1.0 indicate better model performance.

- Root Mean Square Error (RMSE): Measures the average magnitude of prediction errors in the same units as the target variable (kN), providing interpretable accuracy assessment [26]:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - y_i)_2}$$

  Lower RMSE values indicate better prediction accuracy, with values closer to zero representing perfect predictions.

- Mean Absolute Error (MAE): Calculates the average absolute difference between predicted and actual values, providing a robust measure less sensitive to outliers [26]:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - y_i|$$

  MAE provides intuitive interpretation of average prediction deviation in kN units.

- Cross-Validation: Cross–validation is a statistical method used to assess model performance and generalization capability by partitioning data into multiple subsets. In k–fold cross–validation, the dataset is divided into k equal parts, where k-1 parts are used for training and one part for validation, with this process repeated k times to ensure each subset serves as validation data once [28]. This technique provides more robust performance estimates than simple train–test splits and helps detect overfitting by evaluating model stability across different data partitions. 5–fold cross–validation R² scores assess model stability and generalization capability across different data subsets, ensuring robust hyperparameter selection and preventing overfitting.

These regression metrics enable comprehensive evaluation of the model's ability to accurately predict continuous wheel damage values, essential for precise maintenance scheduling and threshold exceedance forecasting.

2.11 Uncertainty Quantification

Uncertainty quantification is essential in predictive maintenance to provide confidence bounds around predictions and enable risk–informed decision making. Monte Carlo simulation is a computational technique that uses repeated random sampling to model uncertainty propagation and estimate probability distributions of outcomes [29].

In predictive maintenance applications, Monte Carlo methods enable the quantification of prediction uncertainty by incorporating variability in model parameters, input features, and degradation processes. By running multiple simulations with randomly sampled variations, the technique produces probability distributions of future equipment conditions, providing decision makers with both point estimates and confidence intervals for maintenance planning.

Page 19

3. Related Work

Doshi and Shah [15] conducted a comparative study of machine learning techniques for predictive maintenance in industrial systems, evaluating multiple algorithms including Isolation Forest and XGBoost for anomaly detection and fault prediction. Isolation Forest achieved 86.5% precision for anomaly detection, while XGBoost achieved 93.5% accuracy for fault classification. Their study highlighted the effectiveness of these algorithms for processing high-dimensional sensor data in maintenance applications.

The relevance of this research lies in its validation of Isolation Forest and XGBoost for predictive maintenance, though their application differs from the current study: while Doshi and Shah applied XGBoost for fault classification, this study employs XGBoost for regression-based damage level prediction. Their findings provide foundational support for the algorithm selection in railway wheel maintenance applications.

McKinnon et al. [16] applied Isolation Forest to monitor wind turbine pitch systems using SCADA data, focusing on detecting anomalies that could indicate impending failures. Their study demonstrated that Isolation Forest effectively identified anomalies months before failures occurred, achieving a precision of 85%. This early detection capability was critical for minimizing maintenance costs and operational downtime. The approach taken by McKinnon et al. is relevant to this thesis as it highlights the utility of Isolation Forest for anomaly detection in systems that rely on complex sensor data for predictive maintenance.

Majidiparast et al. [30] present a prescriptive analytics framework for optimal intelligent predictive maintenance of railway tracks. The study utilizes Graph Convolutional Networks (GCNs) to analyze spatial and temporal data collected from railway systems, including geometric measurements and historical maintenance records. The model effectively predicts potential track failures before their occurrence, aiming to reduce downtime and extend the lifespan of railway assets. A significant contribution of this research is the modeling of the railway network as a bipartite graph, capturing the relationships between different segments of the railway infrastructure. This approach allows the GCN to efficiently propagate information across the network, enhancing the prediction accuracy for maintenance needs. The study reports a high level of effectiveness in its computational experiments, demonstrating the practical applicability of advanced machine learning techniques in railway maintenance. While this study focuses on track infrastructure, its methodology offers valuable insights for predictive maintenance in other components of railway systems, such as locomotive wheels. The use of GCNs to model spatial relationships and predict maintenance needs can inform approaches that analyze RFID sensor data for wheel maintenance. Incorporating similar graph-based modeling techniques could enhance the accuracy and efficiency of predictive maintenance strategies in different contexts within the railway industry.

Yang and Létourneau [5] developed a methodology to predict train wheel failures using operational and maintenance data from the railway industry. They addressed critical challenges including automatic labeling, feature extraction, and model building, achieving a 97% prediction rate of wheel failures while maintaining a reasonable false alert rate of 8%. Their approach yielded promising results in large-scale experiments, demonstrating significant potential for predictive maintenance in railways. The study successfully utilized Wheel Impact Load Detector data, showing that even with a focused data source, effective prediction models can be developed when the right machine learning techniques are applied. Their work established an important foundation for sensor-based predictive maintenance in railway applications.

Daniyan et al. [31] implemented artificial intelligence techniques for predicting wheel bearing failures in railcar learning factories. Their research utilized temperature data to monitor bearing degradation, establishing prediction models with continuous usage patterns over 480 days. The study achieved high prediction accuracy with R-values closer to 1, successfully identifying potential failure periods for wheel bearings, with first signs of degradation appearing after approximately 330 days of operation. While the approach demonstrated good practical value for maintenance planning, it was constrained by its focus on temperature as the sole degradation indicator, potentially missing other important failure modes.

Page 20

Shangguan and Xie [32] addressed the challenge of limited wheel degradation samples by developing a time series generator adversarial network (TimeGAN) to generate synthetic wheel degradation data. Their innovative approach used a sliding window technique to expand the input dataset and incorporated a stationary gamma process to improve generated data quality. The research demonstrated effective prediction of wheel diameter wear using the Gated Recurrent Unit (GRU) network with high accuracy. Though advanced in its methodology, the study's reliance on synthetic data introduces questions about how well the models would translate to varied real-world conditions.

Theissler et al. [4] surveyed the use of machine learning for predictive maintenance in the automotive industry, examining various applications including engine, battery, and transmission systems. They evaluated different ML techniques across these domains, finding Support Vector Regression particularly effective for predicting remaining useful life of components. Their comprehensive review highlighted application patterns and common challenges. The work provides valuable insights from an adjacent transportation sector, though its automotive focus means some railway-specific considerations are not addressed.

Davari et al. [33] surveyed data-driven predictive maintenance approaches for the railway industry, analyzing machine learning and deep learning techniques for monitoring temporal behavior and fault events in railway systems. Their comprehensive review identified a significant trend toward supervised learning methods in railway applications, with an almost equal division between anomaly detection and prediction tasks. While thorough in classification of approaches, they noted limitations in existing supervised learning solutions for the dynamic operational context of railway systems.

Le-Nguyen [34] studied real-time learning approaches for predictive maintenance of railway systems, developing a pipeline employing online machine learning to address the challenges of fast-paced data streams. Their implementation demonstrated the potential of online learning for automated data preprocessing in railway maintenance. The research showed promising results for real-time application, though full integration with existing maintenance protocols presented ongoing challenges.

Wang et al. [6] proposed a novel transformer-based framework with multiplex local-global temporal fusion (LGF-Trans) for high-speed train wheel wear prediction utilizing vibration signals. Their methodology combined attention mechanisms with both local and global temporal information to accurately predict wheel wear curves. The experimental results using real operational data from high-speed trains demonstrated superior performance compared to other deep learning methods. However, the complexity of the transformer architecture requires significant computational resources and large datasets for optimal performance.

Kunst et al. [35] conducted a systematic literature review on machine learning and reasoning for predictive maintenance in Industry 4.0, analyzing articles published between 2015 and 2020. Their review identified key frameworks, architectures, and tools in predictive maintenance, classifying research into integration issues, big data analysis, machine learning approaches, and reasoning with ontologies. This comprehensive analysis provided valuable context on the state-of-the-art and challenges in the field, but highlighted the need for more research on predictive maintenance models specifically designed for railway systems.

Kang et al. [20] developed a machine learning-based approach for predicting the remaining useful life (RUL) of equipment in production lines using artificial neural networks. Their methodology employed a two-stage learning process where linear regression was used for interpolation of degradation data, followed by multilayer perceptron neural networks for final RUL prediction. The study compared multiple algorithms including linear regression, random forest, and support vector regression for modeling equipment degradation patterns.

The authors found that linear regression achieved the best interpolation results for capturing degradation trends, noting that "the results of other algorithms do not show a regular degradation

trend, and as such, it is difficult to fit with a polynomial curve." Their approach was validated using NASA turbofan engine datasets, demonstrating the effectiveness of linear regression for modeling predictable degradation patterns in industrial equipment.

This research is relevant to the study performed in this report as it validates the use of linear regression for trend analysis in predictive maintenance applications, supporting the selection of linear methods for capturing wheel damage progression patterns in railway systems.

Taşcı et al. [22] developed a machine learning-based approach for predicting the Remaining Useful Life (RUL) of production lines in manufacturing using real-world IoT sensor data. Their study compared multiple regression algorithms including Random Forest, XGBoost, Multilayer Perceptron, and Support Vector Regression for equipment failure prediction. Among the evaluated methods, Random Forest achieved the best performance, followed closely by XGBoost for regression-based RUL prediction. The implemented XGBoost regression model successfully prevented approximately 42% of actual production line failures in real-time deployment, demonstrating the practical effectiveness of XGBoost for continuous value prediction in predictive maintenance applications.

This research validates the application of XGBoost for regression-based damage level prediction, similar to the current study's approach for wheel damage forecasting, though applied to different industrial equipment and sensor types.

These studies collectively demonstrate the evolving landscape of machine learning applications in predictive maintenance for railway systems, with particular focus on wheel degradation and failure prediction. The research highlights both significant advances in methodology and persistent challenges in data availability, model integration, and real-world implementation.

4. Problem Formulation

This research focuses on utilizing data from RFID sensors collected from DPC detectors placed on train tracks to monitor the forces acting on locomotive wheels in real-time. By applying machine learning algorithms to this sensor data, the aim is to predict the optimal time for wheel replacement, thereby minimizing downtime and reducing maintenance costs.

To narrow the scope and provide actionable insights, this study will specifically focus on three selected locomotives operating on the "Norra Stålpendeln" route between Borlänge and Boden. These locomotives have been chosen due to their consistent operational patterns and critical role in freight logistics, allowing for a detailed and controlled analysis. While this study focuses on locomotives, the goal is to develop a model that can later be extended to wagons and entire trains.

This thesis will investigate the following key problems:

- RQ1: How can Machine Learning algorithms be applied to sensor data, specifically force impact data, to predict the need for wheel replacement?
- RQ2: How can the most significant features be extracted to optimize wheel replacement prediction?

Motivation: Railway maintenance is a critical aspect of ensuring safety, reliability, and cost efficiency in freight operations. Locomotive wheels experience significant wear and tear, and the ability to accurately predict their replacement needs can lead to substantial operational improvements. The choice of XGBoost, Isolation Forest and Linear Regression aligns with this motivation, as these models are well-suited for handling large datasets with imbalances and complex feature interactions. While this study focuses on locomotives, extending the model to wagons and entire trains has the potential to significantly enhance predictive maintenance practices across the railway industry.

Goals: To achieve the purpose of this research and address the identified research questions, the following specific goals have been established:

- Develop and implement a machine learning pipeline to process and analyze RFID sensor data from the selected locomotives.
- Identify the most significant data features for predicting wheel replacement.
- Evaluate the performance of the prediction model using metrics such as R_2, Root Mean Square Error and Mean Absolute Error.
- Assess the adaptability of the model for future use with wagons and entire trains.
- Provide recommendations for integrating predictive maintenance solutions into existing railway operations.

Assumptions and Limitations: The study assumes that the RFID sensor data is sufficiently accurate to reflect real-world conditions for the purposes of this analysis. However, the data may still contain noise or inaccuracies that could impact the model's performance. The exclusion of wagons limits the immediate generalizability of the findings, but the methodology is designed to be adaptable for future studies involving entire trains. Furthermore, the analysis is geographically and operationally specific to the Norra Stålpendeln route, which may restrict its broader applicability.

Page 23

5. Method

5.1 Data Collection and Preparation

This study utilized RFID sensor data collected from the track infrastructure along the Norra Stålpendeln route between Borlänge and Boden. The sensors measure vertical force impacts as train wheels pass over them, providing continuous monitoring of wheel condition. The dataset encompasses 15,000 measurements from three locomotives (M4004, M4008, M4016), each equipped with six axles with approximately 834 data points.

The raw RFID measurements include peak force values and mean force values for both left and right wheels, recorded with timestamps and associated vehicle identification. These force measurements directly relate to wheel condition, as damaged wheels create larger impact forces when passing over the sensors. The axle weight measurements were also included to account for load variations that might influence damage progression.

5.2 Feature Engineering Methodology

The feature engineering process was designed to capture multiple aspects of wheel degradation. Dynamic features were created by calculating the difference between peak and mean values for each wheel:

$$\text{Dynamic}_{\text{wheel}} = \text{Peak}_{\text{wheel}} - \text{Mean}_{\text{wheel}} \qquad (1)$$

This metric directly quantifies localized damage severity. When wheels are healthy, forces are distributed evenly around the wheel circumference, resulting in peak values close to mean values (small dynamic). When damage occurs at specific locations, those points experience significantly higher forces, creating large dynamic values.

Additional engineered features include:

- Rolling statistics using adaptive window sizes: Calculated using a sliding window approach where window size = min(10, total measurements available). For each measurement, the rolling mean is computed as:

$$\text{Rolling Mean}_i = \frac{1}{w} \sum_{j=\max(1,i-w+1)}^{i} \text{Dynamic}_j \qquad (2)$$

where w is the window size and i is the current measurement index.

- Days since start and measurement index: Days since start calculated as:

$$\text{Days}_i = \frac{\text{Timestamp}_i - \text{Timestamp}_{\text{first}}}{86400 \text{ (seconds per day)}} \qquad (3)$$

Measurement index is simply the sequential position (1, 2, 3, ...) of each measurement within each vehicle-axle combination.

- Encoded vehicle and axle identifiers: Categorical variables converted to numerical values using Label Encoding, where each unique vehicle/axle receives a sequential integer identifier (Vehicle A=0, Vehicle B=1, etc.).

- Rolling mean of dynamic values: Computed using the same sliding window approach as rolling statistics above, applied specifically to the dynamic damage values to reduce short-term noise while preserving underlying degradation trends.

The window size of 10 was selected to balance capturing recent trends while maintaining statistical stability. With the observed measurement frequency, this approximates one week of operational history.

Page 24

5.3 Anomaly Detection Approach

Isolation Forest was selected for anomaly detection due to its effectiveness with high-dimensional data and its ability to identify outliers without assuming any underlying data distribution. The contamination parameter was set to 0.05, identifying the most extreme 5% of measurements as anomalies.

The 5% contamination rate reflects the operational reality that extreme wheel force conditions are infrequent but critical in railway systems. Given the 15,000-measurement dataset, this identifies 750 anomalies which, when duplicated in the enhancement strategy, results in approximately 9.5% of the final training dataset consisting of extreme operational conditions - providing sufficient representation of rare but critical patterns while maintaining normal operational data dominance (90.5%).

Rather than removing these anomalies, the implementation enhanced the training data by duplicating them. This decision was based on the precautionary principle: if these extreme values represent pre-failure conditions, removing them would eliminate the model's ability to learn these critical patterns.

5.4 Hybrid Prediction Approach: Linear Regression and XGBoost

The forecasting methodology employs a novel hybrid approach combining linear regression with XGBoost predictions. This dual-model strategy addresses the inherent limitations of each method when used in isolation.

Linear regression captures long-term degradation trends by fitting:

$$\text{Peak}_{\text{damage}}(t) = \beta_0 + \beta_1 \cdot t + \epsilon \qquad (4)$$

where t represents days since first measurement.

This provides stable, physically interpretable trend projections essential for long-term planning.

XGBoost excels at capturing complex non-linear patterns but requires known feature values for prediction.

The hybrid approach implements a time-dependent weighting scheme:

$$\text{Prediction}_{\text{final}}(t) = w_{\text{linear}}(t) \cdot \text{LR}(t) + w_{\text{XGB}}(t) \cdot \text{XGB}(t) \qquad (5)$$

where:

$$w_{\text{linear}}(t) = \min\left(0.7, \frac{t}{1095}\right) \qquad (6)$$

$$w_{\text{XGB}}(t) = 1 - w_{\text{linear}}(t) \qquad (7)$$

This creates a smooth transition: early predictions (days 0-150) rely primarily on XGBoost, middle-term predictions blend both approaches, and long-term predictions (days 750+) rely primarily on linear trends, where 1095 represents the total forecast horizon in days (3 years = 3 × 365 days). The 0.7 cap was selected to ensure XGBoost always contributes at least 30% to predictions, preventing complete reliance on linear extrapolation which may not capture sudden changes in degradation patterns or operational anomalies that could affect wheel condition.

Hybrid Integration Example: The time-dependent weighting creates a smooth transition from pattern recognition to trend stability:

Day	Linear Weight	XGBoost Weight
1	0%	100%
250	23%	77%
500	46%	54%
750	69%	31%
1095	70%	30%

Page 25

The mathematical progression: $w_{\text{linear}}(t) = \min(0.7, t/1095)$ demonstrates the gradual shift from short-term accuracy to long-term reliability.

Alternative Model Comparison: Random Forest was rejected due to bootstrap sampling being unsuitable for temporal forecasting where recent patterns should receive higher weight than historical averages. Neural Networks were excluded due to insufficient training data per axle (834 measurements per axle) and black-box nature incompatible with safety-critical applications. Support Vector Regression was less effective with mixed categorical and numerical features compared to tree-based methods.

5.5 Model Architecture and Training Strategy

XGBoost was chosen as the primary predictive model due to its proven effectiveness in handling heterogeneous features, its built-in regularization to prevent overfitting, and its ability to provide feature importance metrics crucial for answering the research questions.

A global training strategy was adopted after empirical testing of alternatives:

- Individual axle models: Failed due to insufficient data (approximately 834 data points per axle)
- Vehicle-specific models: Limited cross-vehicle learning
- Global models: Leveraged all available data while maintaining asset-specific information through encoded features

5.6 Hyperparameter Optimization

GridSearchCV with 5-fold cross-validation systematically explored the hyperparameter space:

$$\text{Total combinations} = 3 \times 3 \times 3 \times 2 \times 2 = 108 \tag{8}$$

The parameter grid encompassed:

- n_estimators: [300, 500, 700] – Below 300 showed underfitting; above 700 provided diminishing returns
- max_depth: [4, 6, 8] – Captures 4-8 levels of feature interactions without excessive complexity
- learning_rate: [0.05, 0.1, 0.15] – Balances convergence speed with training stability
- subsample: [0.8, 0.9] – Introduces mild stochasticity while using most data
- colsample_bytree: [0.8, 0.9] – Feature sampling to reduce overfitting

These ranges balance comprehensive search with computational feasibility, resulting in 540 model fits per wheel.

5.7 Damage Threshold Definition

The study employed three damage thresholds for maintenance planning:

- 425 kN: Official Trafikverket threshold for mandatory wheel replacement
- 350 kN: Trafikverket warning threshold for maintenance planning
- 275 kN: Additional threshold introduced for early warning

The 275 kN early warning threshold was selected to provide Green Cargo with internal alerts before Trafikverket's regulatory warnings are triggered. This enables maintenance teams to analyze damage progression patterns and trends leading up to potential threshold exceedance, supporting risk-informed maintenance decisions rather than purely reactive threshold responses. The 75 kN step-down from the regulatory 350 kN warning level maintains consistency with Trafikverket's framework while providing additional lead time for proactive maintenance planning.

5.8 Evaluation Methodology

The evaluation framework employs multiple complementary approaches:

5.8.1 Cross-Validation Strategy

5-fold cross-validation with temporal awareness ensures:

- Each fold maintains chronological order (no future data in training)
- Performance stability across different data subsets
- Robust hyperparameter selection

The choice of 5 folds is a commonly accepted compromise that balances bias and variance in model evaluation. Using fewer folds (e.g., 3) can lead to higher bias and less reliable estimates, while more folds (e.g., 10) increase computational cost and may reduce the size of validation sets, especially with limited data. With temporal data, 5 folds provide enough data in each training set to capture temporal patterns while preserving a strict forward-looking evaluation.

5.8.2 Performance Metrics

Primary evaluation metrics include:

Regression metrics:

- Coefficient of determination (R_2) for scale-independent performance comparison
- Root Mean Square Error (RMSE) for absolute prediction accuracy in kN units
- Mean Absolute Error (MAE) for average prediction deviation
- Residual analysis for model validation and bias detection

Operational metrics:

- Lead time before threshold exceedance for maintenance scheduling
- Prediction interval coverage for uncertainty assessment

5.8.3 Feature Importance Analysis

Feature importance analysis addresses Research Question 2 by quantifying the relative contribution of each input feature to model predictions. XGBoost calculates importance scores during tree construction, where importance represents the average gain contributed by each feature when used for splitting across all trees in the ensemble.

Since separate models are trained for left and right wheels with equivalent but distinct feature sets (e.g., Left_Dynamic vs Right_Dynamic), feature importance is calculated through equivalent feature mapping and averaging:

$$\text{Unified Importance} = \frac{\text{Left Model Importance} + \text{Right Model Importance}}{2} \tag{9}$$

This approach provides unified feature rankings while accounting for both wheel positions, enabling comprehensive guidance for maintenance system optimization.

5.9 Uncertainty Quantification

Monte Carlo simulation with 100 iterations calculates prediction uncertainty. This number balances computational time (approximately 15 minutes for full fleet) with adequate sampling of the uncertainty space. The approach:

1. Adds stochastic noise scaled to 30% of historical volatility: Random noise is added to the model forecasts in each Monte Carlo iteration to simulate variability in operational conditions and measurement errors. The noise level is calibrated to reflect realistic variability observed in the training data, though not explicitly calculated from historical volatility as standard deviation of damage changes over time. Instead, the noise magnitude is through testing chosen to balance realism and signal clarity. The noise simulates unpredictable factors such as varying load conditions, track irregularities, and measurement uncertainty that affect wheel damage progression.
2. Propagates uncertainty through both model components: The added noise affects both the linear regression trend projection and the XGBoost feature inputs. For linear regression, noise is added directly to the trend-based damage progression. For XGBoost, the noisy values become inputs for dynamic features, rolling statistics, and temporal variables, allowing the model to predict how uncertainty in current conditions affects future damage levels. This ensures that uncertainty is consistently carried through the entire hybrid prediction process.
3. Calculates percentile-based confidence intervals (10th and 90th): After 100 simulation iterations, the 10th and 90th percentiles of the prediction distribution are extracted to create 80% confidence intervals. These percentiles were chosen to provide meaningful uncertainty bounds without being overly conservative (95% intervals) or overly narrow (50% intervals). The resulting intervals indicate that 80% of potential outcomes fall within the specified range, providing maintenance planners with realistic uncertainty estimates for scheduling decisions.

The 30% scaling prevents unrealistic uncertainty growth over the 3-year forecasting period while maintaining meaningful variation.

5.10 Maintenance Decision Framework

The methodology implements practical maintenance logic:

- Both wheels on an axle replaced simultaneously (operational constraint): This constraint is implemented by treating axle replacement as a single decision event. When either wheel reaches a damage threshold, the entire axle (both wheels) is scheduled for replacement. The algorithm identifies the replacement timing but applies it to the complete axle unit, reflecting real–world maintenance practices where replacing both wheels at the same time is a very common approach.

- First wheel to exceed threshold triggers replacement: The implementation iterates through each day of the 3-year forecast for both wheels, identifying the earliest day when either wheel's predicted damage exceeds a threshold. The replacement schedule uses 'min(left_exceed_days, right_exceed_days)' to determine axle replacement timing. For example, if the left wheel is predicted to exceed 275 kN on day 800 and the right wheel on day 950, the axle replacement is predicted for day 800.

- Three threshold scenarios provide sensitivity analysis: The algorithm processes each axle through three threshold levels (275 kN early warning, 350 kN regulatory warning, 425 kN mandatory replacement) sequentially. For each threshold, it determines: (a) which wheel will fail first, (b) the exact timing in days, and (c) whether replacement is needed within the 3-year forecast horizon. This generates three maintenance scenarios per axle, allowing comparison of proactive versus reactive maintenance strategies.

This framework transforms model predictions into actionable maintenance decisions aligned with railway operational realities.

6. Ethical and Societal Considerations

This research explores machine learning approaches for predictive maintenance of freight train wheels, presenting several ethical and societal considerations worthy of examination.

6.1 Research Ethics and Data Confidentiality

While this study does not involve human subjects, it does utilize sensitive industrial data that requires ethical handling:

- Confidential Data Management: Several aspects of the data used in this research are confidential, including specific sensor measurement values, actual maintenance costs and wheel lifespan information. This research respects these confidentiality requirements by avoiding the publication of raw data values, specific cost figures, or proprietary technical specifications of wheel components. All results are presented in a manner that protects Green Cargo's business interests while maintaining scientific validity.

- Data Integrity: The accuracy and reliability of sensor data analysis are essential, as inaccurate predictions could potentially affect maintenance planning and operations.

- Transparency and Reproducibility: While protecting confidential information, methodology has been documented to ensure scientific reproducibility of the research approach.

6.2 Societal Impact

The potential implementation of machine learning-based predictive maintenance carries several societal implications:

6.2.1 Economic Considerations

- Potential Resource Optimization: This research aims to develop models that could help predict more optimal wheel replacement timing, potentially reducing instances of premature replacement.

- Operational Planning: By forecasting maintenance needs more accurately, railway operators may be able to better plan maintenance activities, potentially reducing unplanned downtime.

6.2.2 Environmental Sustainability

- Resource Conservation: If successful, more precise maintenance scheduling could contribute to more efficient use of materials and potentially reduce waste from unnecessary replacements.

- Support for Rail Transport: Improving maintenance practices supports rail freight as a transportation mode, which generally has lower environmental impact per ton-kilometer compared to road transport.

6.2.3 Safety Implications

- Proactive Maintenance: Early detection of wheel degradation patterns could potentially help identify issues before they lead to operational problems, contributing to overall railway safety.

6.3 Ethical Considerations in Implementation

- Algorithmic Transparency: The selected machine learning models (XGBoost and Isolation Forest) offer a balance between predictive capability and interpretability, allowing for human understanding of maintenance recommendations.

- Human Decision-Making: The research proposes systems to support rather than replace human decision-making in maintenance planning, maintaining essential human oversight of safety-critical processes.

This research aims to explore technological innovations for maintenance optimization while respecting confidentiality requirements and considering the broader economic, environmental and safety implications of the approach.

7. Implementation

7.1 Implementation Architecture and Strategy

The implementation uses raw DPC sensor measurements into actionable 3–year maintenance forecasts through a four–stage pipeline designed to address the specific challenges of railway wheel damage prediction. The challenges primarily includes limited per-axle data, heterogeneous fleet characteristics and the need for long-term forecasting reliability.

Stage 1: Data Preparation and Feature Engineering The pipeline begins by calculating dynamic damage indicators immediately upon data loading, as these engineered features form the foundation of all subsequent analysis. The implementation prioritizes this calculation because the peak-minus-mean metric directly works as a wheel–damage indicator.

Rolling statistics implementation uses adaptive windowing to handle the reality of varying data availability across the fleet. Newer axles with potentially limited operational history, receive proportionally smaller rolling windows, while established axles utilize the full 10-measurement context. This design ensures all axles can be processed while maximizing the use of available historical information.

Stage 2: Global Model Strategy Empirical testing revealed that individual axle models failed due to insufficient data (approximately 834 data points per axle), the implementation adopts a global approach that pools all 15,000 measurements. This strategy addresses the fundamental data insufficiency challenge while preserving asset-specific information through categorical encoding, enabling the model to learn both universal damage patterns and vehicle-specific characteristics.

Stage 3: Anomaly-Enhanced Training The implementation employs a novel anomaly enhancement strategy rather than traditional anomaly removal. Isolation Forest identifies extreme operational conditions (5% contamination threshold), but rather than removing these samples, the system doubles their representation in the training set. This decision is based of the safety-critical nature of wheel maintenance, where rare extreme conditions may represent pre–failure states that the model must learn to recognize.

Stage 4: Hybrid Forecasting Framework Long–term forecasting combines XGBoost pattern recognition with linear trend extrapolation through a time–dependent weighting scheme. This hybrid approach addresses the fundamental challenge that XGBoost requires feature projection into the future, which becomes increasingly uncertain over extended time horizons.

7.2 Software Environment and Libraries

The data processing, modeling, and analysis were implemented using Python 3.10. The following libraries and their respective versions were utilized to ensure reproducibility:

- pandas 2.2.3
- numpy 2.1.3
- xgboost 3.0.0
- scikit-learn 1.6.1
- matplotlib 3.10.1

Development was performed using the PyCharm IDE for code management and debugging.

7.3 Critical Implementation Decisions and Rationale

Global vs. Individual Modeling Strategy The choice of global modeling directly determines the exceptional performance metrics reported in Section 8.1. Individual axle approaches were

tested but failed to achieve stable predictions due to insufficient training data. The global approach enables R_2 scores exceeding 0.99 by leveraging the full 15,000-measurement dataset while maintaining asset–specific learning through encoded features.

Feature Engineering Implementation Dynamic damage calculation serves as the primary feature because it is a direct indicator of wheel damage. The implementation computes this as:

```
df [ ’Left_Dynamic ’ ] =
        df [ ’ Left Wheel Damage Peak Value ’ ] − df[ ’Left Wheel Damage Mean Value’]
df [ ’Right_Dynamic ’ ] =
        df [ ’ Right Wheel Damage Peak Value ’ ] − df[ ’Right Wheel Damage Mean Value’]
```

Listing 1: Core damage indicator calculation

This engineering choice directly contributes to the 58.3% feature importance of dynamic indicators reported in Section 8.4, validating the implementation’s alignment with physical understanding.

Anomaly Enhancement Strategy Traditional anomaly detection removes outliers, but railway maintenance requires learning from extreme conditions. The implementation increases anomaly representation:

```
X_right_enhanced = np. vstack ([ X_right , X_right [ right_anomalies ] ] )
y_right_enhanced = np. concatenate ([ y_right , y_right [ right_anomalies ]])
X_left_enhanced = np. vstack ([ X_left , X_left [ left_anomalies ]])
y_left_enhanced = np. concatenate ([ y_left , y_left [ left_anomalies ]])
```

Listing 2: Anomaly enhancement implementation

This increases anomaly weight from 5% to approximately 9.5% of the training set, ensuring rare patterns receive enough learning attention without overwhelming normal operational patterns.

Hybrid Forecasting Logic The time-dependent weighting implementation addresses XGBoost’s limitation in long–term forecasting:

```
weight_linear = min (0.7, day / forecast_days)                          # forecast_days = 1095
weight_xgb = 1 − weight_linear
final_prediction = weight_linear ∗ linear_trend + weight_xgb ∗ xgb_prediction
```

Listing 3: Hybrid prediction weighting

This creates a smooth transition from XGBoost-dominated short–term predictions (high accuracy) to linear–dominated long-term predictions (high stability), directly enabling the reliable 3-year maintenance schedules reported in Section 8.3.2.

7.4 Model Training and Optimization Implementation

GridSearchCV Strategy Hyperparameter optimization explores the full parameter space using 5-fold cross-validation with R_2 scoring. The implementation discovers significantly different optimal parameters for left and right wheels:

- Left Wheel: 700 estimators, depth 6, learning rate 0.15
- Right Wheel: 500 estimators, depth 4, learning rate 0.05

These differences indicate that left wheels exhibit more complex damage patterns requiring greater model capacity, a finding that emerges directly from the data–driven optimization process.

Cross-Validation Implementation The 5–fold cross–validation strategy ensures robust hyperparameter selection while preventing overfitting. The implementation forms the splits to maintain representative vehicle-axle distributions across folds, contributing to the strong generalization performance (CV R_2 scores of 0.9975 and 0.9873) reported in Section 8.1.

7.5 Forecasting Implementation Framework

Feature Projection Strategy Long–term forecasting requires projecting features into the future. The implementation handles different feature types appropriately:

- Static features (vehicle/axle encoding): Remain constant
- Baseline features (mean damage values): Remain constant
- Temporal features: Increment linearly (days_since_start += 1 per day)
- Dynamic features: Project using historical trends with stochastic variation

Monte Carlo Uncertainty Quantification The implementation uses 100–iteration Monte Carlo simulation to quantify prediction uncertainty. Each iteration adds stochastic noise scaled to 30% of historical volatility:

```
for sim in range (100):
    left_noise = np.random.normal(0, left_volatility ∗ 0.3)
    right_noise = np.random.normal(0, right_volatility ∗ 0.3)
    projected_damage += trend_progression + noise
```

Listing 4: Uncertainty quantification implementation

The 30% volatility scaling prevents unrealistic uncertainty explosion over 3 years while maintaining realistic operational variation. This implementation enables the confidence intervals (±0.7 kN at 95% confidence) reported in Section 8.1.

Physical Constraint Enforcement The implementation enforces realistic bounds throughout forecasting:

```
# Prevent negative dynamics (ensuring peak is greater than mean)
sim_left_dynamic = max(0.1 , sim_left_dynamic)
sim_left_peak = max(current_left_mean , sim_left_peak)
```

Listing 5: Physical constraint implementation

These constraints ensure predictions remain physically meaningful, contributing to the realistic maintenance forecasts that show only one axle requiring replacement within three years.

7.6 Maintenance Decision Logic Implementation

Threshold Exceedance Detection The implementation processes each forecast through three threshold levels (275, 350, 425 kN) to provide sensitivity analysis. For each threshold, the system identifies the first day when either wheel exceeds the limit:

```
for day , value in enumerate ( forecasts [ ’ right_projections ’ ]):
    if value >= threshold:
        right_exceed_days = day + 1
        break

for day , value in enumerate ( forecasts [ ’ left_projections ’ ]):
    if value >= threshold:
        left_exceed_days = day + 1
        break

axle_replacement_days = min(left_exceed_days , right_exceed_days)
first_wheel = ’Left’ if left_exceed_days <= right_exceed_days else ’Right’
```

Listing 6: Maintenance schedule logic

This logic implements the realistic operational constraint that both wheels are replaced when either one exceeds a threshold, directly generating the maintenance schedules reported in the Section 8.3.2.

Fleet-wide Analysis Pipeline The implementation processes all 18 vehicle–axle combinations systematically, generating individual forecasts and aggregating fleet-wide maintenance requirements. This comprehensive approach enables the results in Section 8.3.2, finding that only one axle across the entire fleet requires replacement within three years at the 275 kN threshold.

7.7 Validation and Output Generation

Performance Validation Implementation The implementation employs rigorous validation through train–test splits (80-20) with vehicle-stratification to ensure representative sampling. Cross-validation scores validate hyperparameter selection, while residual analysis confirms model assumptions. This validation framework directly supports the exceptional performance metrics ($R_2 > 0.99$, RMSE < 1 kN) reported in Section 8.1.

Comprehensive Output Generation The system produces multiple output formats:

- Individual forecast plots for each vehicle-axle combination with uncertainty bands
- CSV maintenance schedules with replacement timing for each threshold scenario
- Summary statistics aggregating fleet-wide maintenance needs
- Performance metrics and model diagnostics

Computational Performance The implementation achieves practical deployment performance:

- Data preparation: <10 seconds for 15,000 measurements
- Model training: approximately 30 minutes including GridSearchCV optimization
- Forecast generation: <5 minutes for complete fleet analysis
- Total pipeline: <35 minutes from raw data to actionable maintenance schedules

This performance enables regular fleet assessment and maintenance planning updates in operational environments.

7.8 Implementation Validation

From Implementation to Performance Metrics The exceptional model performance ($R_2 > 0.99$) reported in Section 8.1 emerges directly from implementation choices. Global modeling strategy providing sufficient training data, feature engineering capturing physical damage phenomena, and anomaly enhancement ensuring learning from extreme conditions. Alternative approaches tested during implementation (individual axle models, standard anomaly removal) failed to achieve comparable performance.

From Feature Engineering to Feature Importance The dominance of dynamic damage indicators (58.3% importance) in Section 8.4 directly validates the implementation’s focus on dynamic calculations. The rolling statistics contributing 10.7% importance demonstrate how the adaptive windowing implementation captures temporal patterns.

From Hybrid Approach to Maintenance Predictions The minimal maintenance requirements identified in Section 8.3.2 (one axle replacement in three years) result from the hybrid forecasting implementation, balancing pattern recognition with trend stability. Pure XGBoost approaches tested during development produced less stable long–term predictions, while pure linear extrapolation missed complex damage patterns, validating the hybrid implementation strategy.

This implementation framework transforms raw sensor measurements into reliable long–term maintenance forecasts, directly provides insights for a potential transition from reactive threshold-based maintenance to proactive predictive planning demonstrated in Section 8..

8. Results

8.1 Model Performance and Validation

The XGBoost models achieved near-perfect performance across all evaluation metrics (Table 3). The left wheel model demonstrated a perfect test R_2 score of 1.0000 with an RMSE of 0.08 kN, while the right wheel model achieved a test R_2 of 0.9993 with an RMSE of 0.36 kN. Cross-validation scores of 0.9975 (left) and 0.9873 (right) confirm strong model stability and generalization across data folds.

Table 3: Comprehensive Model Performance Summary

Metric	Left Wheel	Right Wheel
Training R_2	1.0000	0.9993
Test R_2	1.0000	0.9993
Cross-Validation R_2	0.9975	0.9873
Training RMSE (kN)	0.08	0.36
Test RMSE (kN)	0.08	0.36
Training MAE (kN)	0.06	0.22
Test MAE (kN)	0.06	0.22
Best n_estimators	700	500
Best max_depth	6	4
Best learning_rate	0.15	0.05

8.1.1 Performance Differences Between Left and Right Wheel Models

The analysis reveals notable performance differences between wheel positions, with left wheel models achieving superior accuracy (R_2 = 1.0000, RMSE = 0.08 kN) compared to right wheel models (R_2 = 0.9993, RMSE = 0.36 kN). The right wheel model shows 4.5 times higher prediction error, suggesting systematic factors affecting wheel position measurements differently.

Several potential contributing factors may explain these differences. As described in Section 2.3.1, the conical wheel geometry means that detectors capture different surface areas of left versus right wheels depending on wheel oscillation patterns and contact positioning during passage. Different operational load distributions, noted as factors not included in this thesis, could create asymmetric wear patterns between wheel positions. Additionally, track–specific wear characteristics may result in different deterioration rates on each side of the rails, while certain wheel surface areas may be more prone to damage development – aspects not investigated in this study. The higher variability in right wheel predictions (RMSE 0.36 kN vs 0.08 kN) suggests that right wheels may exhibit more complex damage patterns or that the detector measurements capture less consistent surface conditions for this wheel position.

8.1.2 Critical Assessment of Model Performance

While these results appear outstanding, several factors warrant careful consideration. The extremely high R_2 values (>0.99) raise potential concerns about overfitting, despite showing no generalization gap between training and test performance. However, several factors support the validity of these results:

- Physical consistency: The RMSE values (0.08-0.36 kN) are small relative to the damage threshold ranges (275-425 kN), representing approximately 0.08-0.09% prediction error
- Cross-validation stability: The modest gap between test R_2 and CV R_2 (0.0025 for left, 0.0120 for right) indicates genuine predictive capability rather than data leakage
- Feature interpretability: The model’s reliance on physically meaningful features (dynamic damage indicators) supports mechanistic validity

The parity plots (Figure 5) demonstrate strong correlation between predicted and actual values across the full range of damage levels, with prediction intervals providing realistic uncertainty bounds of ±0.7 kN at 95% confidence.

Figure 5: Parity plots showing predicted vs. actual wheel damage values with 95% confidence intervals for both left and right wheel models. The strong correlation along the diagonal indicates excellent predictive accuracy.

8.2 Data processing and Anomaly Detection Results

Global Modeling Strategy Success

The global modeling strategy proved essential for handling the heterogeneous fleet data and provides a key methodological strength. Individual axle-specific models failed due to insufficient data (approximately 834 points per axle), while the global approach leveraged all 15,000 measurements for training while preserving asset-specific patterns through categorical encoding. This strategy provides substantial training data volume (15,000 samples) while enabling predictions for individual axles, effectively solving the data insufficiency problem that would plague axle–specific approaches.

8.2.1 Feature Engineering Effectiveness

The dynamic damage indicator (Peak − Mean) proved to be the most critical engineered feature, representing 58.3% of model importance. This metric successfully captures the fundamental physics of wheel damage: healthy wheels distribute forces evenly (small dynamic values), while damaged wheels create localized force spikes (large dynamic values).

8.2.2 Anomaly Detection Integration

Isolation Forest identified 750 anomalies per wheel (5% of data) representing extreme operational conditions. Rather than removing these outliers, the enhancement strategy increased their representation to 9.5%, ensuring the model learns from rare but potentially critical pre–failure patterns. The isolation forest analysis (Figure 6) shows these anomalies cluster around extreme peak values and unusual feature combinations.

Figure 6: Isolation Forest anomaly detection results showing the distribution of normal data points versus detected anomalies across different feature dimensions. Anomalies (red points) cluster around extreme values and unusual feature combinations.

8.3 Predictive Maintenance Forecasting Results

The 3–year forecasting analysis across all 18 vehicle-axle combinations reveals different degradation patterns (Figure 7). The implementation generated forecasts for 1,095 days, revealing significant variation in wheel condition trajectories:

Figure 7: Linear regression trends for all 18 vehicle-axle combinations showing historical damage progression patterns. Each subplot represents one axle, with separate lines for left (blue) and right (red) wheels, illustrating the diversity of degradation patterns across the fleet.

8.3.1 Fleet-Wide Damage Trends

Analysis of historical trends across all 18 axles shows mixed patterns across the fleet:

- Improving conditions: 9 left wheels and 7 right wheels show decreasing damage trends
- Deteriorating conditions: 9 left wheels and 11 right wheels show increasing damage trends
- Average trends: -0.0091 kN/day (left), -0.0109 kN/day (right)

Figure 8: Historical linear trends for M4004 Axle 1 showing decreasing damage patterns (-0.0132 kN/day left, -0.0185 kN/day right) from current levels of 142 kN (left) and 116 kN (right).

Figure 9: Historical linear trends for M4008 Axle 3 showing increasing damage patterns (0.1456 kN/day left, 0.1338 kN/day right) from current levels of 128 kN (left) and 124 kN (right), representing the degrading condition that leads to predicted threshold exceedance.

The selected examples illustrate this diversity: M4004 Axle 1 (Figures 8 and 10) represents the improving condition scenario with negative trends leading to no predicted maintenance needs, while M4008 Axle 3 (Figures 9 and 11) demonstrates how positive historical trends translate into future threshold exceedance predictions.

8.3.2 Maintenance Scheduling Outcomes

The forecasting analysis identified minimal immediate maintenance needs:

- Total axle replacements needed: 1 out of 18 axles within 3 years
- Threshold triggering replacement: 275 kN (earliest warning level)
- Replacement timing: Year 3 of forecast period

Page 39

- First wheel to fail: Left wheel (M4008, Axle 3)

This outcome suggests excellent current fleet condition. The single predicted replacement corresponds to M4008 Axle 3, which exhibits the highest positive damage trend (0.1456 kN/day for left wheel).

8.3.3 Forecast Diversity Examples

Figure 10 & Figure 11 illustrates representative prediction scenarios from the 18 individual axle forecasts:

Figure 10: Example forecast showing no threshold exceedance (M4004 Axle 1). Current damage levels of 116-142 kN with slight negative trends result in no predicted threshold crossings within the 3-year horizon, representing the majority case across the fleet.

Page 40

Figure 11: Example forecast showing threshold exceedance (M4008 Axle 3). Positive damage trends (0.1456 kN/day left, 0.1338 kN/day right) from current levels of 124-128 kN lead to predicted 275 kN threshold exceedance in Year 3, representing the single replacement case identified.

No threshold exceedance (majority case): Most axles, such as M4004 Axle 1 (Figure 10), show slight negative trends (-0.0132 to -0.0185 kN/day), resulting in no predicted threshold crossings within the 3–year forecasting horizon.

Threshold exceedance case: M4008 Axle 3 (Figure 11) demonstrates positive trends (0.1456 kN/day left, 0.1338 kN/day right) from current levels (128-124 kN), leading to predicted 275 kN threshold exceedance in the final forecast year.

Stable condition cases: Several axles like M4016 Axle 4 show minimal trends (-0.0170 to +0.0006 kN/day) from moderate baselines (112-121 kN), indicating stable long–term condition.

8.3.4 Uncertainty Quantification in Forecasting

The Monte Carlo simulation reveals variation in prediction uncertainty across axles, directly reflecting their historical behavioral patterns. Assets with consistent operational history (M4004 Axle 1) produce narrow 80% confidence intervals suitable for reliable maintenance planning, while axles showing inconsistent damage (M4008 Axle 3) generate appropriately wider uncertainty bounds. This asset–specific uncertainty quantification enables different maintenance strategies, where high–confidence predictions support standard scheduling while high–uncertainty predictions indicate the need for enhanced monitoring protocols.

8.4 Feature Significance Analysis

The feature importance analysis reveals a clear hierarchy of predictive factors (methodology described in Section 5.8.3), with dynamic behavior dominating the prediction model (Figure 12).

Figure 12: Feature importance analysis showing the relative contribution of different input features to the XGBoost model predictions. Dynamic damage indicators dominate the model with 58.3% importance, followed by baseline damage levels and rolling statistics.

8.4.1 Primary Predictive Features

1. Dynamic Behavior (58.3% importance): The difference between peak and mean damage values turned out to be the dominant predictor across both wheel models, confirming that localized damage severity is the primary indicator of wheel condition

2. Baseline Damage Level (25.2%): Historical damage levels provide essential context for interpreting dynamic spikes

3. Rolling Statistics (10.7%): Temporal smoothing of dynamic values captures medium-term degradation trends

4. Load Factors (3.8%): Axle weight influences damage progression but represents a secondary effect

5. Asset-Specific Factors (0.9%): Vehicle and axle encoding show minimal but measurable influence

8.5 Model Consistency Analysis

Both left and right wheel models demonstrate consistent feature importance patterns, with dynamic damage indicators dominating predictions (59.3% for left wheels, 57.3% for right wheels). This consistency validates the global modeling approach and confirms that wheel damage mechanisms are similar across wheel positions.

The strong agreement between models (difference of only 2.0% for dynamic behavior importance) provides confidence in the reliability of the feature importance rankings and supports the unified analysis approach used throughout this study.

Primary monitoring focus: The dominance of dynamic features (58.3%) indicates that real–time spike detection should be the cornerstone of any monitoring system. Maintenance schedules should prioritize wheels showing increasing peak-to-mean differences rather than absolute damage levels alone.

Secondary indicators: Mean damage values (25.2%) provide essential baseline context, suggesting that maintenance decisions should consider both current spike severity and historical condition trends.

Load management: The modest influence of axle weight (3.8%) suggests that while loading affects damage progression, it is not the primary driver of replacement timing decisions.

8.6 Model Residuals and Uncertainty Analysis

The residuals analysis (Figure 13) reveals well-behaved prediction errors with minimal systematic bias:

Figure 13: Residuals analysis showing prediction errors for both wheel models. The approximately normal distribution with minimal systematic bias indicates good model performance, though right wheel predictions show higher variability.

The residual analysis validates the regression metrics described in Section 5.8.2:

8.6.1 Left Wheel Model Residuals

- Mean residual: -0.002 kN (essentially unbiased)
- Standard deviation: 0.076 kN
- Range: -0.402 to +0.555 kN
- Distribution: Approximately normal with minimal skewness

8.6.2 Right Wheel Model Residuals

- Mean residual: +0.004 kN (minimal bias)
- Standard deviation: 0.361 kN (higher variability than left wheel)

- Range: -3.673 to +3.277 kN
- Distribution: Normal with some outliers at extreme values

The larger residuals for right wheels suggest either inherently more complex damage patterns or potential systematic differences in measurement conditions between wheel positions.

8.7 Validation Against Linear Baseline

Linear regression analysis provides validation context for the XGBoost predictions and demonstrates how historical trends inform future forecasting. The representative examples (Figures 8 and 9) show how different historical patterns lead to different maintenance outcomes:

- Modest average degradation rates: <0.01 kN/day
- High variability between individual axles

M4004 Axle 1 exemplifies the majority scenario where decreasing historical trends (-0.0132 and -0.0185 kN/day) from moderate baseline levels (142 and 116 kN) result in continued so called "improvement", leading to no maintenance requirements within the 3-year forecast horizon (Figure 10).

M4008 Axle 3 represents the minority deteriorating case where positive historical trends (0.1456 and 0.1338 kN/day) from similar baseline levels (128 and 124 kN) drive the system toward the 275 kN threshold, resulting in the single predicted replacement out of 18 axles for a 3–year forecasting period (Figure 11).

These examples demonstrate how the hybrid forecasting approach successfully translates historical linear trends into actionable maintenance predictions, with the linear component providing long–term stability while XGBoost captures complex patterns in the shorter term.

8.8 Model Interpretability

Understanding why specific features influence wheel replacement predictions requires examining the causal mechanisms linking sensor measurements to wheel condition.

Dynamic damage indicators (58.3% importance) directly predict replacement needs because higher peak–minus–mean values indicate localized wheel defects that create force spikes during detector passage. As damage severity increases, these spikes become more pronounced, providing a direct pathway from feature value to replacement necessity.

Baseline damage levels (25.2% importance) influence predictions by providing context for interpreting dynamic spikes - the same dynamic value has different implications depending on the underlying wheel condition. Rolling statistics (10.7% importance) capture degradation trends that distinguish temporary anomalies from progressive deterioration requiring intervention.

While XGBoost feature importance reveals which features matter most, it cannot explain individual prediction decisions. Advanced explainable AI techniques such as SHAP (SHapley Additive exPlanations) could provide instance–level explanations, enabling maintenance teams to understand specific factor contributions for each wheel. Studies like Li [36] demonstrated successful application of SHAP for interpreting XGBoost models, showing how local interpretation methods can extract meaningful explanations from complex ensemble models. Similar application could enhance Green Cargo's maintenance decision–making by providing detailed rationale for individual wheel replacement recommendations.

9. Discussion

9.1 Research Question Answers

9.1.1 RQ1: Machine Learning Application to RFID Sensor Data

How can Machine Learning algorithms be applied to sensor data, specifically force impact data, to predict the need for wheel replacement??

This study demonstrates that machine learning can successfully process DPC detector data to predict wheel replacement needs, but with important methodological considerations. The global XGBoost approach proved essential for handling the heterogeneous fleet data, achieving exceptional performance metrics ($R_2 > 0.99$) by leveraging all 15,000 measurements while preserving asset–specific patterns through categorical encoding.

The finding of the dynamic damage indicator (Peak − Mean) as the dominant feature (58.3% importance) directly validates the theoretical framework established in Section 2.. As described in Section 2.3.1, the conical wheel geometry creates concentrated contact points that produce "specific force patterns" where "damaged wheels create irregular patterns as different surface conditions pass over the detectors." The model's reliance on this feature confirms that ML can effectively capture these physics–based damage signatures.

The usage of Isolation Forest for anomaly enhancement rather than removal addresses a critical challenge in railway maintenance: rare but potentially critical pre–failure conditions. By increasing the representation of the most extreme 5% of measurements, the approach ensures the model learns from unusual patterns that might represent the "irregular patterns" characteristic of damaged wheels described in the Section 2..

However, the application faces significant limitations related to data availability and validation challenges, which will be discussed in detail below.

9.1.2 RQ2: How can the most significant features be extracted to optimize wheel replacement prediction?

This study demonstrates a systematic feature extraction and optimization methodology through three key approaches:

Feature Engineering Process: Significant features were extracted through physics–based engineering, creating dynamic damage indicators (peak–minus–mean) that capture the fundamental wheel–rail contact mechanics. This engineering approach transforms raw sensor measurements into meaningful predictive signals.

Global Modeling for Feature Assessment: The global modeling strategy enables robust feature importance evaluation by leveraging data from all assets while preserving individual characteristics through categorical encoding. This approach provides sufficient data volume for reliable feature ranking despite limited individual asset histories.

XGBoost-Based Feature Ranking: The model's built–in feature importance calculation during tree construction provides quantitative, data–driven identification of predictive factors. This method ranks features based on their actual contribution to prediction accuracy rather than assumptions.

Validation Through Performance: Feature significance was validated through exceptional model performance ($R_2 > 0.99$), confirming that the extraction methodology successfully identified meaningful predictive patterns.

The methodology reveals that dynamic damage indicators dominate predictions (58.3% importance), followed by baseline conditions (25.2% importance) and temporal patterns (10.7% importance), providing clear guidance for maintenance system optimization.

9.2 Comparison with Current Green Cargo Practices

The results reveal notable differences between the ML–based predictive approach and Green Cargo's current threshold–based system described in Section 2.. Currently, Green Cargo operates with Trafikverket's regulatory limits (425 kN mandatory replacement, 350 kN warning) in a reactive framework that "only responds to exceedances of predefined limits, rather than forecasting when failures are likely to occur."

The predictive analysis identifies only one axle requiring replacement within three years at the 275 kN early warning threshold, while no axles approach the current 350-425 kN operational thresholds. This suggests either:

- The fleet condition is significantly better than anticipated based on the "significant wear and tear" described in the background
- The predictive approach enables maintenance scheduling well before current reactive triggers

The ML approach provides several advantages over the current threshold–based system:

- Proactive forecasting: 3–year maintenance schedules versus reactive threshold responses
- Uncertainty quantification: Confidence intervals for maintenance planning versus binary threshold alerts
- Pattern recognition: Complex damage signatures versus simple threshold exceedance
- Fleet-wide optimization: Coordinated maintenance scheduling versus individual asset responses

9.3 Physical System Validation

The results strongly validate the theoretical understanding of wheel-rail dynamics established in the background 2.. The dominance of dynamic features (58.3%) aligns perfectly with the physics of limited wheel–rail contact, where "localized damage or wear" creates detectable force spikes even when damage occurs on "portions of the wheel that don't make regular contact with the rail" during individual passes.

The hybrid forecasting approach addresses the background observation that "multiple measurements needed" by combining short–term XGBoost pattern recognition with long–term linear trend extrapolation. This methodology effectively handles the geometric constraint where "wheel damage might not always align with the sensor during a single pass."

The identification of M4008 Axle 3 requiring replacement demonstrates that the approach can detect "irregular patterns as different surface conditions pass over the detectors," validating the theoretical framework while providing actionable maintenance insights.

9.4 Critical Limitations and Unexpected Findings

9.4.1 Data Availability and Validation Challenges

The most significant limitation of this study is the lack of actual wheel replacement records for validation. This limitation is particularly critical given the background emphasis on the Transmontana locomotives' "metal-against-metal" braking systems that "cause significant wear and tear on the wheels." The analysis was designed around the expectation that these systems would "generate substantial wear and tear data, offering a rich dataset for modeling and analysis" with "pronounced wear" providing "clear signals of degradation".

Page 46

9.4.2 Temporal and External Factors Influencing Degradation Predictions

While the dataset includes approximately 834 data points per axle, providing a robust basis for model training, it may not fully capture the long–term degradation cycles or seasonal variations expected under the described harsh operational conditions. Although the models leverage all available data, the temporal span of the observations may restrict their ability to predict patterns manifesting over significantly longer operational periods.

Additionally, the relatively low predicted replacement rate, only 1 axle across 18 axles over three years, may result from factors beyond the scope of this thesis, including:

- Operational Factors: Current practices, such as effective load balancing and preemptive maintenance, may mitigate wear more effectively than initially anticipated.
- Measurement Limitations: The geometric constraints of wheel–rail contact may prevent detection of certain damage types, potentially underestimating degradation.
- External Influences: Unaccounted factors such as weather conditions, track quality variations, and seasonal temperature effects could significantly impact damage progression.

These considerations highlight the importance of incorporating more temporal data and broader contextual variables in future analyses to ensure the predictions align more closely with long–term operational realities.

9.4.3 Model Performance Interpretation

While the extremely high R_2 values (>0.99) appear exceptional, they warrant careful interpretation. The strong cross–validation performance (0.9975 and 0.9873) and substantial training dataset (15,000 measurements) support genuine predictive capability rather than overfitting. However, these metrics may reflect the relative stability of wheel condition within the observation period rather than the model's ability to predict actual failure events.

The exceptional performance might indicate that wheel degradation follows more predictable patterns than anticipated, or conversely, that the current dataset lacks the variability expected from "significant wear and tear" operations.

9.5 Implications for Railway Maintenance

9.5.1 Strategic Maintenance Planning

The results suggest significant opportunities for Green Cargo to potentially transition from reactive threshold–based maintenance to proactive forecasting. The ability to generate 3–year maintenance schedules with quantified uncertainty enables:

- Resource optimization: Coordinated scheduling of maintenance activities across the fleet
- Inventory management: Predictable wheel replacement needs for supply chain planning
- Operational continuity: Advance scheduling to minimize service disruptions
- Cost reduction: Elimination of emergency maintenance through proactive intervention

9.5.2 Sensor System Optimization

The feature importance findings provide guidance for optimizing Green Cargo's monitoring investments. The dominance of dynamic damage indicators suggests that sensor systems should prioritize:

- High-resolution force spike detection over absolute load monitoring
- Pattern recognition capabilities for identifying irregular contact signatures

Page 47

- Historical trend analysis rather than instantaneous threshold comparison
- Integration of multiple measurement points to address geometric detection limitations

9.5.3 Threshold Recalibration Needs

The minimal maintenance requirements predicted within Trafikverket's threshold framework suggest potential opportunities for threshold optimization. Green Cargo could consider:

- Implementation of lower early-warning thresholds (e.g., 275 kN) for enhanced lead time
- Development of asset–specific thresholds based on operational history and model predictions
- Integration of uncertainty bounds into threshold-based decision making

9.6 Achievement of Research Objectives

This study achieved its primary objective of developing a machine learning approach for predictive wheel maintenance in railway operations. Both research questions were answered:

RQ1 was addressed through the successful implementation of a global XGBoost modeling strategy that processes DPC sensor data to predict wheel replacement needs, validated through performance metrics and practical forecasting outcomes.

RQ2 was answered through systematic feature extraction methodology combining physics-based engineering, global modeling assessment, and data-driven ranking techniques, successfully identifying the most predictive factors for wheel replacement optimization.

The hybrid forecasting methodology enables Green Cargo to potentially transition from reactive threshold–based maintenance to proactive 3–year maintenance planning, directly addressing the stated purpose of improving decision–making and reducing operational risks.

However, the absence of actual failure validation data means that while the technical objectives were achieved, operational validation remains incomplete. The study demonstrates the feasibility and potential of ML–based predictive maintenance but requires additional validation against confirmed failure events.

This study contributes several methodological insights to railway predictive maintenance:

9.6.1 Global Modeling Strategy

The global modeling approach addresses the challenge of insufficient individual asset data while preserving asset–specific learning through categorical encoding. This strategy could be applied to other fleet–based maintenance challenges.

9.6.2 Hybrid Forecasting Framework

The combination of XGBoost pattern recognition with linear trend extrapolation provides both short–term accuracy and long–term stability. The time–dependent weighting scheme offers a practical solution for long–horizon forecasting in equipment condition monitoring.

9.6.3 Anomaly Enhancement Strategy

The decision to enhance rather than remove anomalies addresses a critical challenge in safety–critical systems where rare events may represent the most important learning opportunities. This approach could inform anomaly handling strategies in other predictive maintenance applications.

9.7 Future Research Directions

9.7.1 Validation Against Actual Failures

Future research should prioritize obtaining actual wheel replacement and failure records to validate prediction accuracy against real operational outcomes rather than threshold exceedance. This would enable optimization of both model parameters and threshold values based on confirmed failure events.

9.7.2 Extended Temporal Coverage

Longer observation periods (3-5 years) would enable validation of long-term forecasting accuracy. Extended monitoring would also capture seasonal and operational variations not present in the current dataset.

9.7.3 Multi-Modal Sensor Integration

Integration of additional sensor types (vibration, temperature, acoustic) could address the geometric limitations of force–based detection and provide more comprehensive condition assessment. This would particularly benefit detection of damage types that may not create consistent force signatures.

9.7.4 Operational Context Integration

Future models should incorporate operational variables (weather, track conditions, loading patterns, operational changes) to improve prediction accuracy and provide more comprehensive understanding of degradation factors.

9.8 Broader Applicability and Generalization

The methodological contributions of this study extend well beyond Green Cargo’s specific operational context, offering several generalizable approaches for predictive maintenance in asset–intensive industries:

9.8.1 Fleet-Based Asset Management

The global modeling strategy addresses a challenge across industries where individual asset histories are insufficient for reliable predictions. This approach could, for example, be applied to:

- Aircraft engine maintenance in aviation fleets
- Wind turbine blade monitoring across wind farms
- Mining equipment maintenance in large operations
- Maritime vessel component management

9.8.2 Sensor-Based Condition Monitoring

The hybrid forecasting framework combining ML pattern recognition with linear trend extrapolation provides a robust approach for long–horizon predictions in any sensor–based monitoring system. The methodology is particularly applicable to industries where:

- Equipment operates under similar conditions but with asset–specific variations
- Sensor measurements are irregular or infrequent
- Geometric or operational constraints limit sensor effectiveness

9.8.3 Safety-Critical Systems

The anomaly enhancement strategy offers valuable insights for safety–critical applications where rare events may represent the most important learning opportunities. This approach has relevance for example these areas:

- Nuclear facility monitoring systems
- Medical device predictive maintenance
- Infrastructure health monitoring (bridges, tunnels)
- Industrial safety system maintenance

9.8.4 Threshold-Based to Predictive Transitions

Many industries currently rely on threshold–based maintenance similar to Green Cargo’s approach. The demonstrated transition methodology provides a framework for organizations seeking to evolve from reactive to predictive maintenance while maintaining regulatory compliance and operational safety.

9.8.5 Railway Network Scalability

The methodology demonstrates strong potential for scaling across Green Cargo’s entire fleet and broader railway operations, as the standardized DPC detector infrastructure provides consistent data collection regardless of operational context.

The approach can be directly applied to all Green Cargo wagons and locomotives operating across different routes, as the DPC detectors collect identical data types (peak load, mean load, timestamps) regardless of location. The global modeling strategy successfully integrates data from multiple assets while preserving asset–specific characteristics through categorical encoding, enabling seamless expansion to the entire fleet without requiring route–specific model development.

For high–speed passenger services, the core methodology remains applicable since these systems utilize the same DPC detector infrastructure and data collection protocols. While high–speed trains with advanced braking systems generate different force magnitude patterns, the fundamental data structure and feature engineering approaches remain unchanged. The higher operational frequency could potentially enhance model training through increased data density.

The methodology can be scaled across European railway networks where standardized DPC detector infrastructure exists. Since the data collection methodology is consistent across different railway systems, the same global modeling approach can be applied to international operations, enabling coordinated predictive maintenance strategies across borders while leveraging the standardized sensor infrastructure.

The key advantage is that the data structure remains constant across all applications - only the actual force values differ based on operational conditions, making the methodology highly scalable without fundamental changes to the approach.

The study's emphasis on uncertainty quantification and confidence intervals addresses a critical need in maintenance decision–making across industries where safety and reliability are paramount, making the approach broadly applicable beyond railway operations.

For Green Cargo to implement this predictive maintenance approach effectively, several practical considerations must be addressed:

9.8.6 Data Infrastructure

Implementation requires robust data collection, processing and analysis infrastructure capable of handling real–time sensor data across the fleet while maintaining the computational efficiency demonstrated (35-minute processing time).

Page 50

9.8.7 Integration with Existing Systems

The predictive approach must integrate with Green Cargo's current maintenance management systems while complementing rather than replacing regulatory threshold monitoring required by Trafikverket.

9.8.8 Regulatory Compliance

Any implementation must maintain compliance with Trafikverket's mandatory threshold limits while leveraging predictive capabilities for enhanced operational efficiency within regulatory constraints.

9.9 Limitations and Critical Assessment

Several limitations warrant acknowledgment:

9.9.1 Potential Overfitting Concerns

The extremely high R_2 values, while supported by cross–validation and the substantial 15,000-measurement training dataset, may indicate that the current dataset exhibits more regularity than typical operational conditions. However, the global modeling approach provides sufficient data volume to support the model complexity and the strong cross–validation performance (CV R_2 of 0.9975 and 0.9873) suggests genuine predictive capability rather than overfitting.

9.9.2 Impact of Operational Variability on Model Robustness

Several operational factors not captured in this study could significantly affect model robustness. Seasonal weather conditions such as winter snow, ice, and temperature variations alter wheel–rail contact characteristics and may create different damage patterns than those observed in the current dataset. Similarly, varying braking patterns across different operational scenarios - including emergency braking frequency and different braking intensities - could produce wheel wear characteristics not represented in the current analysis.

Route–specific characteristics of the Borlänge-Boden line present another limitation. Different routes have distinct gradient profiles, curve distributions, track conditions, and operational speeds that create unique damage progression patterns. The current model's training on a single route limits its exposure to this diversity of operational stresses.

The limitation to three locomotives operating a single route significantly constrains the model's generalization capabilities. While this approach provides deep insights into specific operational conditions, it restricts exposure to the broader spectrum of locomotive behaviors, operational practices, and infrastructure variations present in railway systems.

Additional data sources could substantially improve generalization. Incorporating data from multiple routes with diverse track characteristics would expose the model to different operational stresses. Extended temporal coverage spanning multiple seasons would capture weather–related variations in damage patterns. Including locomotives with different operational roles, braking patterns, and maintenance histories would enhance the model's ability to handle diverse fleet characteristics. Integration of weather data and route–specific information could provide additional context for understanding damage progression variability across different operational environments.

9.10 Practical Implementation Readiness

The results demonstrate significant practical value for Green Cargo's maintenance operations:

- Computational efficiency: 35-minute total processing time enables regular fleet assessment

Page 51

- Actionable outputs: Clear maintenance schedules with confidence intervals
- Uncertainty quantification: Monte Carlo simulation provides realistic prediction bounds
- Scalability: Global modeling approach can accommodate fleet expansion

10. Conclusions

This thesis successfully developed and validated a machine learning approach for predictive wheel maintenance in railway operations, making several important contributions to the field.

10.1 Methodological Advances

My approach introduced three main innovations. First, I developed a global modeling strategy that solves the problem of limited data for individual assets by using a single global model that captures asset–specific patterns through categorical encoding. Secondly, I created a hybrid forecasting framework that combines XGBoost pattern recognition with linear trend forecasting to achieve reliable long–term predictions. Thirdly, instead of removing extreme operational conditions from the data like most approaches do, i used an anomaly enhancement method that learns from these extreme conditions, which is crucial for safety–critical railway systems.

10.2 What I Found

The analysis revealed that dynamic damage indicators are by far the most important predictive feature, accounting for 58.3% of the model's importance. This finding provides clear guidance for optimizing sensor systems in future implementations. The model also successfully measures prediction uncertainty for individual assets, which helps maintenance teams make better risk–informed decisions. Most importantly, this approach enables a shift from reactive threshold–based maintenance to proactive 3–year forecasting with quantified uncertainty levels.

10.3 Practical Results

The model achieved excellent predictive accuracy with R_2 values above 0.99 while maintaining computational efficiency with just 35–minute processing times. When applied to real fleet data, it generated actionable maintenance schedules that identified minimal immediate needs, 1 out of 18 axles requires replacement within the next 3 years.

10.4 Future Work

Looking ahead, several areas need more work, particularly incorporating actual failure data and extending observation periods to validate the long–term forecasting capabilities. The approach could also benefit from integrating additional sensor types and operational variables. The hybrid forecasting framework shows promise for other maintenance challenges across different industries. This work provides a foundation for transforming maintenance strategies from reactive to proactive approaches across the transportation industry.

References

[1] Adortech, Challenges and innovations in railroad maintenance, Online, Available: https://adortech.com/blog/challenges-and-innovations-in-railroad-maintenance#Heavy-Usage-and-Constant-Wear-and-Tear, Nov. 2023.

[2] A. Larsson and R.-M. Johansson, Järnvägsnätsbeskrivning 2023 (tdok 2020:0074), Accessed: 2025-04-15, 2023. [Online]. Available: https://bransch.trafikverket.se/contentassets/a2b42dceee5f45968c50dba050588793/tdok_2020-0074_jnb2023.pdf.

[3] R. K. Mobley, An Introduction to Predictive Maintenance, Second. Burlington, MA, USA: Butterworth-Heinemann, 2002. [Online]. Available: https://books.google.se/books?hl=sv&id=SjqXzxpAzSQC.

[4] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes and G. Elger, 'Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry,' Reliability Engineering and System Safety, vol. 215, p. 107864, 2021. doi: 10.1016/j.ress.2021.107864.

[5] C. Yang and S. Létourneau, 'Learning to predict train wheel failures,' in Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '05, Chicago, Illinois, USA: ACM, 2005, pp. 516–525, isbn: 1-59593-135-X. [Online]. Available: https://www.researchgate.net/publication/221654313_Learning_to_Predict_Train_Wheel_Failures.

[6] H. Wang, T. Men and Y. Li, 'Transformer for high-speed train wheel wear prediction with multiplex local–global temporal fusion,' IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–12, 2022. doi: 10.1109/TIM.2022.3154827.

[7] T. Chen and C. Guestrin, 'XGBoost: A scalable tree boosting system,' in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, isbn: 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785.

[8] S. Weisberg, Applied Linear Regression, 3rd. Hoboken, NJ: Wiley, 2005, isbn: 978-0-471-66379-4.

[9] F. T. Liu, K. M. Ting and Z.-H. Zhou, 'Isolation forest,' in 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[10] Järnväg.net, Mb locomotive, Image credited to Green Cargo. [Online]. Available: https ://www.jarnvag.net/lokguide/mb.

[11] Trafikverket, Trafikverkets detektorsystem uhtto, https : / / bransch . trafikverket . se / contentassets/aa399d15a6d94be48023a6efc62b1aea/trafikverkets_detektorsystem_uhtto_241118.pdf, 2018.

[12] Trafikverket, Image of dynamic pressure check (dpc) detector, Online, 2025. [Online]. Available: https://bransch.trafikverket.se/for-dig-i-branschen/teknik/anlaggningsteknik/Detektorer/.

[13] K. Mattsson, 'Wheel-rail impact loads generated by wheel flats: Detector measurements and simulations,' Master's thesis in Mobility Engineering, Chalmers University of Technology, Gothenburg, Sweden, 2023. [Online]. Available: http://hdl.handle.net/20.500.12380/306554.

[14] A. L. Samuel, 'Some studies in machine learning using the game of checkers,' IBM Journal of research and development, vol. 3, no. 3, pp. 210–229, 1959.

[15] J. V. Doshi and K. K. Shah, 'Enhancing industrial predictive maintenance through anomaly detection in multivariate sensor data using machine learning techniques,' Journal of Engineering Science, vol. 17, pp. 120–132, 2022. [Online]. Available: https://journal.esrgroups.org/jes/article/view/5399.

[16] C. McKinnon et al., 'Investigation of isolation forest for wind turbine pitch system condition monitoring using scada data,' Energies, vol. 14, no. 20, p. 6601, 2021. doi: 10.3390/en14206601. [Online]. Available: https://www.mdpi.com/1996-1073/14/20/6601.

[17] I. Belcic and C. Stryker, What is supervised learning? https://www.ibm.com/think/topics/supervised-learning, 2024.

[18] IBM, What is unsupervised learning? https://www.ibm.com/think/topics/unsupervised-learning, 2021.

[19] T. Chen and C. Guestrin, 'Xgboost: A scalable tree boosting system,' in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794. doi: 10.1145/2939672.2939785. [Online]. Available: https://doi.org/10.1145/2939672.2939785.

[20] Z. Kang, C. Catal and B. Tekinerdogan, 'Remaining useful life (rul) prediction of equipment in production lines using artificial neural networks,' Sensors, vol. 21, no. 3, p. 932, 2021. doi: 10.3390/s21030932.

[21] J. H. Friedman, 'Greedy function approximation: A gradient boosting machine,' Annals of statistics, pp. 1189–1232, 2001.

[22] B. Taşcı, S. Alemdaroğlu, O. Bulut and D. Ö. Duru, 'Remaining useful lifetime prediction for predictive maintenance in manufacturing,' Computers & Industrial Engineering, vol. 184, p. 109 566, 2023. doi: 10.1016/j.cie.2023.109566. [Online]. Available: https://doi.org/10.1016/j.cie.2023.109566.

[23] J. Bergstra and Y. Bengio, 'Random search for hyper-parameter optimization,' Journal of Machine Learning Research, vol. 13, no. 1, pp. 281–305, 2012. [Online]. Available: https://www.jmlr.org/papers/v13/bergstra12a.html.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort et al., 'Scikit-learn: Machine learning in python,' Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[25] M. Kuhn and K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press, 2019.

[26] T. O. Hodson, 'Root-mean-square error (rmse) or mean absolute error (mae): When to use them or not,' Geoscientific Model Development, vol. 15, no. 14, pp. 5481–5487, 2022. doi: 10.5194/gmd-15-5481-2022. [Online]. Available: https://doi.org/10.5194/gmd-15-5481-2022.

[27] C. Lewis-Beck and M. Lewis-Beck, Applied regression: An introduction. Sage publications, 2015, vol. 22.

[28] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd. Springer, 2009.

[29] C. P. Robert and G. Casella, Monte Carlo Statistical Methods, 2nd. Springer, 2004.

[30] S. Majidiparast, R. Neamatian Monemi and S. Gelareh, 'A graph convolutional network for optimal intelligent predictive maintenance of railway tracks,' Decision Analytics Journal, vol. 14, p. 100 542, 2024. doi: 10.1016/j.dajour.2024.100542. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772662224001462.

[31] I. Daniyan, K. Mpofu, M. Oyesola and B. Ramatsetse, 'Artificial intelligence for predictive maintenance in the railcar learning factories,' Procedia Manufacturing, vol. 45, pp. 13–18, 2020. doi: 10.1016/j.promfg.2020.04.032.

[32] A. Shangguan, G. Xie, R. Fei, L. Mu and X. Hei, 'Train wheel degradation generation and prediction based on the time series generation adversarial network,' Reliability Engineering & System Safety, vol. 229, p. 108 816, 2023. doi: 10.1016/j.ress.2022.108816.

[33] N. Davari, B. Veloso, G. d. A. Costa, P. M. Pereira, R. P. Ribeiro and J. Gama, 'A survey on data-driven predictive maintenance for the railway industry,' Sensors, vol. 21, no. 17, p. 5739, 2021. doi: 10.3390/s21175739.

Page 55

[34] M.-H. Le-Nguyen, F. Turgis, P.-E. Fayemi and A. Bifet, 'Real-time learning for real-time data: Online machine learning for predictive maintenance of railway systems,' Transportation Research Procedia, vol. 72, pp. 171–178, 2023. doi: 10.1016/j.trpro.2023.11.391.

[35] R. Kunst, L. Avila, A. Binotto, M. Schmitt, S. Bocker and C. Rolim, 'Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges,' Computers in Industry, vol. 123, p. 103 298, 2020. doi: 10.1016/j.compind.2020.103298.

[36] Z. Li, 'Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost,' Computers, Environment and Urban Systems, vol. 96, p. 101 845, 2022. doi: 10.1016/j.compenvurbsys.2022.101845. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0198971522000898.