Time series data mining for railway maintenance

Afonso Pinho Lourenço

Master's thesis

Supervisor at FEUP: Prof. Luís Guimarães
Supervisor at GECAD: Prof. Goreti Marreiros

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Master in Industrial Engineering and Management

2023-06-26

Watch your signal, for it becomes data;
watch your data, for it becomes information;
watch your information, for it becomes knowledge;
watch your knowledge, for it becomes wisdom;
watch your wisdom, for it becomes your future.

Time series data mining for railway maintenance

Abstract

By integrating advanced sensor technologies and leveraging machine learning algorithms, artificial intelligence has brought about a revolution in wayside and on-board monitoring within the railway sector. While various algorithms have been put forward to detect track irregularities and wheel out-of-roundness, they often prove inadequate when applied in practical scenarios. This thesis focuses on the challenges and opportunities of using artificial intelligence for condition-based maintenance in railway systems, particularly at the wheel-track interface. To achieve this goal, the thesis investigates the most common techniques for data acquisition and how to extract knowledge from sensor signals with machine learning algorithms and effective pre-processing techniques for time series. Additionally, it explores how maintenance models are adapted to account for streaming data and automated maintenance decision making, with a comprehensive evaluation of performance, computational expense, interpretability, and cost considerations to improve trust and adoption by maintenance personnel. Lastly, the thesis investigates the key opportunities and challenges in scaling up the use of artificial intelligence for condition-based maintenance in railway systems and proposes strategies that can be employed to overcome these challenges. These questions were addressed through a comprehensive literature review, and novel efficient computer algorithms were developed to address the identified research gaps. The algorithms offer comprehensive fault diagnosis, parallelized implementation, automatic hyperparameter tuning, and processing techniques for data streaming environments. Rrenesults show the proposed models successfully rank the defects of individual wheels, establishing a useful priority sequence for replacement interventions. As these methods were validated with physical-based modelling, the next efforts should lie in ensuring there are not too many unfulfilled assumptions and rough approximations. Thus, future work includes a field trial to further evaluate the proposed methodology based on on-site measurements. Furthermore, the robustness and efficiency of the methodology will be assessed under distinct track environments, particularly in the presence of bridges, tunnels and other under passing structures.

Extração de conhecimento de dados em séries temporais para monitorização de condições ferroviárias

Através da integração de tecnologias avançadas de sensores e algoritmos de aprendizagem automática, a inteligência artificial revolucionou a monitorização na via e a bordo no setor ferroviário. Apesar de terem sido propostos vários algoritmos para detetar irregularidades nas vias e desgaste irregular das rodas, estes muitas vezes revelam-se insuficientes quando aplicados em cenários reais. Esta tese centra-se nos desafios e oportunidades do uso da inteligência artificial para manutenção preditiva em sistemas ferroviários, particularmente na interface roda-carril. Para atingir este objetivo, a tese investiga as técnicas mais comuns para aquisição de dados e como extrair conhecimento a partir dos sinais de sensores com algoritmos de aprendizagem automática e técnicas eficazes de pré-processamento para séries temporais. Além disso, explora como os modelos de manutenção são adaptados para lidar com dados em tempo real e tomada de decisão automatizada, com uma avaliação abrangente de desempenho, custos computacionais, interpretabilidade e considerações de custo para melhorar a confiança e adoção pelos técnicos de manutenção. Por fim, a tese investiga as principais oportunidades e desafios na ampliação do uso da inteligência artificial para manutenção preditiva em sistemas ferroviários e propõe estratégias para superar esses desafios. Estas questões foram abordadas através de uma revisão bibliográfica abrangente e do desenvolvimento de novos algoritmos computacionais eficientes para abordar as lacunas identificadas na literatura. Os algoritmos oferecem diagnóstico abrangente de falhas, implementação paralelizada, ajuste automático de hiperparâmetros e técnicas de processamento para ambientes de dados em streaming. Os resultados evidenciam que os modelos propostos classificam com sucesso as deficiências das rodas individuais e via férrea, estabelecendo uma sequência de prioridade útil para intervenções de substituição. Uma vez que estes métodos foram validados com base em modelagem física, esforços futuros devem garantir de que não existam demasiadas pressuposições não cumpridas e aproximações grosseiras. Assim, o trabalho futuro inclui um ensaio de campo para uma avaliação mais aprofundada da metodologia proposta, com base em medições no local. Além disso, será feita uma avaliação da robustez e eficiência da metodologia em diferentes ambientes de via, nomeadamente na presença de pontes, túneis e outras estruturas de passagem subterrânea.

Acknowledgments

It gives me an immense pleasure to present the thesis report in its completed form.

The present work has been developed under project FERROVIA 4.0 (POCI/01/0247/FEDER/046111) and WAY4SafeRail (NORTE/01/0247/FEDER/069595). It has received Portuguese National Funds through FCT (Portuguese Foundation for Science and Technology) under project UIDB/00760/2020.

First of all, I would like to thank my main supervisor Prof. Goreti Marreiros for her guidance, but more importantly for her moral support through this research work.

Without her timely advice and through knowledge in machine learning, the research would not have been accomplished with such a great success. I am extremely thankful for her support.

I thank my co-supervisor Prof. Luís Guimarães for his support and help. He guided me in this endeavor with just the right mixture of affirmation and criticism.

I thank my colleagues PhD Candidates, Marta Fernandes and Jorge Meira, for proofreading parts of this work.

I thank Master student Carolina Ferraz who came to GECAD to work together in fault diagnosis and maintenance. I also thank my colleagues in CONSTRUCT-LESE for provision and support on the data.

Last, but the most important, I am extremely grateful to my parents and friends for being there for me all these years, for their patience, support and encouragement.

Porto, June 2023

Afonso Lourenço

Table of Contents

Table of Figures

Table of Tables

1 Introduction

The railway sector relies on reliability and maintainability to ensure the smooth and secure transportation of passengers and freight (Hidirov & Guler, 2019). With trains designed to operate for decades and maintenance costs accounting for a significant portion of the total expenditure, there is an urgent need to keep safety and quality of service at an optimal level while minimizing the operational costs. In fact, with high traffic levels, massive axle loads, and constantly changing conditions, even minor flaws in the railway track can snowball into significant problems. In particular, the wheel-rail interface has received a lot of attention in the literature, as it incurs most of the cost of maintenance for both railway vehicles and infrastructure (Fröhling & Hettasch, 2010). Examples of such components include wheels, rails, sleepers, fastenings, and ballast, among others. These components are exposed to various environmental and operational conditions, which can cause corrosion, cracks and other forms of damage (Gonçalves et al., 2023; Guedes et al., 2023; Mosleh, Montenegro, Alves Costa, et al., 2021). In addition, surfaces of the wheel-rail are subjected to high stick, sliding, and contact stresses in rolling contact, with the increasing hazard function being a common pattern observed in the failure rates of these mechanical components (Chong et al., 2010; Mohammadi et al., 2023). To address this increasing probability of failure increase over time, it is imperative to perform maintenance actions proactively.

By implementing a preventive maintenance strategy, railway operators can reduce the likelihood of downtime, increase equipment lifespan, and avoid costly emergency repairs. The railway industry often uses a periodic preventive maintenance based either on tonnage or kilometres travelled for vehicles and accumulated load or passing traffic for tracks (Lagnebäck, 2007). However, these maintenance interventions are intuitively affected by expert knowledge on the effect of structural parameters on degradation rate. For example, change in rail profiles, closeness to switches and direct fastening systems increases the rate of degradation. In addition, the railway industry tends to have a comprehensive maintenance plan, by proactively lubricating, refurbishing, calibrating, tamping and drive-by visual inspecting equipment on a regularly scheduled basis. However, these plans tend to be too conservative to compensate for the fact that they fail to tackle the specific root cause, which not only results in very high maintenance costs but can also reduce the component life. For instance, it has been proven that despite tamping improves the condition of the track geometry, excessive ballast settlement can degrade the structural parameters of railway systems (D. Barke & Chiu, 2005). Ideally, the maintenance action should be performed just before the machine failure. To reach this goal, both onboard and wayside condition monitoring are increasingly embraced with analysis tools to evaluate the railway load (Alemi et al., 2017), as illustrated in Figure 1.

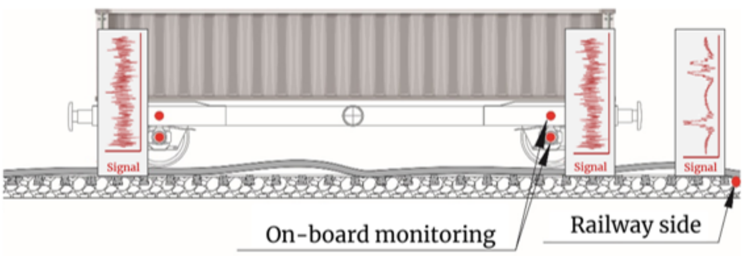


Figure 1 – Types of condition monitoring for proactive maintenance: on-board and railway side

The condition of a system is quantified by obtaining periodically or even continually information from the sensors mounted on the components, taking actions only when there is evidence of abnormal behaviors. However, condition-based maintenance strategies are more expensive, and thus only beneficial for crucial components whose failure does not happen instantaneously and leads to severe function loss and safety risk. In recent years, the railway industry has witnessed the emergence of low-cost and easily maintainable condition monitoring systems, enabled by digital automation systems, Big Data, and Industrial Internet of Things technologies. This has led to a growing trend of using artificial intelligence for condition-based maintenance, with numerous publications exploring this approach (Xie et al., 2020).

1.1 Research questions

Despite these efforts, the implementation of data mining techniques for railway maintenance has proven challenging due to the lack of a common framework and standards. To address this issue, this thesis investigates the challenges and opportunities of employing artificial intelligence for condition-based maintenance in railway systems, with a particular emphasis on the wheel-track interface, and to propose potential solutions to address these challenges. With this central research question in mind, the following specific questions (SQ) were posed:

(SQ1) What are the most common types of sensors used to collect time series data in railway systems, particularly for the wheel-track interface?

(SQ2) What are the most used machine learning algorithms for condition-based maintenance in railway systems, and how do they compare in terms of accuracy and computational efficiency?

(SQ3) What are the most effective pre-processing techniques for time series data in the context of railway maintenance, and how can they be optimized for different types of sensors and maintenance tasks?

(SQ4) How are artificial intelligence-based maintenance models adapted to account for streaming data and automated maintenance decision making?

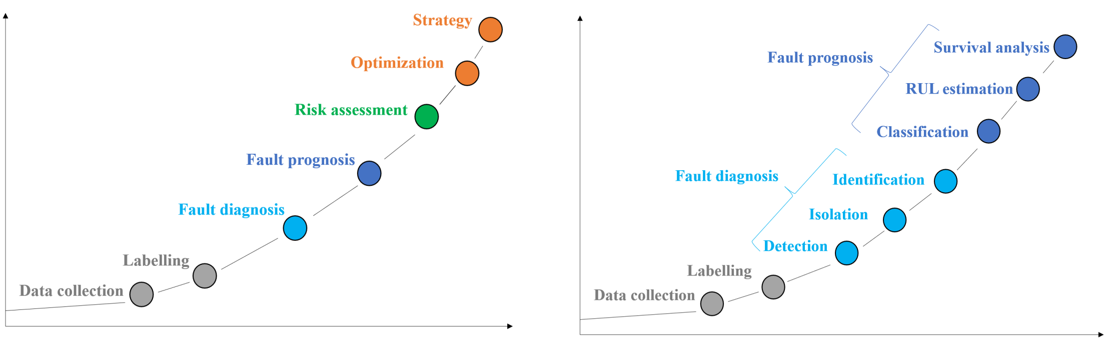
(SQ5) How is computational expense, interpretability and cost considerations incorporated into artificial intelligence-based maintenance models to improve trust and adoption by maintenance personnel and decision makers?

(SQ6) What are the key opportunities and challenges in scaling up the use of artificial intelligence for condition-based maintenance in railway systems, and what strategies can be employed to overcome these challenges?

These questions were addressed in the form of a comprehensive literature review. Furthermore, leveraging on the identified research gaps, novel efficient computer algorithms were developed with characteristics that only a few works in the current literature have achieved, namely:

- comprehensive fault diagnosis with the three necessary objectives: (1) detection of aberrant train behavior; (2) isolation of specific defective wheels; (3) identification of the severity;
- parallelized implementation of machine learning algorithms that scale well to very large databases, enabling efficient data processing across multiple clusters;
- automatic hyperparameter tuning process, saving time, and reducing the potential for errors, while making the algorithm accessible to non-expert users in the machine learning field;
- processing techniques for data streaming environments that automatically update themselves and address issues such as concept drift.

The developed strategies were validated with 3D numerical train-track dynamic interaction simulations, carried out with the in-house software Vehicle-Structure Interaction Analysis (Montenegro et al., 2015). By defining unknown variables one by one, while observing how the methodology deals with the effect of these unknown parameters in the responses, the decision support system can be deemed reliable. Ultimately, the proposed methodologies aim to be generic and with successful validation have the potential to be applied to real experimental data considering different types of trains and wheel defects. Regarding the scope and aim of this thesis, it is worth noting that the maintenance paradigm encompasses a broad range of activities, from data collection to maintenance scheduling and strategic decision-making supported by cost analysis. However, this study concentrates exclusively on fault diagnosis and prognosis tasks, as illustrated in Figure 2.



(a) Maintenance paradigm in the literature (b) Scope covered in this thesis

Figure 2 – Contextualization of the thesis' scope within the maintenance paradigm

1.2 Contributions

This thesis's primary contributions were published, with the majority of the research work and writing being performed by the first author.

Online adaptive learning for out-of-round railway wheels detection

Authors: Afonso Lourenço, Jorge Meira and Goreti Marreiros

Published in Proceedings ACM SAC Conference, Tallinn, Estonia, March 27 - March 31, 2023. DOI: 10.1145/3555776.3577860

Wheel out-of-roundness is one of the most important causes of train accidents. With the Internet of Things rapidly expanding, many opportunities arise for the implementation of real-time condition-based maintenance techniques. The purpose of this work was building an online adaptive learning algorithm for wheel polygonization detection with different wavelengths. Using efficient pre-processing techniques to deal with the open-ended and non-stationary nature of an accelerometer data stream, this method automatically detected train passages, segmented groups of axles and ranked damage severity, accounting for variability in train speed and the unevenness profile of rail. Experimental results show the proposed model successfully ranks the polygonization of individual wheels, establishing a useful priority sequence for replacement interventions.

Automated green machine learning for condition-based maintenance

Authors: Afonso Lourenço, Carolina Ferraz, Jorge Meira and Goreti Marreiros

Published in Proceedings ESANN Conference, Bruges, Belgium, October 4 – October 6, 2023

Within the big data paradigm, there is an increasing demand for machine learning with automatic configuration of hyperparameters. Although several algorithms have been proposed for automatically learning time-changing concepts, they generally do not scale well to very large databases. In this context, this paper presents an automated green machine learning approach applied to condition-based maintenance with automatic data fusion and density-based anomaly detection based on locality sensitivity hashing. Experiments on numerical simulations of train-track dynamic interactions demonstrate the utility of the approach to detect railway wheel out-of-roundness. This unlocks the full potential of scalable machine learning, paving the way for environment-friendly systems and automated decision-making.

Adaptive time series representation for out-of-round railway wheels fault diagnosis in wayside monitoring

Authors: Afonso Lourenço, Carolina Ferraz, Diogo Ribeiro, Araliya Mosleh, Pedro Montenegro, Cecília Vale, Andreia Meixedo and Goreti Marreiros

Published in Engineering and Failure Analysis Journal, Special Issue in Smart Condition Assessment of Railway Infrastructure, 2023. DOI: 10.1016/j.engfailanal.2023.107433

Through the integration of advanced sensor technologies and machine learning algorithms, artificial intelligence has revolutionized wayside monitoring in the railway sector. Although several algorithms have been proposed for detecting out-of-roundness, i.e., flats, wear treads and polygonization, they generally fall short in isolating the root cause of the wheel's issue in a train passage. In this context, the paper presents a novel approach for wheel out-of-roundness diagnosis with (1) detection of aberrant train behavior; (2) isolation of specific defective wheels; (3) identification of the severity. For this, the methodology automatically segments a strain gauge signal, capturing the complex nature and temporal dependence of vibration patterns. This segmentation allows the extraction of localized accelerometer features in both the time and frequency domain, as well as implicit axle count and labelling of each wheel passage. Moreover, a single-value damage indicator based on anomaly detection algorithms was proposed. To validate the effectiveness of the proposed methodology, experiments on a set 3D numerical train-track dynamic interaction simulations are performed for different wheel profiles, track irregularities, train speeds, sensor placement and noise, associated to other environmental and operational variations. This demonstrates the potential of artificial intelligence for real-time assessment of wheels without interfering with normal service conditions, suggesting the possibility of automated fault diagnosis.

A survey on time series data mining for railway condition monitoring

Authors: Afonso Lourenço and Goreti Marreiros

Ongoing work, to be submitted to Artificial Intelligence Review Journal

1.3 Thesis outline

The present thesis is divided in 6 Chapters. This Chapter 1 introduces the topics of railway and the wheel-track interface and maintenance as drivers for the posed research questions in Section 5.1 and published contributions in Section 5.2. Chapter 2 describes the state of the art in time series data mining for railway condition monitoring, structured with 5 Sections. Section 2.1 defines the research methodology, while Section 2.2 provides an overview of the search results, in terms of related work. Section 2.3 covers the kind of data being used, namely the parts and type of defects subject to predictive maintenance (PdM), measurement methods, typical characteristics of sensor data and support with physical-based models. Section 2.4 provides a thorough overview of the approaches developed to deal with time series sensor data, with different goals, learning tasks, methods, and implementation components. Section 2.5 studies the most frequent evaluation procedures, along four dimensions: performance, interpretability, cost and data streaming. Finally, Section 2.6 concludes the literature review, with open challenges and future research directions. Chapters 3-5 zoom-in on the application of time series data mining for out-of-roundness fault diagnosis. Chapter 3 explains the process of simulating train-track dynamic interaction in Section 3.1 and numerical modelling to represent two virtual monitoring systems: wayside in Section 3.2 and on-board in Section 3.3. Chapter 4 proposes time series data mining methodologies for condition-based maintenance. Sections 4.1-4.4 provided a detailed description of the main approach for fault diagnosis. Section 4.5 provides an adaptation to streaming data with a fixed sliding window, while Section 4.6 proposes an alternative query by content algorithm based on time series numeric distance. Section 4.7 includes parallelize implementation and automatic hyperparameter tuning process. Section 8 studies algorithms intrinsically prepared for data streaming environments. Chapter 5 presents and discusses the results for the various methodologies proposed. Full details are provided for the fault diagnosis approach in Section 5.1, while the remaining methodologies are briefly discussed in Sections 5.2-5.5 to keep the thesis concise. Finally, Chapter 6 summarizes main conclusions and describes ideas for future work.

2 Literature Review

The theoretical framework presented in this chapter is paramount for the reader to grasp state of the art in machine learning applications, particularly in the realm of railway condition monitoring. Through Section 2.1, the foundation is laid for the research methodology with the Kitchenham's methodology for transparent reporting of systematic reviews and meta-analyses being used (Kitchenham, 2004). The following steps are discussed:

- Definition of research question: it guides the search for relevant papers in the literature;
- Search process: it presents the research strategy and the scientific sources used in the search;
- Selection process: definition of the criteria applied to select relevant papers.

Section 2.2 provides a comprehensive overview of related work, highlighting the main differences. Sections 2.3-2.5 describe the state of the art in time series data mining for railway condition monitoring, Section 2.3 covers the kind of data being used, namely the parts and type of defects subject to predictive maintenance, measurement methods, typical characteristics of sensor data and support with physical-based models. Section 2.4 provides a thorough overview of the approaches developed to deal with time series sensor data, with different goals, learning tasks, methods, and implementation components. Section 2.5 studies the most frequent evaluation procedures, along four dimensions: performance, interpretability, costs and data streaming. Finally, Section 2.6 concludes the literature review, with open challenges and future research directions.

2.1 Research methodology

The recent advent of digital automation systems coupled with advances in Big Data and the Industrial Internet of Things technologies greatly improves the transparency for system health condition, maintenance scheduling and sensor placement selection. Furthermore, the pace at which this development takes place is fast and has resulted in a growing number of publications in which artificial intelligence is used for condition-based maintenance. Despite the increasing efforts in applying data mining techniques for railway maintenance, the results have not been fruitful due to the absence of a common framework and standards. This comprehensive survey relies on the fact that to create an effective maintenance plan, it is necessary to ensure four requirements:

(R1) condition-based maintenance strategy is appropriate;

(R2) available data;

(R3) problem is appropriately formulated;

(R4) data mining methods are evaluated correctly.

As for (R1), it has been explained in Section 1.1 how for the mechanical components in the wheel-rail interface, condition-based maintenance is the right strategy. Regarding (R2), the dynamic context of the railway system is exceptionally challenging. Numerous measurement methods are utilized for bogies and tracks, including 3D-laser cameras, ultrasonic devices, eddy current, magnetic flux leakage, manual inspection, and others. As the data available impacts the selection and success of data mining methods, this survey focuses solely on time series, not only because it is the most common format of sensor data (Alemi et al., 2017), but also for the complexity in dealing with the temporal structure of data. The most prominent problems arise from the high dimensionality of time-series data and the difficulty of defining a form of similarity measure based on human perception. Regarding (R3), it is necessary to select effective pre-processing techniques to deal with the temporal dimension, such as for data representation, similarity measurement and indexing. In addition, the modelling strategy can be posed as a fault diagnosis or prognosis task. Finally, for (R4) there are also significant gaps in evaluation standards, characterized by the absence of proper performance metrics, interpretability, cost incorporation and considerations for streaming environments. Figure 3 illustrates these requirements.

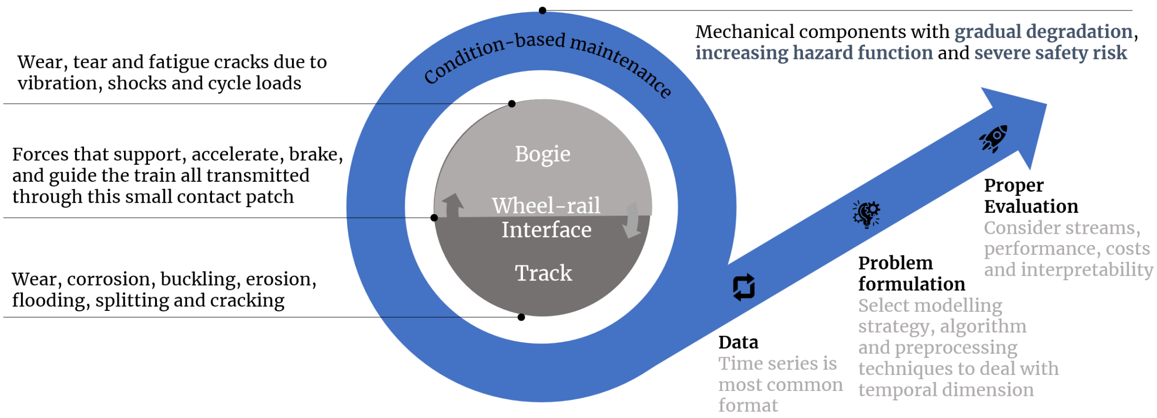


Figure 3 – An effective maintenance plan requires the maintenance strategy to be appropriate, available data, problem to be appropriately formulated, and data mining methods to be evaluated correctly

Within this context, this literature review poses the central question: What are the key challenges and opportunities in applying artificial intelligence to condition-based maintenance of railway systems, particularly in the context of the wheel-track interface, and how can they be addressed? To consider relevant papers, the choice of keywords for building the search strings was based on terms commonly found in the literature:

(“PdM” OR “CBM” or “diagnosis” or “prognosis” or “condition monitoring” or “RUL” or “PHM”) AND (“railway” OR “rail” OR “track” OR “wheel-track”) AND (“data-driven” OR “big data” OR “AI” OR “ML” OR “DL” OR “statistical” OR “time series”)

Note that this survey exclusively focused on primary research papers published in English between 2010 and 2022. Papers that utilized measurement systems unrelated to time series or did not primarily focus on wheel and track condition monitoring were excluded from consideration. Following the stated search strategy, a total of 136 publications were thoroughly reviewed. These reviewed papers were sourced from 86 different journals, including Transportation Research Part C: Emerging Technologies (n = 9), Structure and Infrastructure Engineering (n = 6), Journal of Rail and Rapid Transit (n = 6), Mechanical Systems and Signal Processing (n = 4), and Sensors (n = 4). Multiple databases were assessed during the research process, including ScienceDirect, SAGE, IEEE Xplore, Taylor&Francis, MDPI, and Springer.

To record the information from the selected studies in a structured manner, the following data extraction fields were used:

- Study information, including:

D1. Authors	D4. Database
D2. Keywords	D5. Journal/conference
D3. Year of publication	D6. Number of citations

- Data characteristics, including:

D7. Monitoring object	D11. Environmental conditions
D8. Failure mode	D12. Operational conditions
D9. Measurement Technique	D13. Textual maintenance data
D10. Measurement Sensor	D14. Data stream

- Time series data mining framework, including:

D15. Modelling strategy	D18. Preprocessing
D16. Task	D19. Structure representation
D17. Algorithm	D20. Distance representation

- Evaluation methodology, including:

D21. Performance metrics	D23. Interpretability
D22. Computational expense	D24. Cost consideration

- Main outcomes, trends and new lessons learned that advanced the state of the art:

D25. Outcomes	D26. Trends

- Limitations that could affect the validity or generalizability of the findings and future research opportunities:

D27. Limitations	D28. Future research opportunities

2.2 Related work

In the realm of related work, early literature reviews focused on discussing different types of sensors, measurement objectives, and conditions. One notable contribution in 2011 proposed a taxonomy that categorized monitoring systems into infrastructure-based infrastructure monitoring, infrastructure-based rolling stock monitoring, rolling-stock-based infrastructure monitoring, and rolling-stock-based rolling stock monitoring (Ward et al., 2011). Another comprehensive overview conducted in 2014 examined the use of wireless sensors for monitoring structures, vehicles, and machinery, shedding light on their practical applications (Hodge et al., 2014). However, the most widely adopted taxonomy formed the basis for a review conducted in 2017, which further classified monitoring systems into in-service categories, including on-board and wayside measurement, as well as in-workshop assessment (Alemi et al., 2017). In terms of wayside detection, monitoring systems were extensively reviewed, in 2005 focusing on vibration-based systems (D. Barke & Chiu, 2005), and in 2019 focusing on image processing for visual inspection (S. Liu et al., 2019). In terms of in-service vehicles, a literature review was published in 2015, focusing on track geometry condition (Weston et al., 2015). Furthermore, the literature has given considerable attention to both track geometry and wheel out-of-roundness. A comprehensive review published in 2000 provided a thorough analysis of the causes and modeling techniques of wheel out-of-roundness (Nielsen & Johansson, 2000). An extension to this review in 2003 included causes of wheel corrugation and low-order polygonal wear, along with mitigation measures to tackle these challenges (Nielsen et al., 2003). Additionally, a 2005 review examined the effects of wheel out-of-roundness on vehicle and track components, contributing to a better understanding of the implications (D. W. Barke & Chiu, 2005). Since then, numerous studies have emerged over the past decade, continuously updating the state-of-the-art knowledge in this field (Tao et al., 2020). Regarding methods for condition monitoring, early surveys primarily focused on the application of physical-based degradation prediction models (Falamarzi et al., 2019a). However, the advent of Big Data has led to a growing trend of using artificial intelligence for condition-based maintenance, with numerous publications exploring this approach (Chenariyan Nakhaee et al., 2019). Despite these advancements, there remains a scarcity of surveys in railway maintenance that specifically focus on the application of artificial intelligence. Two notable surveys conducted in 2020 and 2021 addressed maintenance applications, measurement methods, data-driven models, and evaluation considerations, providing valuable insights into these areas (Davari et al., 2021; Xie et al., 2020). However, these surveys covered a wide range of railway subdomains, sensor types, and failures, which somewhat limited their depth of analysis. In contrast, the survey presented in this paper aims to bridge the research gap by focusing on the development and evaluation of data mining methods, particularly in the context of time series sensor data and wheel-track interface components. By providing an extensive categorization, identifying open challenges, and outlining future research directions, this survey offers a more comprehensive and focused examination of the subject matter, contributing to the advancement of the field.

2.3 Data sources

To make optimal maintenance decisions, a vast amount of dynamic data from different sources is required, including service failure data, signal data, ballast history, grinding history, remedial action history, traffic data, and inspection data. The collection of this data depends on the specific components related to each type of failure, the degradation rate, and the desired accuracy of measurements. The main monitoring objects can be categorized into distinct areas of focus: wheel geometry, bogie structural components, track geometry, and track structural components. In addition, there are four main measurement methodologies: manual inspection, recording car, on-board systems, and wayside systems. Different sensors are used depending on the methodology. This survey focuses specifically on sensors presented as time series data, e.g., eddy current, vibration, displacement, impact load detectors, fiber and ultrasonic.

2.3.1 Monitoring objects

In the realm of railway system literature, the monitoring of the wheel-track interface comprises several critical components that need continuous surveillance, for which understanding the concept of failure and its associated failure modes is paramount. Failure occurs when an item is no longer able to perform its required function, resulting in a fault or state of inability, while each failure mode represents a distinct way the performance requirement fails. Given the complexity of the wheel-track interface and the diverse range of potential failure modes, this thesis focuses on aggregating trends at the component level while highlighting the most common failure modes. These monitoring objects can be categorized into distinct areas of focus: wheel geometry, bogie structural components, track geometry, and track structural components.

Firstly, wheel geometry monitoring entails the assessment of essential wheel parameters such as diameter, flange height, and tread profile. By closely monitoring these geometric attributes, deviations or abnormalities can be promptly identified, enabling timely maintenance actions to be taken. One significant concern it the presence of out-of-round wheels. These include flats formed when a wheel set locks and skids along the rail, polygonization caused by repetitive loading and unloading cycles, and corrugated treads formed by excessive friction or contact with foreign objects. In addition to wheel geometry, monitoring the health and performance of other associated bogie components is vital. Among these components, the axle bearing serves a critical role by facilitating the transfer of loads and minimizing friction during the movement of the train. Table 1 summarizes the works according to monitoring objects focusing on wheel geometry and bogie structural components.

Table 1. Categorization of works based on monitoring objects: wheel geometry and axle bearing

Object	Failure mode	Works
Wheel Geometry	Flats due to sliding and skidding, corrugated treads and polygonization with uneven wear, material fall-out, such as shelling or spalling	(Bollas et al., 2013; Braga & Andrade, 2021; S. Chen et al., 2022; X. Gao et al., 2011; Gonçalves et al., 2023; Guedes et al., 2023; Jia & Dhanasekar, 2007; Kwon et al., 2011; Mohammadi et al., 2023; Mosleh et al., 2022c; Mosleh, Montenegro, Alves Costa, et al., 2021; Mosleh, Montenegro, Costa, et al., 2021; Ni & Zhang, 2018; Shan et al., 2019; Vrignat et al., 2015; Yang et al., 2012)
Axle bearing	Material loss due to wear, inadequate lubrication and fatigue	(Fumeo et al., 2015; Jayaswal et al., 2011; Kamlu & Laxmi, 2019)

Track geometry monitoring is a critical process that focuses on evaluating various irregularities that can have negative effects on train stability and passenger comfort. These irregularities include track misalignment, unevenness, and deviations from the desired rail geometric parameters. Key parameters of concern are longitudinal level, cant for centrifugal force, twist for unevenness, alignment for lateral force, wide gauge, as well as other horizontal and vertical deformities. These failures typically initiate as small wear and cracks, but quickly propagate due to the substantial shear and normal stresses on the rail caused by rolling-sliding contact loading. As cracks progress, they can result in spalling, leading to material detachment from the rail surface.

Nonetheless, it should be noted that track geometry irregularities not only are responsible for train accidents, but they can also lead to the birth of structural defects. These defects encompass several critical aspects, including the condition of joints, turnouts/switches, rolling element bearings, missing components, fasteners, sleepers, ballast, and subgrade. Joints are used to connect two rail sections together. They provide flexibility to accommodate temperature changes and allow the rail to expand and contract without significant damage.

Turnouts, also known as switches, enable trains to change tracks or move from one line to another. Rolling element bearings are used in various parts of the railway system to reduce friction and allow smooth rotation. Fasteners, such as clips or bolts, are used to secure the rail to the sleepers that provide support and stability to the track. The ballast, a layer of crushed stones or gravel, provides support, drainage, and stability to the track, while the subgrade refers to the underlying soil or formation on which the track is constructed. The failure modes associated with these structural defects primarily involve wear, predominantly in curves, fatigue in the form of surface or subsurface cracks, and plastic flow resulting in rail corrugation. Table 2 summarizes the monitoring objects focusing on track geometry and track structural components.

Table 2. Categorization of works based on monitoring objects: track geometry, joint, turnout, rolling bearing, ballast and fastener

Object	Failure mode	Works
Track Geometry	Rail head, breaks, longitudinal level, twist and cant	(Arasteh khouy et al., 2016; Bai et al., 2015a,; Cárdenas-Gallo et al., 2017; Costa et al., 2022; Falamarzi et al., 2019b; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Falamarzi, Moridpour, Nazem, & Hesami, 2018; S. Gao et al., 2018; Guler, 2014; Jamshidi et al., 2016; Jiang et al., 2019; Khajehei et al., 2019; Khouy et al., 2014; Kwon et al., 2011; Lam et al., 2018; Lasisi & Attoh-Okine, 2018, 2019; Lederman et al., 2017; Lee et al., 2020; Q. Li et al., 2019; Mercier et al., 2012; Molodova et al., 2014, 2016; Oukhellou et al., 2008; Sadeghi & Askarinejad, 2010, 2012; Schalk et al., 2017; Sharma et al., 2018; Tsunashima, 2019; Wei et al., 2017)
Joint	Misalignment and wear	(Sun et al., 2019)
Turnout	Misalignment	(Bukhsh et al., 2019; Huang et al., 2017; Zhou et al., 2016)
Rolling bearing	Wear and overheating	(Ahmad et al., 2019; Chaolong et al., 2012; Jayaswal et al., 2011)
Fastener and Sleeper	Corrosion and exposure to moisture	(Trinh et al., 2012; Wei et al., 2017; Xia et al., 2010)
Ballast	Degradation and fouling	Audley & Andrews, 2013; Lam et al., 2018; Soleimanmeigouni et al., 2018)

The existing studies in this field have predominantly adopted a single component perspective, which oversimplifies the modeling process by isolating interconnected system effects (Lourenço, Fernandes, Canito, et al., 2022). As a result, it often leads to flawed system-level conclusions. It is important to recognize that railway systems are complex and heterogeneous, composed of interdependent components such as rails, sleepers, joints and fasteners, ballast, and subgrade. Therefore, a more comprehensive approach is necessary, considering the multi-component system dependence categorized into economic, stochastic, and structural dependence (Nicolai & Dekker, 2008).

While a few studies have addressed the economic dependence aspect by exploring the cost benefits of simultaneous maintenance of multiple components (Caetano & Teixeira, 2013; Gustavsson, 2015; Pargar et al., 2017; Verbert et al., 2017), the literature remains largely insufficient in addressing both structural dependence, which involves maintenance connections between parts, and stochastic dependence, where the failure of one component affects the performance of other parts of the system. In fact, the prevailing multi-component perspective found in the reviewed literature merely offers limited explanations of the dynamic interaction between rolling stock and track components. Some studies focus on examining the interaction between friction on sharp curves, gauge corners, and track conditions with defects (Dwyer-Joyce et al., 2013; Liang et al., 2013; Matsumoto et al., 2014; Palo et al., 2014). Others investigate the impact of overloads, weight-in-motion, and unbalanced loads on the rail (Mosleh et al., 2020; Pan et al., 2013; Pintão et al., 2022; Silva et al., 2023). Furthermore, recent works have utilized a simplified track quality index that integrates various factors, including dynamic effects, speeds, loads, and rail geometry parameters (Filograno et al., 2011; Lasisi & Attoh-Okine, 2018; H. Li et al., 2014; Sadeghi & Askarinejad, 2010), and key features of rail geometry parameters, such as gauge and twist, represented with standard deviations (Lasisi & Attoh-Okine, 2018; Ma et al., 2019; Sadeghi & Askarinejad, 2012).

2.3.2 Measurement techniques

There are three main categories for measurement techniques: manual inspection, onboard and wayside measurement methods. Manual inspection involves a walking patrol looking for signs of track defects. Onboard techniques involve installing sensors on the vehicle to monitor the condition of the track and the wheels themselves. These can be mounted on in-service vehicles that cover long lengths of track, which provides real-time data or in-workshop inspection vehicles, normally carried out periodically and rarely on busy routes. Alternatively, wayside measurement systems allow the extraction of significant amounts of data referring to all the operating vehicles with a reduced number of sensors. Table 3 highlights the works based on the four measurement techniques.

Within these various measurement techniques, various sensors are available to efficiently monitor and control different process parameters, including force, machine vision, optical geometry, acoustic, fiber bragg grating, pressure transmitter, eddy current signal, ultrasonic, magnetic and thermal. Force measurement detectors are widely used to monitor the impact between wheels and the track structure. These detectors employ vibration-based systems to accurately capture the forces involved (Fumeo et al., 2015; Gómez et al., 2018; H. Hu et al., 2017; Jayaswal et al., 2011; Liang et al., 2015; Luo et al., 2020; Ma et al., 2019; Palo et al., 2014; Tsunashima, 2019).

Table 3. Categorization of works based on measurement: manual inspection, recording car, in-service vehicles and wayside sensors

Technique	Description	Works
Manual inspection	Walking patrol looking for signs of potential failure	(Famurewa et al., 2017; C. Hu & Liu, 2016; Sadeghi & Askarinejad, 2012)
Recording car	Mechanized track patrol carried out with sparse frequency	(Andrade & Teixeira, 2013; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Falamarzi, Moridpour, Nazem, & Hesami, 2018; X. Gao et al., 2011; Kwon et al., 2011; Lee et al., 2020; Peng et al., 2011; Sharma et al., 2018)
In-service vehicles	On-board sensors providing real-time data	(Bergmeir et al., 2013; Dwyer-Joyce et al., 2013; Fumeo et al., 2015; Lederman et al., 2017; Z. Li & He, 2015; Liang et al., 2013, 2015; Matsumoto et al., 2014; Molodova et al., 2016; Rabatel et al., 2011; Sammouri et al., 2014; Schenkendorf et al., 2016; Tsunashima, 2019; Yin & Zhao, 2016)
Wayside sensors	Fixed sensors providing real-time data	(Bollas et al., 2013; Brizuela et al., 2010, 2011; Filograno et al., 2011, 2013; Lai et al., 2012; H. Li et al., 2013, 2014; Palo et al., 2014; Pan et al., 2013; Yang et al., 2012)

Vertical loads are typically measured in zones where the train is not accelerating or braking, while lateral loads are best measured in narrow curves. A commonly used sensor is the wheel impact load detector, which weighs each wheel multiple times and provides dynamic impact load information at the wheel level (Fumeo et al., 2015; Gómez et al., 2018; H. Hu et al., 2017; Jayaswal et al., 2011; Liang et al., 2015; Luo et al., 2020; Ma et al., 2019; Palo et al., 2014; Tsunashima, 2019). In addition, many systems deal directly with acceleration data (Falamarzi et al., 2019b; Karimpour et al., 2018; Molodova et al., 2016; P. Salvador et al., 2016; Sun et al., 2019; Wei et al., 2017). Machine vision technology utilizes advanced computer algorithms to process digital image data obtained from railcar underframes and side-frames (H. Li et al., 2014; S. Liu et al., 2019). By analyzing this information, machine vision systems can capture various wheel features, e.g., flange height, flange thickness, rim thickness, and diameter, as well as the condition of railcar components. Laser-based systems also play a significant role in track and wheel monitoring (Jiang et al., 2019; Lasisi & Attoh-Okine, 2018; H. Li et al., 2014; Yang et al., 2012). For example, optical geometry detectors employ cameras mounted on tangent tracks to calculate various parameters for each set of axles based on the angle of attack and tracking position. These systems enable precise measurements of the track's elevation, curvature, and alignment, which are crucial for ensuring safe and efficient railway operations. Longitudinal level, a commonly measured parameter, helps maintain track stability and smoothness (Andrade & Teixeira, 2013;

Audley & Andrews, 2013; Mercier et al., 2012; Quiroga & Schnieder, 2010). Acoustic emission monitoring involves processing the unique noise signatures emitted (Amini et al., 2016; Bollas et al., 2013; Thakkar et al., 2012). By analyzing these acoustic signals, anomalous patterns can be detected. In addition, ultrasonic sensors are extensively employed in track and wheel monitoring applications (Brizuela et al., 2010, 2011; Dwyer-Joyce et al., 2013; S. Gao et al., 2018; X. Gao et al., 2011; Jamshidi et al., 2016; Peng et al., 2011; Schalk et al., 2017). They utilize ultrasonic waves to measure parameters such as distance, thickness, and defects in materials. Fiber Bragg grating sensors are widely utilized in track and wheel monitoring due to their immunity to electromagnetic interference, multiplexing capabilities, long reach, lightweight nature, and high signal fidelity (Filograno et al., 2011, 2013; Lai et al., 2012; Ni & Zhang, 2018; Pan et al., 2013; Q.-A. Wang & Ni, 2019). These measure parameters such as strain (Falamarzi et al., 2019b; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Falamarzi, Moridpour, Nazem, & Hesami, 2018; Liang et al., 2013; Palo et al., 2014; Sadeghi & Askarinejad, 2012), temperature (Fumeo et al., 2015; Rabatel et al., 2011), and pressure (H. Li et al., 2014). Eddy current testing involves inducing electrical currents in conductive materials to evaluate their integrity and detect flaws or discontinuities (S. Gao et al., 2018; Oukhellou et al., 2008; Schalk et al., 2017; Zhou et al., 2016). Magnetic are also used to detect the changes in the signature of metallic objects (Kwon et al., 2011; Matsumoto et al., 2014).

2.3.3 Type of data

Furthermore, combining contextual data, e.g., maintenance inspection history and traffic volume, with sensors is crucial for more accurate analysis, trend identification, and decision-making. The use of both real and simulated data is prevalent in railway condition monitoring. Simulated data provides researchers with several benefits. One significant advantage is the ability to create controlled and repeatable experiments, allowing researchers to manipulate variables and observe their effects without the constraints of real-world data collection. However, it is important to recognize the limitations and the need for validation through field trials using real data. Filed trials provide valuable insights into the performance, reliability, and generalizability of models, algorithms, and interventions. They help researchers identify potential issues, assess the impact of confounding factors, and uncover unanticipated outcomes. However, there is still a lack of algorithms presented as in use with real data, translating into most studies adopting a complete offline task. However, in real world applications with wayside or on-board sensors data comes in the form of streams. Some efforts were made to address this, with algorithms built to perform computationally demand training parts off-line with model ready for fast real-time predictions. However, the literature is still missing the practical application of complete online algorithms. Table 4 encapsulates these types of data.

Table 4. Categorization of works based on characteristics of data: simulated, real, contextual and stream

Type	Description	Works
Simulated data	Controlled and repeatable experiments	(Gonçalves et al., 2023; Guedes et al., 2023; H. Hu et al., 2017; Kang et al., 2018; Lee et al., 2018; Matsumoto et al., 2014; Mohammadi et al., 2023; Mosleh et al., 2022c; Mosleh, Montenegro, Alves Costa, et al., 2021; Mosleh, Montenegro, Costa, et al., 2021; Pintão et al., 2022)
Real data	Acknowledge all uncontrollable interferences	(Falamarzi et al., 2019b; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Falamarzi, Moridpour, Nazem, & Hesami, 2018; Fumeo et al., 2015; Gerum et al., 2019; Guler, 2014; Karimpour et al., 2018; Lai et al., 2012; Lasisi & Attoh-Okine, 2018; H. Li et al., 2013, 2014; Z. Li & He, 2015; Quiroga & Schnieder, 2010; Rabatel et al., 2011; Sadeghi & Askarinejad, 2010)
Contextual data	Failures, ballast history, traffic, inspections, and interventions	(Andrade & Teixeira, 2013; S. Gao et al., 2018; Gerum et al., 2019; Lee et al., 2018; H. Li et al., 2014; Oukhellou et al., 2008; Sadeghi & Askarinejad, 2010)
Data stream	Recognize high-rate frequency of continuous flow of data	(Filograno et al., 2011; Fumeo et al., 2015; Lee et al., 2020; H. Li et al., 2014; Meixedo et al., 2022; Mosleh et al., 2022a; Ni & Zhang, 2018; Rabatel et al., 2011; Sammouri et al., 2014; Wei et al., 2017; Yin & Zhao, 2016)

2.4 Time series data mining

It is crucial to consider not only the available data but also the desired output that the model should provide, and timeframe involved. To illustrate this process, let's consider the example of rail wear caused by the friction between train wheels and the tracks. It is important to formulate specific goals based on the business requirements or the maintenance personnel's needs. For instance, are there requirements to detect rail sections with an uneven longitudinal level in real-time? Or perhaps the goal is to predict sections nearing the end of their useful lifespan and schedule replacements for the next year, treating it as a classification problem? Another approach could involve survival analysis to predict the probability of the contact rail having a level below a specific threshold within the next three years. Alternatively, the goal might be to forecast the longitudinal level in one year's time. Once the problem formulation aligns with the available data, it becomes crucial to select appropriate data mining methods as well as key implementation components representative of time-series data. For example, if the focus is primarily on condition monitoring, a state space model might be adequate for capturing the dynamics of the system. However, if there is a lot of contextual information available, a machine learning model might be better suited to handle the multidimensionality.

2.4.1 Modelling strategy

Predictive maintenance involves six modeling strategies, serving distinct purposes and relying on specific types of data. These strategies can be grouped into fault diagnosis and fault prognosis. When a machine experiences a breakdown, a substantial amount of time is typically spent on pinpointing the causes of the failure, while only a minor portion is allocated to the actual repair.

Fault diagnosis involves identifying the factors responsible for the deterioration observed in equipment with three stages involved: detection, isolation, and identification. Detection involves anomaly detection and benchmarking machine behavior to identify deviations. Isolation determines the specific kind, location, and time of fault occurrence, such as identifying the exact component of the track, bogie or defective wheel. Identification focuses on quantifying the size and behavior of the fault. Fault prognosis aims to answer questions related to the timing and extent of failure or degradation in observed processes or equipment, in three different ways. Failure prediction involves classifying failures within a specified time window, which helps optimize maintenance strategies. Survival analysis is used to assess the probability of failure over time, providing insights into equipment reliability and longevity. Finally, it is possible to estimate the remaining useful life (RUL) of machines or components. Table 5 highlights works with fault diagnosis and prognosis modelling strategies.

Table 5. Categorization of works based on modelling strategies: fault diagnosis, failure prediction, RUL estimation, survival analysis

Strategy	Problem formulation	Works
Fault Diagnosis	Is there a potential failure? Where? What? How severe?	(Alves et al., 2015; Amini et al., 2016; Brizuela et al., 2011; Du et al., 2019; Famurewa et al., 2017; Filograno et al., 2013; X. Gao et al., 2011; Gómez et al., 2018; Gonçalves et al., 2023; Guedes et al., 2023; C. Hu & Liu, 2016; Huang et al., 2017; Jayaswal et al., 2011; Jia & Dhanasekar, 2007; Kang et al., 2018; Kim et al., 2016; Krummenacher et al., 2017; Lam et al., 2018; Lourenço et al., 2023; Luo et al., 2020; Meixedo et al., 2021a, 2022; Mohammadi et al., 2023; Molodova et al., 2014; Mosleh et al., 2022b, 2022c, 2022a; Mosleh, Montenegro, Alves Costa, et al., 2021; Mosleh, Montenegro, Costa, et al., 2021; Ni & Zhang, 2018; Rabatel et al., 2011; Sadeghi & Askarinejad, 2012; Sun et al., 2019; Trinh et al., 2012; Vileiniskis et al., 2016; Xia et al., 2010; Yin & Zhao, 2016; Zhou et al., 2016)
Failure Prediction	Will a failure happen in the next year?	(Andrade & Teixeira, 2013; Bai et al., 2015a; Bergmeir et al., 2013; Gerum et al., 2019; H. Hu et al., 2017; Jiang et al., 2019; Karimpour et al., 2018; Lasisi & Attoh-Okine, 2019; Lee et al., 2020; H. Li et al., 2013, 2014; Sadeghi & Askarinejad, 2010; Sammouri et al., 2014; Sharma et al., 2018; Vrignat et al., 2015)

RUL Estimation	When will the next failure happen?	(Bai et al., 2015a; Bergmeir et al., 2013; Chaolong et al., 2012; Z. Chen et al., 2019; Costa et al., 2022; Dwyer-Joyce et al., 2013; Falamarzi et al., 2019b; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Fumeo et al., 2015; Guler, 2014; Lederman et al., 2017; Lee et al., 2018, 2020; Q. Li et al., 2019; Z. Li & He, 2015; Liang et al., 2013, 2015; Matsumoto et al., 2014; Molodova et al., 2016; Rabatel et al., 2011; Sammouri et al., 2014; Schenkendorf et al., 2016; Tsunashima, 2019; Q.-A. Wang & Ni, 2019; Yin & Zhao, 2016)
Survival Analysis	What is the probability of failure over time?	(Bollas et al., 2013; Brizuela et al., 2010, 2011; Filograno et al., 2011, 2013; Lai et al., 2012; H. Li et al., 2013, 2014; Palo et al., 2014; Pan et al., 2013; Yang et al., 2012)

2.4.2 Learning method

The modelling strategy highly influences the learning method of choice. The most common associations are: query by content (Lourenço et al., 2023), anomaly detection which involves ranking a set of instances, from the most anomalous to the most normal, solely based on intrinsic properties of the dataset (Kang et al., 2018; Meixedo et al., 2021a; Rabatel et al., 2011) and clustering by finding intrinsic groups of types or severities of damage (Du et al., 2019; Mosleh et al., 2022b; Schalk et al., 2017). The most used techniques are artificial neural networks (ANN) (Gómez et al., 2018; H. Hu et al., 2017; Sadeghi & Askarinejad, 2012; Sammouri et al., 2014; Yin & Zhao, 2016), Bayesian method (Andrade & Teixeira, 2013; Kang et al., 2018; Lam et al., 2018; Oukhellou et al., 2008; Sammouri et al., 2014) and k-means (Gerum et al., 2019; Q. Li et al., 2019). Conversely, despite classification being used in fault diagnosis tasks, it is mostly performed for failure prediction based on known training examples. Regarding RUL estimation, forecasting and regression models are the most used for explicitly modelling variable dependencies to predict a subset of the variables from others, while for survival analysis both probability distribution models (Audley & Andrews, 2013; Oukhellou et al., 2008) and stochastic processes (Chaolong et al., 2012; Quiroga & Schnieder, 2010; Soleimanmeigouni et al., 2018) are used. Specifically, the literature has shown a tendency for the usage of kNN (Z. Li & He, 2015; Nadarajah et al., 2018; Sammouri et al., 2014), ANN (Lee et al., 2018), decision trees (H. Li et al., 2014; Z. Li & He, 2015; Nadarajah et al., 2018), logistic regression (Cárdenas-Gallo et al., 2017; Kang et al., 2018; Sadeghi & Askarinejad, 2010), random forest (RF) (Falamarzi et al., 2019b; Gerum et al., 2019; Jamshidi et al., 2016; Z. Li & He, 2015), adaptive boosting (Ada-boost) (Trinh et al., 2012; Xia et al., 2010) and support vector machines (SVM) (Falamarzi et al., 2019b; S. Gao et al., 2018; C. Hu & Liu, 2016; Jayaswal et al., 2011; Jiang et al., 2019; Lasisi & Attoh-Okine, 2018; Lee et al., 2018; H. Li et al., 2014; Nadarajah et al., 2018; Sammouri et al., 2014; Tsunashima, 2019; Zhou et al., 2016). Table 6 encapsulates the link between common learning methods and tasks.

Table 6. Categorization of works based on learning methods: query by content, anomaly detection, clustering, classification, forecasting, regression, probability distribution model and stochastic process

Learning Task	**Common Learning Methods**	**Works**
Fault Diagnosis	Query by content, anomaly detection, clustering	(Lourenço et al., 2023, Kang et al., 2018; Meixedo et al., 2021a; Rabatel et al., 2011, Du et al., 2019; Mosleh et al., 2022b; Schalk et al., 2017)
Failure Prediction	Classification	(Bai et al., 2015a; Bergmeir et al., 2013; Gerum et al., 2019; H. Hu et al., 2017; Jiang et al., 2019; Karimpour et al., 2018; H. Li et al., 2013, 2014; Sadeghi & Askarinejad, 2010; Sammouri et al., 2014)
RUL Estimation	Forecasting, regression	(Bai et al., 2015a; Chaolong et al., 2012; Z. Chen et al., 2019; Costa et al., 2022; Falamarzi et al., 2019b; Falamarzi, Moridpour, Nazem, & Cheraghi, 2018; Fumeo et al., 2015; Guler, 2014; Lee et al., 2018, 2020; Q. Li et al., 2019; Z. Li & He, 2015; Q.-A. Wang & Ni, 2019)

Survival Analysis	Probability distribution model, stochastic process	(Chaolong et al., 2012; Quiroga & Schnieder, 2010; Soleimanmeigouni et al., 2018, Audley & Andrews, 2013; Oukhellou et al., 2008)

2.4.3 Implementation components

When analyzing time series sensor data, it is also important to consider several factors that can impact the success of data mining methods. In IoT environments with numerous low-cost sensors, noise, inconsistencies, missing, or duplicated data are common challenges. Fluctuations in reported values may result from minor variations in sensor sensitivity, unrelated events near the sensor, or transmission errors. In railway condition monitoring, various signal preprocessing techniques are used to address these issues. Noise filtering is handled by using traditional signal processing techniques like digital filters or wavelet thresholding. Scaling differences between time series are handled with normalization, transformation invariances and techniques that intrinsically handle these differences. Data fusion is performed to deal with the inclusion of more contextual information for enhanced detection capabilities. This contextual information in IoT data comes from sensors distributed throughout the monitored environment, i.e. the temporal correlation between observations over time, spatial relationships between multiple sensors deployed in the same system, which becomes more complex when sensors are mobile and the external context of weather, soil type, and others. typical redundancy in terms of sensor types and locations while satisfying certain invariance properties (Z. Li & He, 2015). The most used technique is Principal Component Analysis (Famurewa et al., 2017; Guedes et al., 2023; Jiang et al., 2019; Lasisi & Attoh-Okine, 2018; Q. Li et al., 2019; Mohammadi et al., 2023). Also, a recent work proposed Deep Autoencoder to capture and compress the salient features of the input data (Q. Li et al., 2019). Resampling is also typically to smooth the noise in high-frequency signals. Finally, to represent the fundamental shape characteristics of a time-series, three groups of techniques are available (Esling & Agon, 2012). For short time series, shape-based representation is the most used, while for long time series both feature-based and model-based representations are derived and compared.

Feature-based is the most used approach wheel-track condition monitoring. Various advanced signal processing methods have been proposed that extract condensed damage-sensitive knowledge from time-series sensor data. Several works propose both statistical and frequency-domain features, e.g., the Fourier transform (Dwyer-Joyce et al., 2013; Lederman et al., 2017; P. Salvador et al., 2016). However, time-frequency techniques have been the most widely used for dynamic system analysis, for its ability to identify content of a signal in the frequency domain without losing information abouts its time domain characteristics (H. Li et al., 2013; Liang et al., 2013, 2015; P. Salvador et al., 2016). Three commonly used techniques are: the wavelet transform (Gómez et al., 2018; Jayaswal et al., 2011; Krummenacher et al., 2017; Lasisi & Attoh-Okine, 2018; Liang et al., 2013, 2015; Molodova et al., 2014, 2016; Schenkendorf et al., 2016; Wei et al., 2017), Wigner-Ville Transform (Liang et al., 2013, 2015), Short-time Fourier Transform (Liang et al., 2013, 2015; Mohammadi et al., 2023). Furthermore, other approaches were developed to help dealing with the non-linear effects on the suggested time-dependent characteristics. For instance, changes in train speed tend to affect the impact load, but this can vary based on the specific length of wheel flats, not following the pattern of increased impact load with higher train speeds (Dukkipati & Dong, 1999). Some alternatives were then suggested to extract the features, such as a time-frequency kurtosis to reduce surrounding noise and highlight faulty signal patterns of wheel flats (Amini et al., 2016), spectral kurtosis (Mosleh, Montenegro, Costa, et al., 2021), symbolic data (Alves et al., 2015) and empirical mode decomposition to transform the signal into several intrinsic mode functions that isolate the failure signal mode from interferences and a criterion based on signal lag and envelope spectrum (Gonçalves et al., 2023; Mosleh, Montenegro, Alves Costa, et al., 2021).

Model-based representation consists in directly describing the dynamic behavior of the signal, such as autoregressive models (Braga & Andrade, 2021; Costa et al., 2022; Meixedo et al., 2021a; Quiroga & Schnieder, 2010), state space models, e.g., the Hidden Markov Model (Bai et al., 2015a; Z. Chen et al., 2019; Kamlu & Laxmi, 2019; Lam et al., 2018; Mercier et al., 2012; Sharma et al., 2018; Vrignat et al., 2015) and regression (Falamarzi et al., 2019b; Nadarajah et al., 2018). In fact, model-based representations are also frequently used to extract coefficients which are then used in a feature-based approach (Audley & Andrews, 2013; H. Hu et al., 2017; H. Li et al., 2014; Z. Li & He, 2015). Recently, there has also been an increased use of deep learning models. Recurrent Neural Networks (RNN) have become popular architectures used for railway condition monitoring, including Long Short-Term Memory (LSTM) for capturing long-term dependencies in sequential data (Falamarzi et al., 2019b; Guler, 2014; Falamarzi, Moridpour, Nazem, & Hesami, 2018; Gerum et al., 2019). Alternatively, CNNs (Convolutional Neural Networks) despite commonly associated with image recognition tasks, also have been applied to time series data with some modifications (Luo et al., 2020; Sun et al., 2019).

Shape-based representation directly focuses on distinguish or match any pair of time-series with an intuitive distance. To formalize this measure, it is necessary to establish a notion of similarity based on perceptual criteria rather than strict mathematical identity. In railway condition monitoring Dynamic Time Warping (DTW) (Du et al., 2019; Huang et al., 2017; Kim et al., 2016; Lourenço et al., 2023) has been preferred over the Euclidean distance (Vileiniskis et al., 2016), as it allows for the recognition of perceptually similar objects, even if they are not identical in mathematical terms. Table 7 categorizes these algorithms into the three described groups.

Table 7. Categorization of works based on implementation components for time series: feature-based, model-based and shape-based

Representation	Algorithm	Works
Feature-based	Statistical, Time-Frequency, Spectral, Model Coefficients	(Gómez et al., 2018; Lasisi & Attoh-Okine, 2018; Liang et al., 2013, 2015; Molodova et al., 2014, 2016; Schenkendorf et al., 2016; Wei et al., 2017, Dwyer-Joyce et al., 2013; Lederman et al., 2017; P. Salvador et al., 2016, Gonçalves et al., 2023; Mosleh, Montenegro, Alves Costa, et al., 2021)
Model-based	Autoregressive, HMM, Regression, RNN, LSTM	(Bai et al., 2015a; Z. Chen et al., 2019; Kamlu & Laxmi, 2019; Lam et al., 2018; Mercier et al., 2012; Sharma et al., 2018; Vrignat et al., 2015, Falamarzi et al., 2019a; Guler, 2014), Braga & Andrade, 2021; Costa et al., 2022; Quiroga & Schnieder, 2010)
Shape-based	Euclidean, DTW	(Du et al., 2019; Huang et al., 2017; Kim et al., 2016; Lourenço et al., 2023; Vileiniskis et al., 2016)

Finally, the stationarity of time series data must be considered. A stationary time series has consistent mean, variance, and autocorrelation over time. However, IoT time series data often exhibit non-stationarity in the form of concept drifts, seasonality and change points, which is particularly difficult to handle as the operation of this system is dynamic over-functioning time. Concept drift refers to changes in the statistical distribution of data over time, while seasonality involves cyclical changes occurring over longer time scales. Change points represent abrupt and permanent shifts in the normal state of a monitored system. The standard solution involves the identification of change points, then determining functions for curve fitting the intervals between them (S. Chen et al., 2022; Fumeo et al., 2015; Lourenço, Fernandes, Marreiros, et al., 2022).

2.5 Evaluation metrics

Evaluating if an intelligent system can predict when a train requires maintenance goes beyond mere accuracy measurements. In this chapter, multifaceted aspects of evaluation are assessed, including performance targets, the cost implications of false-positives and false-negatives, model interpretability, and the continuous assessment of the system's performance through data stream evaluation. Rail defects datasets often exhibit a highly skewed distribution, with a significant majority of observations belonging to the non-defective class. However, standard metrics used in balanced classification may not be suitable for imbalanced data. Thus, choosing the appropriate evaluation metric is crucial, for which three main types should be considered: threshold metrics, ranking metrics, and probability metrics (Bukhsh et al., 2019; H. Hu et al., 2017; Kang et al., 2018; H. Li et al., 2014; Z. Li & He, 2015; Lourenço et al., 2023; Rabatel et al., 2011; Sharma et al., 2018). Threshold metrics, such as accuracy, sensitivity-specificity, and precision-recall, quantify classification prediction errors. Ranking metrics evaluate classifiers based on their ability to separate classes and include the ROC curve and precision-recall curve. Probabilistic metrics quantify uncertainty in a classifier's predictions, which is useful for penalizing wrong but highly confident predictions. Intrinsically related to this imbalanced distribution is the notion of a false alarm or false safe prediction. From the engineer's perspective, high false alarm prediction usually leads to ineffective and unnecessary decision-making, while false safe prediction can cause huge loss for the railway service suspensions. Despite the need for these to be carefully defined according to the business's needs and the characteristics of the data, only a few studies addressed economic impact (Al-Douri et al., 2016; Arasteh khouy et al., 2016; Khajehei et al., 2019; Khouy et al., 2014). To calculate this cost, both mandatory regulations and operational imperatives of equipment uptime and people safety should be considered, constrained the availability of the maintenance team and track line. Furthermore, most studies exhibit a fundamental barrier between evaluation and optimization of sensor types, placement and repair plans where human operators require support from AI to strike multi-objective trade-offs, where different priorities need to be balanced. However, the exact dynamics of these trade-offs may not be fully known, and decisions must be made on a case-by-case basis. To overcome this incompleteness in the problem formalization, explanations play a crucial role in ensuring trustworthiness (Audley & Andrews, 2013; Famurewa et al., 2017; Gerum et al., 2019; H. Li et al., 2014; Mercier et al., 2012; Palo et al., 2014; Rabatel et al., 2011; Sharma et al., 2018). This includes developing visual analytics techniques, prototypical examples, and deductive argumentative systems for explanations. Finally, the large volumes of sensor data impose on algorithms memory and time constraints. Thus, algorithms must be evaluated on scalability and accuracy over time in data stream settings (Audley & Andrews, 2013; Famurewa et al., 2017; Gerum et al., 2019; H. Li et al., 2014; Mercier et al., 2012; Palo et al., 2014; Rabatel et al., 2011; Sharma et al., 2018). Table 8 summarizes these evaluation metrics.

Table 8. Categorization of works based on evaluation metrics: performance, interpretability, cost and computational expense

Dimension	Types	Works
Performance	Threshold, ranking, probabilistic	(Bukhsh et al., 2019; H. Hu et al., 2017; Kang et al., 2018; H. Li et al., 2014; Z. Li & He, 2015; Lourenço et al., 2023; Rabatel et al., 2011; Sharma et al., 2018)
Interpretability	Visual analytics, prototypical examples, deductive argumentative systems	(Audley & Andrews, 2013; Famurewa et al., 2017; Gerum et al., 2019; H. Li et al., 2014; Mercier et al., 2012; Palo et al., 2014; Rabatel et al., 2011; Sharma et al., 2018)
Cost	Mandatory regulations, operational imperatives	(Al-Douri et al., 2016; Arasteh khouy et al., 2016; Khajehei et al., 2019; Khouy et al., 2014)
Computational expense	Memory requirements, computational time, stream	(H. Li et al., 2014)

2.6 Future challenges and research directions

Based on four key requirements for creating an effective maintenance plan, this survey analyzed the significant progress in low-cost and easily maintainable condition monitoring systems. Despite the vast body of work devoted to this topic in the early years, the research has not been driven so much by actual problems but by an interest in proposing new approaches. Regarding the appropriateness of a condition-based maintenance strategy, it was established that condition-based maintenance is the appropriate strategy for addressing the gradual degradation processes observed in the wheel-rail interface, such as wear and rolling contact fatigue, being the most significant monitoring object in railway. Concerning the availability of data, the survey focused on time series due to its prevalence in sensor data and the complexity involved in handling its temporal structure. Within this context, several recommendations for future research were identified:

- Formulating the problem in detail to properly select effective pre-processing techniques for dealing with the temporal dimension, as well as determining the modeling strategy. Future works should provide a more detailed business-oriented justification of the technique's selection;
- Adopting more comprehensive approaches, considering the multi-component system dependence categorized into economic, stochastic, and structural dependence. The existing studies in this field have predominantly adopted a single component perspective, which oversimplifies the modeling process;
- Developing automatic data labeling methods for large volumes of sensor data with highly imbalanced distribution of faulty samples. This would enable the advent of more valuable fault prognosis approaches for real-world applications;

- Integrating contextual data on failures, ballast history, traffic, inspections, and interventions data. Future efforts should lie in text mining techniques to leverage narrative descriptions commonly used in the railway industry;
- Developing algorithms with scalability, that accommodate increased data volume and adapt to changes in the maintenance requirements;
- Designing algorithms and evaluation metrics for data stream environments in railway maintenance, to enable real-time monitoring and proactive maintenance;
- Incorporating automatic hyperparameter tuning, for models better adapted to the specific characteristics of the data;
- Translating research findings into practical applications with proper performance metrics, cost considerations and interpretability. This includes developing visual analytics techniques, prototypical examples, and deductive argumentative systems for explanations. Involving key stakeholders, such as maintenance personnel, engineers, and management, in the development and implementation of these systems is key.

3 Data

For a data-driven approach, the available data is of key importance. Building a physical-based model provides the opportunity to parameterize and validate the full diagnosis task and interpret artificial intelligence models, since it would be too cost-prohibitive to induce them in a real railway system. In the last two decades, researchers explored the intricate mechanisms of out-of-roundness formation on wheels (Cai et al., 2019a; Johansson & Andersson, 2005), making it possible to induce reliable faults into a simulation (Lourenço, Fernandes, Marreiros, et al., 2022). Numerical modelling allows for a deeper understanding on the prediction of anomalies that may not be feasible to reproduce in experimental measurements affected by external factors, such as environmental disturbances, measurement errors, and electromagnetic interferences.

The data used throughout the thesis is introduced in the current chapter and its distinctive characteristics are described. This thesis is part of a project that is carried out together with the Laboratory of Seismic and Structural Engineering in the Research Unit CONSTRUCT at the Department of Civil Engineering in University of Porto and access to the data generator software has been generously provided by them. Two datasets were constructed representing vibration-based systems with onboard and wayside measurement methods. Onboard techniques involve installing sensors on the vehicle to monitor the condition of the track and the wheels themselves. Alternatively, wayside measurement systems are a more feasible solution to identify wheel defects, as the installation of sensors on the track allows the extraction of significant amounts of data referring to all the operating vehicles. This chapter is structured as follows: description of the process of simulating train-track dynamic interaction in Section 3.1, and numerical modelling to represent two virtual monitoring systems: wayside in Section 3.2 and on-board in Section 3.3.

3.1 Train-track dynamic interaction

Numerical simulations of train-track dynamic interaction were conducted using VSI software, which was previously validated and described in detail in the authors' previous publication (Montenegro et al., 2015). The VSI model coupled the train to the track through a 3D wheel-rail contact model that uses the Hertzian theory to compute normal contact forces and the USETAB routine to assess the tangential forces resulting from rolling friction creep (Hertz, 1882; Kalker, 1996). The VSI software was implemented in MATLAB® and imports structural matrices from both the vehicle and structure, which were previously modelled separately in a finite element (FE) package. A graphical representation of the numerical modelling is presented in Figure 4.

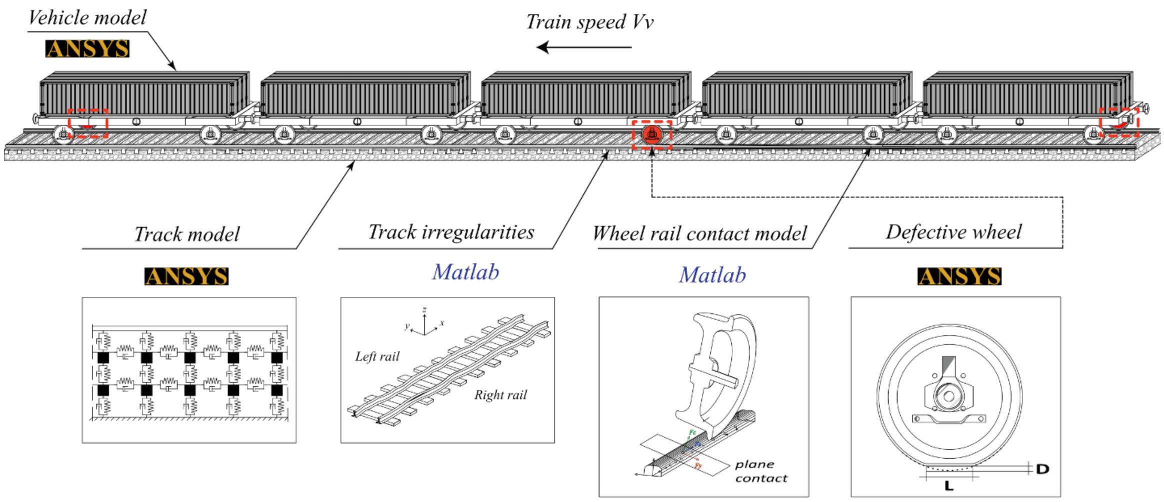


Figure 4 - Train-track dynamic interaction model with ANSYS® and Matlab: the wheels marked in red were defective in the damaged scenarios, namely the first and last on the right side, and the sixth on the left side.

In this study, the track was modelled using beam elements to represent the rails and sleepers, spring-dashpot elements to simulate the behavior of the flexible layers such as the ballast and fasteners/pad, and mass point elements to consider the mass of the ballast. The train was modelled in ANSYS® through a multibody formulation, using spring-dashpot elements to simulate the flexibility of the primary and secondary suspensions, rigid beams to consider the rigid body movements of the vehicle, and mass point elements located at the center of gravity of each body, including carbody, bogies, and wheelsets, to simulate their mass and inertial effects.

3.2 Numerical wayside modelling

The wheel out-of-roundness detection system includes an accelerometer and strain gauge installed along the stretch of the track. To test and validate the methodology presented in this study, a wide range of 3D simulations based on the train-track interaction model were performed. The virtual wayside monitoring system includes measurements evaluated for 24 different locations mounted on the rail between two sleepers. Figure 5 exemplifies 4 sensors installed along the stretch of the track in midspan location.

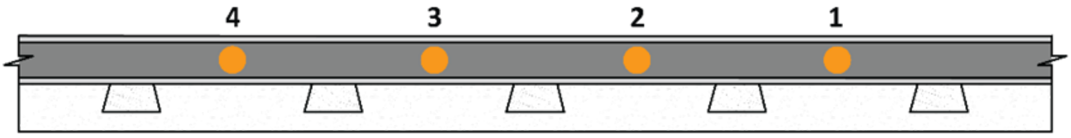


Figure 5 - Virtual wayside monitoring system: sensor placement

A virtual simulation of baseline and damaged wheel scenarios was accomplished to test and validate the automatic out-of-roundness diagnosis presented in this study. After successful validation, this method has the potential to be applied to real experimental data considering different types of trains with several wheel defect conditions. In real conditions, the wheels are not entirely smooth and have imperfections, i.e., flats and polygonization. Despite their small size, these out-of-roundness irregularities can cause extreme variations in the wheel–rail contact forces, producing vibrations on the train and track components. As presented in Figure 4, three wheels were considered as defective: the first and last on the right side, and the sixth on the left side.

For wheel flats, two intervals of flat lengths (L_w) were considered, nominated as L1 and L2. The uniform distributions U (25, 50) mm, and U (50,100) mm define the lower and upper limits of the wheel flat length for each interval L1 and L2, respectively. The wheel flat depth (D_w) is calculated based on the following equation (Zhai et al., 2001):

$$\mathrm{D}_w = \frac{L_w^2}{16R_w} \tag{1}$$

in which R_w is the radius of the wheel. The vertical profile deviation of the wheel flat is defined as:

$$Z = -\frac{\mathrm{D}_w}{2}\left(1 - cos\frac{2\pi\, x_w}{\mathrm{L}_w}\right) \cdot H\left(x_w - (2\pi R_w - \mathrm{L}_w)\right), 0\ \leq x\ \leq 2\pi R \tag{2}$$

where H represents the Heaviside function and x_w is the coordinate aligned with the track longitudinal direction. Figure 6 illustrates the effect of different wheel flat severities.

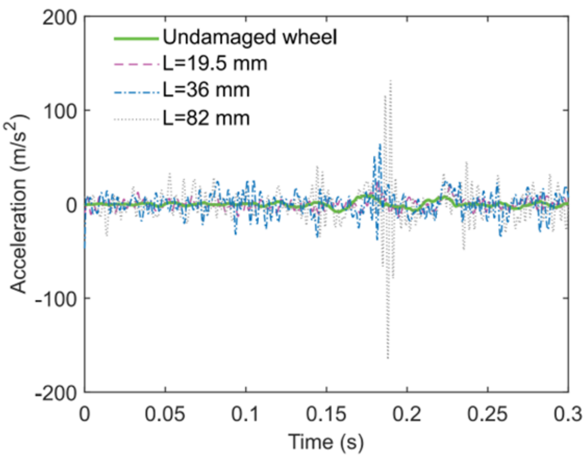


Figure 6 – Acceleration time series for different wheel flat lengths (L_w): 0, 19.5, 36 and 82 mm

For wheel polygonization, a periodic radial tread irregularity around the wheel circumference was considered by varying wavelengths (λ) in the function of the harmonic order (θ) and wheel radius:

$$\lambda = \frac{2\pi R_w}{\theta}, \quad \theta = 1, 2, 3 \dots, n \tag{3}$$

The selected wheel profiles were characterized by the wavelengths in the first 20 harmonics, with the 6th to 8th harmonic orders being dominant and different irregularity wheel profiles generated based on the sum of sine functions (H=20) as follows:

$$w(x_w) = \sum_{\theta=1}^{H} A_\theta sin\left(\frac{2\pi}{\lambda} x_w + \varphi_\theta\right) \tag{4}$$

where A_θ is the amplitude of the sine function for each wavelength, which is calculated by the function:

$$A_\theta = \sqrt{2} \cdot 10^{\frac{L_w}{20}} \cdot w_{ref} \tag{5}$$

with $w_{ref} = 1\ \mu m$. The wheel irregularity level (L_w) values were selected based on the irregularity spectrum in Figure 7, produced with measurement values of four wheels with polygonal damage (Cai et al., 2019a). Considering phase angles to the sine functions that are uniformly and randomly distributed between 0 and 2π, several wheel irregularity profiles were generated with Equation 4 to obtain different damage severities between 0.8 mm and 1.2 mm.

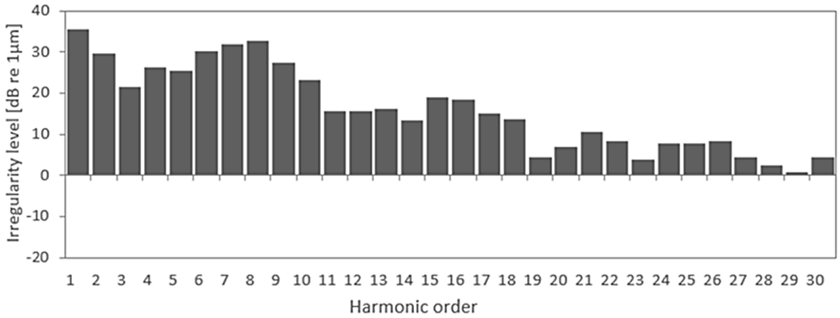


Figure 7 – Wheel irregularity amplitude spectra (L_w) and harmonic order (θ) distribution

In real conditions, the rails also present small imperfections that can significantly affect the wheel-rail contact force values. In this context, 10 unevenness profiles, shown in Figure 8, were generated for wavelengths between 1 m to 30 m, which covers the D1 wavelength interval defined by EN 13848-2 standards, with a sampling discretisation of 0.001 m.

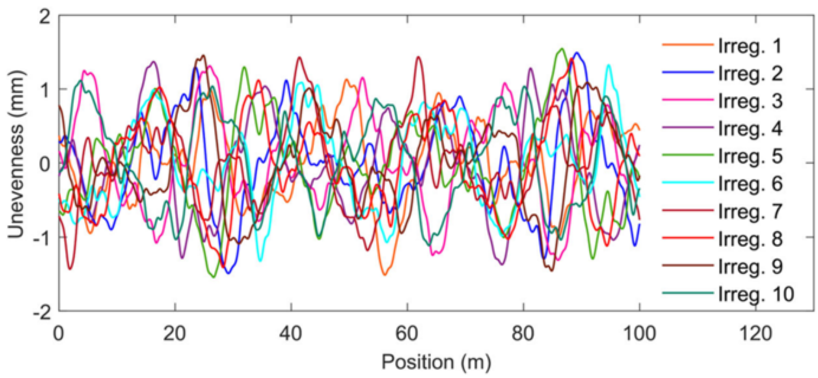


Figure 8 – Simulation of 10 unevenness rail profiles for wavelengths between 1 m to 30 m

The amplitude of the unevenness profile was varied between −2 mm and 2 mm, for a total simulation length of 100 m. These were based on the real track irregularities of the Northern Line of the Portuguese Railway Network, measured with a track inspection vehicle EM120 every six months. It is important to note that the wavelengths used to generate these track irregularities are significantly longer than the wheel flat and polygonization. Thus, the exited frequencies due to the track are much lower than those of a defective wheel. More details about the generation of unevenness profiles can be found elsewhere (Mosleh et al., 2020). By reproducing the responses of the rail considering several speeds, wheel profiles and rail irregularities, different time-series samples were simulated to capture a range of operating conditions (Mosleh et al., 2022c). The dynamic analyses were carried out for one passing vehicle. Table 9 summarizes the 23520-wheel scenario, divided into groups of one axle (J=7440) and two axles (K=16080). Within these groups, further subdivisions were made according to baseline and damaged conditions. The first encompasses the passage of baseline vehicles between 60 and 100 km/h in 10 unevenness profiles, measured by 24 different sets of sensors (accelerometer and strain gauge). The second includes the measurement of vehicles with defective wheels circulating at 80 km/h, for a single unevenness profile and 24 different sets of sensors. To obtain a more reliable reproduction of the rail response, an artificial noise was generated based on 5% of the maximum response of the signal and added to all the original signals. Additionally, all the time series are filtered based on a low-pass Chebyshev type II digital filter considering a cut-off frequency of 500 Hz.

Table 9. Summary of baseline, polygonization and flat wheels in wayside scenarios, considering vehicle used, number of track irregularity profiles, range of speeds, noise ratio, location and amplitude of defects

	Baseline scenarios	Damaged scenarios	
	Perfect Wheel	**Polygonization Wheel**	**Flat Wheel**
Vehicle	5 freight wagons of the Laagrss type		
Irregularity profiles	10	1	1
Speeds	60 – 100 km/h	80 km/h	80 km/h
Noise ratio	5 %	5 %	5%
Location of defect	-	1st wheel right	1st wheel right 6th wheel left 10th wheel right
Amplitude of defect	-	0.8 – 1.2 mm	25 – 50 mm 50 – 100 mm
Total passages of 1 axle	6240	720	480
Total passages of 2 axles	15360	0	720

3.3 Numerical on-board modelling

The track condition detection system includes 7 accelerometers installed in the carbody and wheelset, as depicted in Figure 9. The virtual simulation of baseline train scenarios was identical to the one previously described in the wayside model in terms of the healthy wheels, 10 unevenness rail profiles and the noise distribution based on 5% of the maximum response of the signal.

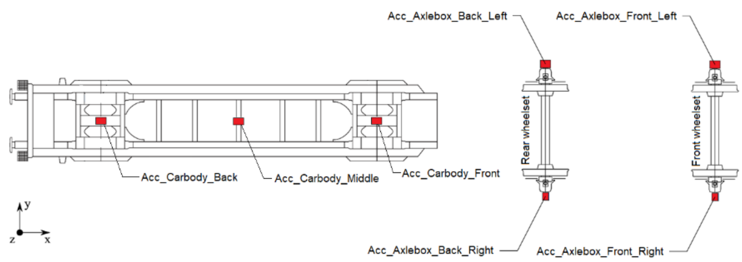


Figure 9 – Virtual on-board monitoring system: sensor placement

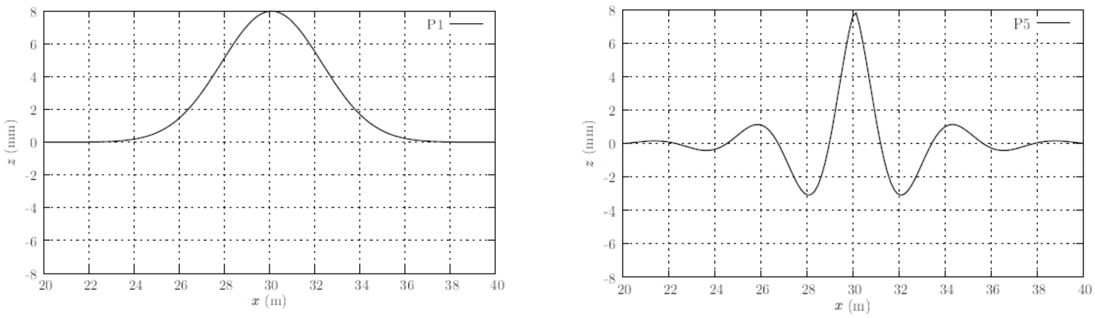
In this case, however, the passage of trains was simulated for speeds between 40 to 120 km/h. Another key difference is that 2 types of loading schemes were considered for a full load of 40.3 ton and empty load of 10.8 ton. For track defects, 2 types were considered, P1 and P5. Regarding P1, the defect was modelled as follows:

$$z = Ae^{(-1/2)(kx)^2} \tag{6}$$

with k ranging from 0.47 to 0.3 for wavelengths of 56 to 25 m, respectively. Regarding P5, the equation was:

$$z = Ae^{-k|x|}cos(\pi kx) \tag{7}$$

With k ranging from 0.65 to 0.3 for wavelengths of 5 and 15 m, respectively. To simulate different defect severities, the amplitude was set to 4 levels: 6 mm as acceptable, 8 mm as a first warning, 10 mm as recommended intervention and 17 mm as a situation of safety risk. Figure 10 presents the individual effect of the defect models.



(a) P1, A = 8 mm

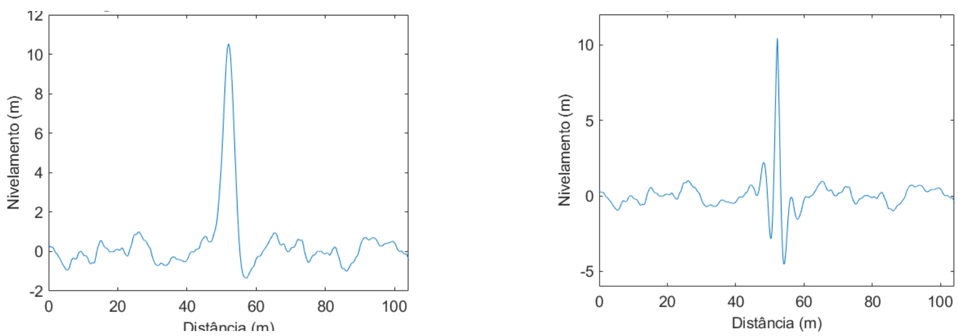
(b) P5, A = 8 mm

Figure 10 – Track defects for different type of irregularities

After combining all these factors, a total of 200 and 64 scenarios were simulated for baseline and damaged scenarios. Figure 11 illustrates two defective signals with irregularity 1 and a defect amplitude of 10 mm. Table 10 summarizes all the scenarios obtained.

Table 10. Summary of baseline and damaged (P1 or P5) on-board scenarios, considering loading schemes, track irregularity profiles, range of speeds, noise ratio, wavelength and amplitude of defects

	Baseline scenarios	**Damaged scenarios**	
	Healthy Track	**P1**	**P5**
Loading schemes	2	1	1
Irregularity profiles	10	1	1
Speeds	40 – 120 km/h	75 km/h	75 km/h
Noise ratio	5 %	5 %	5%
Wavelength		5 – 25 m	15 – 25 m
Amplitude of defect	-	8 lengths: ± 6mm (acceptable) ± 8mm (warning) ± 10mm (intervention) ± 17mm (safety risk)	
Total passages	200	40	24



(a) Irregularity 1, Defect P1, A= 10, k =.65

(b) Irregularity 1, Defect P5, A= 10, k =.47

Figure 11 – Defective track acceleration measurement varying irregularity, wavelength, type and amplitude of defect

4 Fault diagnosis methodologies

This chapter presents the devised time series data mining methods that are used to diagnose wheel out-of-roundness and railway conditions while relying on vibration-based systems. Nonetheless, before devising machine learning models, the itemized problems that need to be tackled are the following:

- The data being studied is static, which means it is not possible to track the wheel and vehicle's lifetime, and thus not feasible to model the underlying degradation process;
- The majority of observations did not experience a failure, which introduces bias in the data distribution and requires proper selection and development of model validation techniques;
- Visualizing time series data larger than several ten thousand observations can be challenging and working with super-high dimensional raw data can be costly in terms of processing and storage costs, requiring a high-level representation or abstraction of the data;
- Time series data mining techniques need to be robust against transformations, such as shifts or scaling, which can occur in the time axis or amplitude of time series data and other noisy components, such as measurement errors and outliers;
- The emergence of a massive amount of data that is too big to store, such as streaming data, requires immediate data reduction to achieve a reasonable storage size of the data.

Amidst the challenges of the context, multiple approaches were developed. Sections 4.1-4.4 focuses on Approach A, which is presented in detail, while the remaining approaches are briefly introduced to maintain concision. Table 11 presents a summary of the main characteristics and research gaps addressed by the various methodologies. Approach A enabled for comprehensive fault diagnosis with the three necessary objectives: (1) detection of aberrant train behaviour; (2) isolation of specific defective wheels; (3) identification of the severity for the respective defect type. Section 4.5 presents Approach B which adapts A to a context of streaming data with a fixed sliding window, while, in Section 4.6, approach C proposes an alternative query by content algorithm based on dynamic time warping and triggered by a concept drift detector to directly capture the shape of the signal. In Section 4.7, approach D includes parallelize implementation of machine learning algorithms that scale well to very large databases, and automatic hyperparameter tuning process, saving time, and reducing the potential for errors, while making the algorithm accessible to non-expert users in the machine learning field. In Section 4.8, approach E studies algorithms intrinsically prepared for data streaming environments, which automatically update themselves and address the issue of concept drift.

Table 11. Summary of the main characteristics and research gaps addressed by the various methodologies (Technique: Wayside, On-board; Stream; Task: Detection, Diagnosis; Method: Anomaly Detection – AD, Query by Content – QC; Algorithm: Isolation Forest – iForest, Locality sensitive hashing technique – LSHAD, XStream; Preprocess: Short-time Fourier transform – STFT; Structure: Hidden Markov model – HMM; Distance: Dynamic Time Warping – DTW; Auto-tune; Time)

	Data		**Time Series Data Mining**						**Evaluation**	
	Technique	**Stream**	**Task**	**Method**	**Algorithm**	**Preprocess**	**Structure**	**Distance**	**Auto-tune**	**Time**
A	Wayside		Diagnosis	AD	iForest	STFT	HMM			
B	On-board	Yes	Detection	AD	iForest	STFT	Window			
C	Wayside	Yes	Detection	QC				DTW		
D	Wayside		Detection	AD	LSHAD	STFT	HMM		Yes	Yes
E	On-board	Yes	Detection	AD	XStream					Yes

Focusing on the suggested approach A, a four-stage process was developed for unsupervised fault diagnosis algorithm of wheel out-of-roundness, with two failure modes: wheel flats and wheel tread polygonization. Figure 12 presents the procedure. The method was intended to solve the challenging nature of localizing out-of-round wheels in a train passage with several wagons and subsequently identify the severity of the defect individually. The initial stage of the approach involves segmentation using Hidden Markov Models (HMM) to capture the temporal structure of the strain gauge signal, which is frequently less impacted by the passage of damaged wheels. This step is vital in enabling the extraction of several localized features from the data, conducted in a second stage. For both statistical and frequency domain characteristics in accelerometer data, the Short-time Fourier Transform (STFT) was selected over the wavelet transform (Cohen, 1995), since it has been often identified as a preferred alternative to address the non-linear environmental and operational effects. The third stage of the approach involves determining a minimal subset of features for fault diagnosis using Principal Feature Analysis (PFA). Finally, a single damage indicator is created with Isolation Forest (iForest) for each group of axles to complete the fault diagnosis task, considering the differences in vibration patterns between the axle groups. Results were compared with other state-of-the-art anomaly detection algorithms, i.e., mahalanobis distance, support vector machines and local outlier factor. It is worth noting that the proposed approach is designed to be practical and efficient in real-world applications, as only the selected features are extracted in the deployment of the monitoring system. The main contributions of this work are threefold:

- development of a methodology capable of wheel fault diagnosis without interfering with normal operating conditions, focusing on the challenging task of fault isolation and localization;
- segmentation of the time series measurements, for implicit indexing and guided extraction of localized features, easily correlated with the damage of an individual wheel;
- test the effectiveness of the proposed method with respect to different types and sizes of out-of-roundness defects with high accuracy, using enough data to confirm generalization.

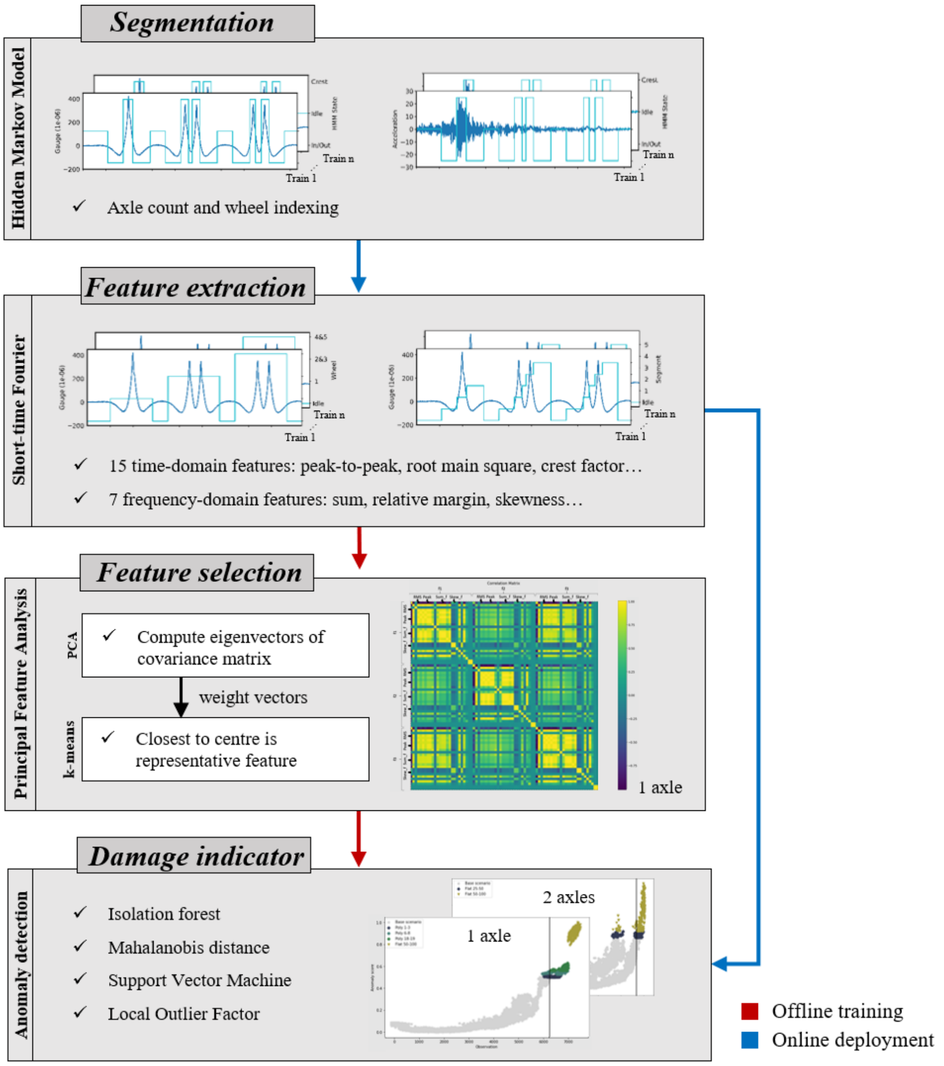


Figure 12 – Framework A of out-roundness wheel fault diagnosis methodology with four main stages: (1) segmentation, (2) feature extraction, (3) feature selection and computation of a (4) damage indicator

4.1 Segmentation (A)

Segmentation can be effectively used as a front-end signal processing technique to partition vibration signals into distinct segments. These segmented portions of the data can then undergo subsequent analysis to extract dynamic features indicative of the wheel-rail dynamics. In this context, the Hidden Markov Model (HMM) was selected as it enables high computational efficiency and provides a way to comprehend the signal's underlying structure by associating sequences of observations with distinct hidden internal states (Rabiner, 1989). To achieve this, the Baum-Welch algorithm is utilized under an iterative Expectation-Maximization approach to estimate model parameters from observed data (Bilmes, 1998). Additionally, the Viterbi algorithm and the Forward-Backward algorithm are used to specify start probability vectors, transition probability matrices, and model likelihoods (Stamp, 2004).

Employing this approach results in consistent and meaningful segments that can be indexed automatically to the corresponding wheel. Ultimately this allows for a more efficient and accurate analysis of the wheel-rail dynamics, identifying cutting points in a way that enables direct comparison of similar patterns in individual wheel passages. For hyperparameter tuning of the number of states, between 2 and 5 gaussian emissions were considered. However, between closely connected wagons, the departure of one wheel interferes with the initial approach of the consecutive wheel's signal. To tackle this partial confoundment of consecutive wheels with one another, a solution was devised by grouping the axles and developing two separate pipelines. The group of one axle is designed to handle individual axles, specifically the first and last wheels of a train. The group of two axles is dedicated to managing the pairs of axles located between closely connected wagons. With this segmentation technique, three main purposes were achieved:

- guided short-time windowing for the STFT, making the approach less sensitive to environmental and operational variations;
- localized feature extraction in both the statistical and frequency domain. This was particularly important in handling the temporal dependence of the data, which is a crucial factor in ensuring a more comprehensive understanding of the system and accurate fault detection;
- implicit indexing of each wheel passage proved to be a crucial aspect in the subsequent fault isolation and localization process. This allowed for more targeted and efficient maintenance, resulting in reduced downtime and cost savings.

4.2 Time-frequency analysis (A)

In damage detection systems, feature extraction is critical in converting the time series data into alternative information sensitive to damage (Javed et al., 2014). However, extracting useful features from raw data poses significant challenges in removing environmental and operational effects from the identified dynamic properties, to make it easier to detect potential damage and take preventative measures (Yen & Lin, 2000).

4.2.1 Short-time Fourier Transform

To address the presence of non-linear factors in the railway environment, the Shot-time Fourier Transform (STFT) was adopted (Griffin & Lim, 1984). By computing the Fourier transform for overlapping short-time windows of a segmented signal, STFT creates a power spectrum of the signal over time, providing a more accurate analysis of the magnitude spectra of each time segment. STFT has been found to yield promising results in identifying different types of faults and handling noise in various applications related to wheel condition monitoring, exhibiting comparable performance when compared to commonly used alternatives such as wavelet and Wigner-Ville transform (Liang et al., 2013; Xin et al., 2021).

In the field of time-frequency representation theory, a well-known trade-off exists between accuracy in one variable and the other. When using wide time windows, the frequency resolution improves, but the time resolution suffers, making it difficult to precisely locate wheel defects. Conversely, narrow time windows offer better time resolution but distort the signal's frequency content within that timeframe. Therefore, determining the window length requires finding a compromise that balances both time and frequency resolutions. To address this resolution trade-off and enhance fault detection, HMM were employed to define the window duration. This approach provided better time and frequency resolution for different parts of the vibration signal, enabling more precise detection of faults. The signal was split into HMM-based sequences, and overlapping adjacent segments generated windows by 50%. The resulting number of windows is twice the number of segments minus one, providing a more comprehensive and detailed analysis of the signal. The outcome after applying this segmentation was a variable sized time series, where m segments were obtained for the K scenarios of one axle and p segments for the J scenarios of two axles.

4.2.2 Feature extraction

The segments were used to extract 21 features based on the windowing technique. Experimental studies have shown the potential of these features in enhancing vibration-based structural damage detection methods, particularly under environmental effects. (Zhang et al., 2022). In the time domain, 14 features were extracted from each segment, including mean, max, min, power, and margin. Envelope metrics, such as peak-to-peak amplitude, root mean square, variance, standard deviation, skewness, kurtosis, crest factor, form factor, and pulse indicator, were also selected to capture the overall shape of the data. In the frequency domain, 7 features were extracted for each window, including mean, peak, sum, and envelope metrics, allowing for detecting magnitude variation over frequency. These included variance, skewness, kurtosis, and relative margin.

This led to the automatic construction of two three-dimensional feature matrices, one for a single axle and the other for two axles. The dimensions of the matrix in groups of one axle K-by-21-by-m, where K is the total number of scenarios of one axle, 21 the total number of time-frequency features and m is the number of segments extracted. Simultaneously, J-by-21-by-p is the matrix for groups of two axles, where J is the total number of scenarios of two axles and p is the number of segments extracted. These matrices contained valuable information for further analysis and damage detection.

4.3 Feature Selection (A)

Identifying the minimum feature subset that characterizes incipient failures is crucial for effective monitoring and model interpretation. To achieve this, the Principal Feature Analysis (PFA) algorithm was chosen due to its computational efficiency, generalizability, and suitability for unsupervised settings. This filter method uses the *k*-means method to select a representative feature set based on the structure of a Principal Component Analysis. The goal is to obtain a principal feature subset with maximum variability in a smaller dimensional space (Lu et al., 2007).

4.3.1 Principal Component Analysis

Environmental factors such as temperature and wind, and operational variables like train speed and type, can significantly influence vibration characteristics. Although the variables are not directly measured, their impact can be detected by analyzing changes in the observed features. To surmount this challenge, a latent variable approach of Principal Component Analysis for feature space resolution can be extremely useful (Jolliffe, 2005). Through analysis of the matrices, containing the features extracted from the dynamic responses of each of the segments, PCA rotates the original coordinate system using a 21-by-21 orthonormal linear transformation matrix. After obtaining the covariance matrix of the extracted features via singular value decomposition, the general rule of cumulative variance reaching 80% is used to select the eigenvalues (Härdle & Simar, 2019).

4.3.2 k-means clustering

To leverage the reduced dimension space for feature selection, *k*-means algorithm is adopted to regroup the data into arbitrary k clusters. The idea is to find k centroids that minimize the Euclidean distance of each vector to its nearest centroid. This method requires k to be defined in advance. However, as we are using it as an auxiliary tool for feature selection, the range to be considered is not critical. Nonetheless, the global silhouette index is used to validate the quality of the resulting feature space, since it revealed a superior performance in previous studies (de Oliveira Dias Prudente dos Santos et al., 2016). Then, we choose the feature vector closest to each centroid as a representative feature. To find the n most distinct features, since *k*-means is a non-deterministic algorithm, 10000 runs varying k between 2 and 10 were made. It is important to note that PFA returns the minimal subset in terms of redundancy, not accounting for predictive power. The outcome is matrices K-by-n-by-m and J-by-n-by-p, where n is the number of selected features.

4.4 Damage index (A)

Anomaly detection involves ranking instances whose intrinsic properties deviate from normal behavior, in the form of a damage index. These methods typically use distance and density metrics as a ranking measure (Goldstein & Uchida, 2014). However, these have difficulty in detecting both scattered anomalies, i.e., polygonization defects, and clustered anomalies, i.e., wheel flats. To deal with this masking effect, a mass-based approach is more adequate as it has no regard for the characteristics of the regions, e.g., shape and size (Ting et al., 2009). In this thesis, separate anomaly detection models were built for groups of one and two axles to deal with the differences in vibration patterns captured with the segment-based features. The former corresponds to matrix of K-by-m for the first and last wheels, while the latter to the matrix J-by-p for consecutive axles of closely connected wagons.

4.4.1 Isolation Forest

Considering the requirement of real-time monitoring and non-linear relationships represented in the vibration signal, the Isolation Forest (iForest) algorithm's score is leveraged for identifying mass-based anomalies. This approach is preferred for its stability under hyperparameter tuning, computational efficiency, and the absence of assumptions on data distribution (Lesouple et al., 2021). The iForest algorithm operates by constructing multiple binary trees and randomly selecting a feature and split value within a specified range until each terminal leaf contains only one instance or instance with the same value (F. T. Liu et al., 2012). The number of splits obtained is used to calculate a score that indicates the individual instances' susceptibility to be isolated. Despite its effectiveness, iForest is predisposed to bias towards groups of correlated variables due to the computation of random splits, which was mitigated with the previous step of feature selection (Puggini & McLoone, 2018). As anomalies represent the events of interest, the iForest ranking score was used for simultaneous fault detection and severity identification while the representative data structure ensured the ability to isolate the location of defects. To inspect the tradeoff between false positives and false negatives, two parameters were considered: the subsampling size ψ from 0.01 to 1 and the number of trees to build t from 50 to 500. Finally, leveraging on the simulated dataset, the isolation paths that led to the decision that a particular data point is an anomaly were analysed. This provides local interpretability by helping uncover the causal relations between specific regions of the feature space and specific fault types, ultimately helping users trust the iForest model.

4.4.2 Benchmark models

To assess the performance of the suggested damage index, its performance was compared to other anomaly detection techniques that have been extensively used in the literature. Three benchmark models were used: the mahalanobis distance (MD), support vector machine (SVM), and local outlier factor (LOF).

LOF works by examining the local density of each instance in the dataset and comparing it to the density of its neighboring instances (Breunig et al., 2000). Instances with low density, which are far from other instances, have a higher likelihood of being outliers. However, LOF may be prone to errors in regions with varying densities and skewed distributions, leading to inaccurate anomaly detection. MD is another statistical technique that is commonly used in damage identification studies (Meixedo et al., 2021a). It calculates the distance between points in a feature space, considering the correlations between the variables and the scale of each variable (De Maesschalck et al., 2000). SVM learns a decision boundary that separates normal observations from anomalous observations in high-dimensional feature space (Sotiris et al., 2010). Once the hyperplane is learned, new observations can be classified as normal or anomalous based on their distance to the hyperplane. Despite being robust to imbalanced datasets, this technology has difficulties with complex and nonlinear data distributions.

4.5 Sliding window feature extraction (B)

Approach B adapts A to a context of on-board data with a fixed sliding window, as illustrated in Figure 13. Two additional hyperparameters needed to be tuned: the window size and sliding rate.

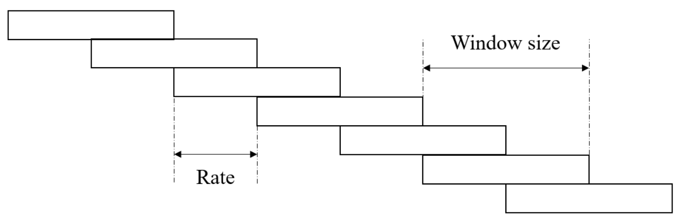


Figure 13 – Framework B with sliding window approach under two hyperparameters: window size and rate

4.6 Numeric distance time series representation (C)

Approach C proposes an alternative solution to deal with the relative scarcity of defective subsequences in real-time. By providing the model with a baseline of normal behaviour, mining rare events solely relies on the characterization of subsequences that stray too far from the model, using a similarity measure adequate for time series. This approach lies in building an automatic classification algorithm for wheel tread polygonization with radial irregularities ranging from 1 to 20 wavelengths. The design chosen follows the generic schema for online adaptive learning illustrated in Figure 14. In summary, a memory module accumulates a fixed-length batch of stationary accelerometer data from either a clear or occupied track status. If it corresponds to an occupied status, i.e., a train passage, data are presented to the learning module, implemented with a hidden Markov model (HMM), to isolate individual wheel behaviour. Then, a query-by-content approach is used to obtain predictions, with Dynamic Time Warping (DTW) scores ranking wheel passages from the most anomalous to the most normal. To check if a model update is necessary, two different change detection modules are used according to the status feedback of the track. If occupied, the loss estimation module is achieved with a simple rule that tracks the stability of states calculated under a updated HMM and indicates the end of a train passage. Alternatively, if the track is clear, a tailored concept drift detection algorithm is used to alarm the passage of a train.

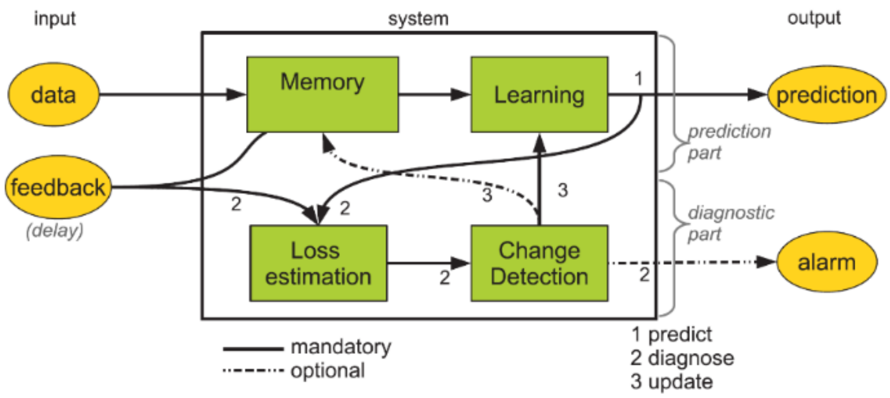


Figure 14 – Framework for online adaptive learning algorithm (Gama et al., 2014) with four main modules: a memory module accumulates a fixed-length batch of stationary accelerometer data, learning module is implemented with hidden Markov model, a loss estimation module is achieved with a simple rule and a tailored change detection algorithm is used to alarm the passage of a train

Learning under non-stationarity data requires updating the model and forgetting the old information. A simple concept drift detector with a fixed-sized sliding window was used when the track was in a clear status to have a secure and reliable detection of a train passage. A drift occurs when the incoming instance is out of the bound of the normal range determined by scaling the historical interquartile range. Compared to its alternatives, this method should provide a good baseline (Gama et al., 2014). It has a more precise localization of the change point than a sequential analysis at the expense of more memory requirements. Also, using a single value to check drift eliminates the delay of comparing two different distributions. Nevertheless, this better responsiveness comes with a trade-off in robustness to system noise. Its main limitation lies in the extensive a priori knowledge of the structure and environment of the data stream, with the user being required to guess a time scale for change. For its deployment, the PersistAD method in the adtk python package was used, with two hyperparameters tuned: the window size and the scale applied to the interquartile range. These must be carefully defined according to the characteristics of the data and the business's needs, e.g., tradeoff between false positives and false negatives. Thus, several models were trained with different values to visually inspect the performance of the algorithm. Window size values equal to 0.1, 1 and 10 seconds were admitted based on reasonable expectations of railway traffic control policies. Scales 2, 2.5 and 3 were used, assuming a Gaussian distribution all translated to more than three standard deviations. Also, as the aim is to detect the first wheel passage, the accelerometer is expected to measure a drop first and only negative drifts are considered.

To isolate a specific wheel as the probable root cause, while dealing with the dynamically changing environment of the wheel-track interface, the HMM technology was then adopted to produce an annotation vector and guide the anomaly search on time series subsequences. For its application, the hmmlearn python package was used. An analysis of the effect of parameters confirmed high accuracy with the default parameters. The HMM is triggered when a concept drift is flagged and applied to a variable-length window comprised of multiple 5 seconds batches. The rationale behind this interval is based once again on railway traffic control. Finally, the end of the train passage is triggered when the last batch is completely labelled with the same state. Upon this alarm, all readings in the sliding window corresponding to clear status of the track are discarded and the remaining are merged to form individual passages. In this merge, a challenge was posed by wagons being closely connected, which caused confoundment of consecutive wheels with one another. The lack of time to reach a steady clear status leads to a partial loss in information of the wheel behaviour. Therefore, strategies based on axle groups and individual wheels were experimented for the binary segmentation into clear and occupied status.

Regarding fault detection, a query-by-content approach was adopted by matching the database of obtained individual time series against a baseline pattern to retrieve the set of most dissimilar solutions. As anomalies represent the events of interest - symptoms of wheel polygonization – an intuitive distance measurement is needed to determine similarity between time series. For this purpose, the Euclidean distance can reveal ineffective in handling warps in the temporal dimension, i.e., shifts, scale differences and acceleration (Esling & Agon, 2012). Despite being dependent on the data, DTW has shown to be the most accurate in handling warps if the time series is small and still overseeable by the unaided eye (Tsai et al., 2011; X. Wang et al., 2013). To reduce computing time, the FastDTW extension was selected since it provides an approximate solution in linear computing time (S. Salvador & Chan, 2007). Using the DTW score as a damage ranking measure, the query is sequentially compared to the wheel passages as they are detected. The scores are obtained with the fastdtw python package, dynamically added to a sorted array and raw data discarded. However, four accelerometers are sensing the same wheel in different positions of the circumference. Thus, instead of transmitting all scores, which causes redundancy, only the maximum distance out of the four is added. This parallel memory mechanism is key since stream data will not stop flowing and cannot be acquired again.

Furthermore, despite DTW adapting well to the nature of the wheel monitoring data, some issues arise in the online ranking of wheel passages from the most anomalous to the most normal due to the influence of the train speed on scale differences. Therefore, a pre-processing stage was added where the train speed is inferred from the obtained states. Leveraging on the fact that the distance among sensors is known, the time of passage of the first group of axles in different sensors is used to estimate speed. Then, a very fast search is performed on a database of baseline patterns under the various admittable train speeds to select the most approximate query. In fact, it would seem pedantic to insist on slow exact matches, considering the non-stationary conditions of the data stream and bounded amount of memory. It is also important to notice that the whole matching strategy is only possible due to the a priori knowledge obtained from the pre-processing procedure. Figure 15 summarizes the five stages of this approach: (1) train detection; (2) attainment individual time series for each group of axles; (3) inference of speed for query selection; (4) similarity measurement; (5) ranking priority sequence for wheel maintenance interventions.

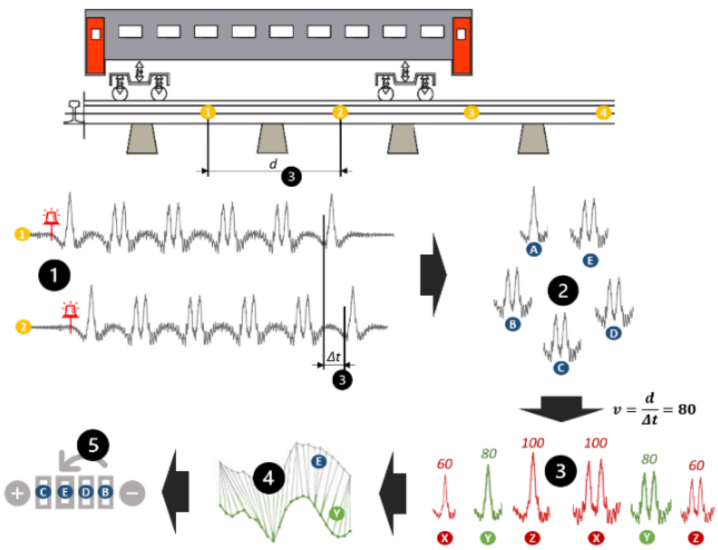


Figure 15 – Framework C with five main stages: (1) train detection; (2) attainment individual time series for each group of axles; (3) inference of speed for query selection; (4) similarity measurement; (5) ranking priority

4.7 Parallelized implementation and automatic hyperparameter tuning (D)

Approach D includes parallelize implementation of machine learning algorithms that scale well to very large databases, and automatic hyperparameter tuning process, saving time, and reducing the potential for errors, while making the algorithm accessible to non-expert users in the machine learning field. Indeed, the railway sector offers a great illustration of how green automated machine learning can be used in real-world applications, due to its huge network of long rail lines and sensors dispersed over great distances. However, few works achieved automated fault detection for a user-friendly solution (Chenariyan Nakhaee et al., 2019a). To bridge this gap, the added value of this paper encompasses a four-stage process, depicted in Figure 16 with: (1) windowing by hidden Markov model (HMM); (2) extraction of localized features with short-time Fourier transform (STFT); (3) data fusion with Principal Component Analysis (PCA); and (4) anomaly detection based on locality sensitive hashing technique (LSHAD).

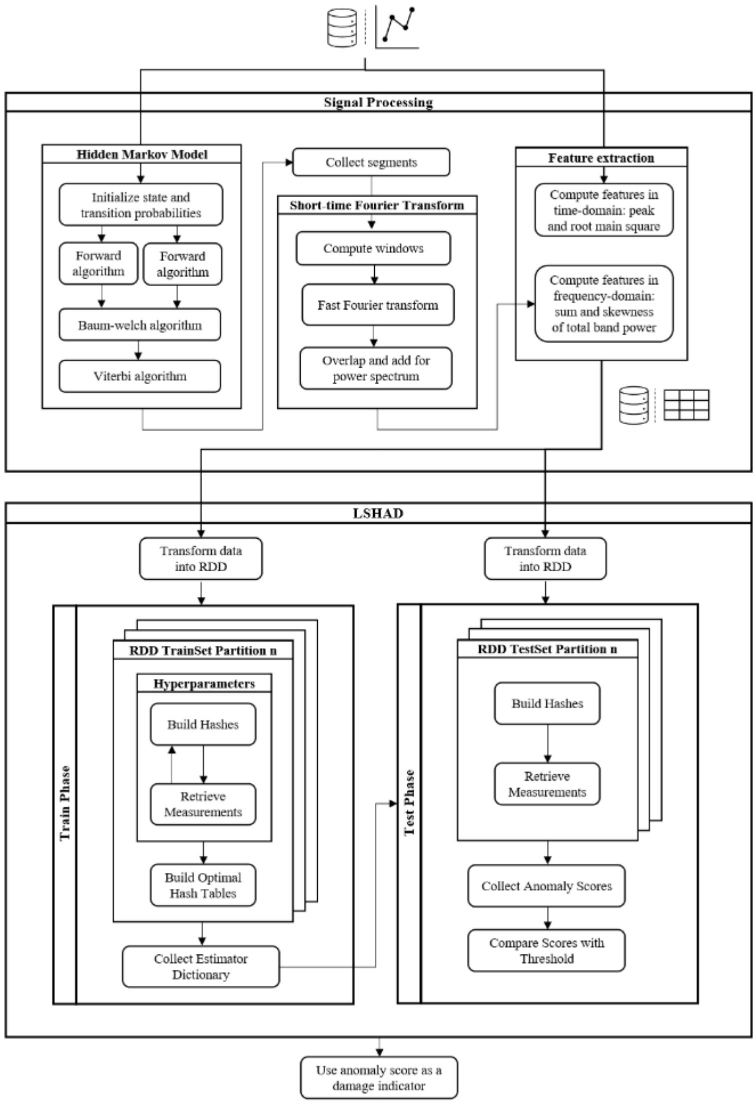


Figure 16 – Framework D with four main stages: (1) windowing; (2) extraction of localized features; (3) data fusion; and (4) anomaly detection based on locality sensitive hashing technique.

Similarly to Approach A, the HMM and STFT were deployed to extract the wheel-rail dynamics from the vibration signals, associating sequence of observations with different hidden internal states and extracting 22 features. To reduce the dimensionality and mitigate the effects of EOVs, a latent variable approach of Principal Component Analysis was adopted (Jolliffe, 2005). The number of components to extract was determined with scree plots and eigenvalue thresholds. In this reduced dimension space, LSHAD was used for fault detection (Meira et al., 2022). The algorithm is highly parallelizable and incorporates an automatic hyperparameter tuning mechanism so that users do not have to implement costly manual tuning. In addition, it has the added advantage of being an unsupervised learning process, that does not need labelled data. The tuning mechanism can adjust its hyperparameters despite the input data domain. Nonetheless, in groups of two axles there is a partial confoundment in the vibration signal of consecutive wheels with one another due to the closely connected wagons. This leads to a different data structure representation, tackled with two separate pipelines of PCA and LSHAD, allowing to better capture these differences. To evaluate the model, two criteria were used: the area under the receiver operating characteristic (AUC) and the relative increase in modelling time with larger datasets. For comparison, the models considered were: iForest, SVM and LOF.

4.8 Data streaming algorithm (E)

Approach E studies algorithms intrinsically prepared for data streaming environments, which automatically update themselves and address the issue of concept drift. Although stream data in the real world is infinite, it can be composed of an infinite number of bounded stream data sequences. Adopting this perspective, the continuous flow of accelerometer readings was simulated by concatenating different train scenarios and smoothing the data. In this thesis, the Hamming window technique was used, where a tapering function is defined to truncate the signal and then convert the discontinuities of the frequency response to transition bands (Podder et al., 2014). Figure 17 shows an exemplary passage constructed from the scenarios described in Table 12.

Table 12. Simulated data stream for the seven on-board sensors, composed of a concatenation of seven scenarios

Scenarios	1st	2nd	3rd	4th	5th	6th	7th
Speed (km/h)	60	70	75	75	75	75	75
Type Defect	-	-	P1 (13)	P1 (16)	-	P1 (19)	-

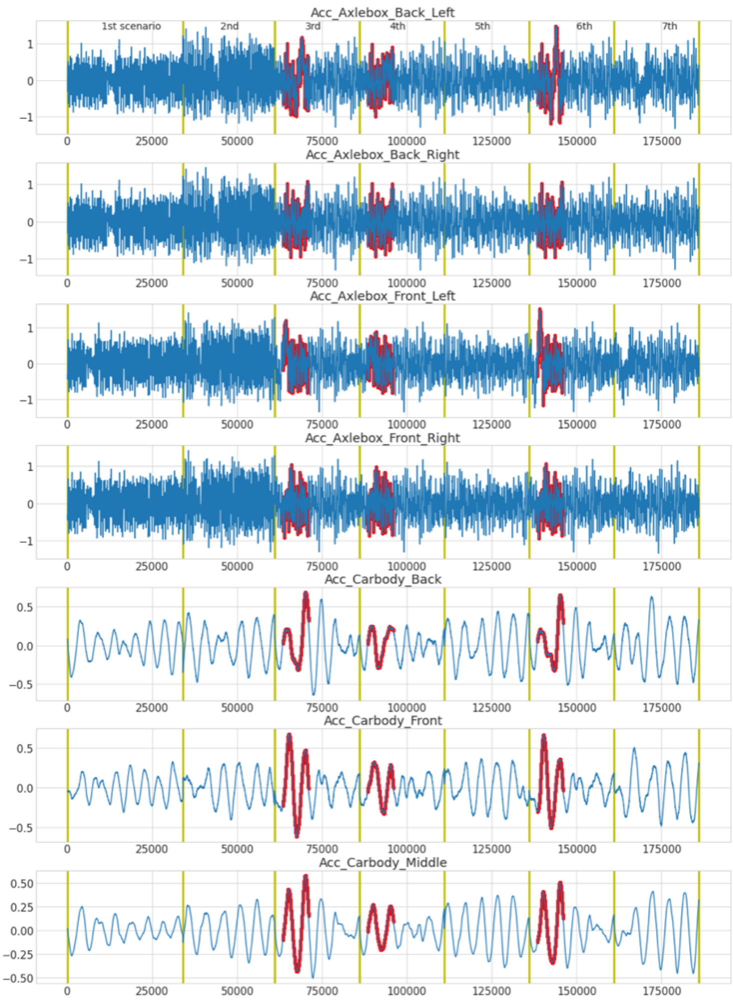


Figure 17 – Hamming smoothing for the seven on-board sensors, being highlighted in red the defective portions of the track

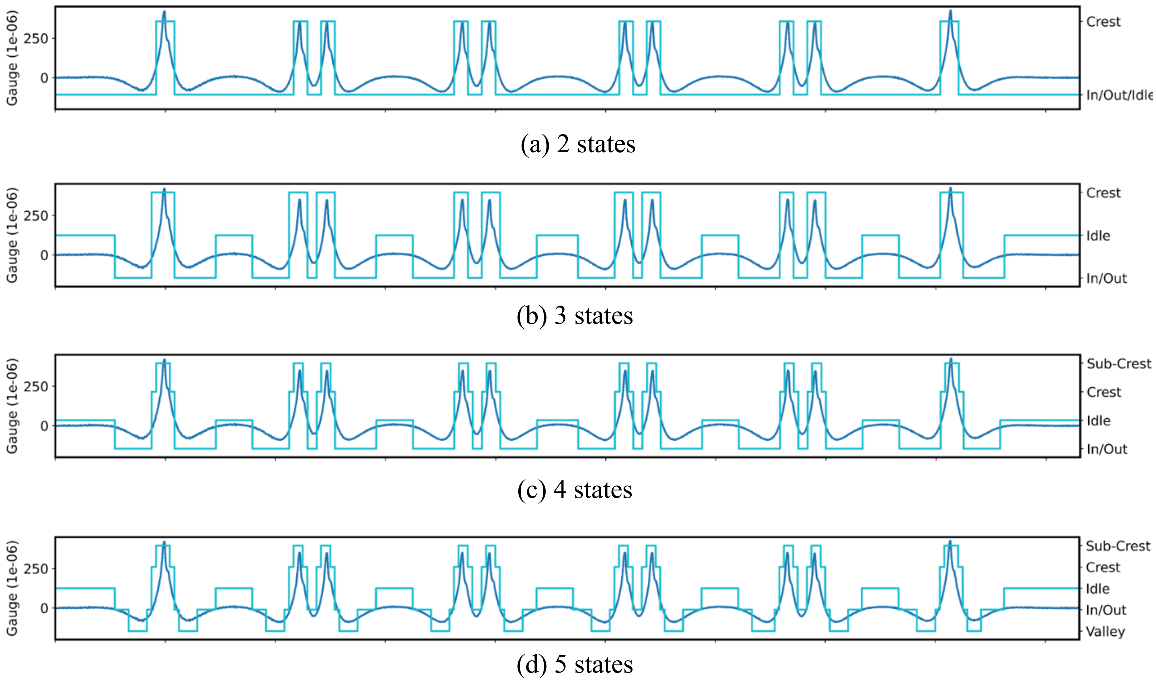
After having the desired data streams built and smoothed, four data stream algorithms were applied for fault detection: XStream, Half-Space Trees (HST), iForest for data streams (iF-DS) and Exact Storm (ES). To evaluate the performance of these models, various metrics were used: confusion matrix, accuracy, precision, recall, and the F1-score. Each of these algorithms has its own caveats. XStream is suitable for the detection of outliers in feature-evolving data streams. Being density-based, it has no need of knowing the feature space a-priori. It also uses multiple scales to count neighbors of a point and derive its density. The algorithm starts by reducing data dimensionality, then estimates the density of observations and finally approaches the problem of evolving data points (Manzoor et al., 2018). Conversely, HST is a fast and simple way of computing anomaly scores, when comparing with density or distance-based algorithms. In evolving data streams, this methodology divides the stream into equally sized windows, and then compares two consecutive windows, the reference and the latest, to gather anomaly degrees in the subspace (Tan et al., 2011). The iF-DS is an adaptation of the iForest with sliding windows, maintaining the perks of low complexity, time, and CPU consumption (Togbe et al., 2020). Finally, the Exact Storm algorithm consists of two-stage procedure: creating a data structure from the initial data and, then, storing nodes associated with different data stream objects. This is then fed to the second part of the algorithm, where the succeeding and preceding neighbors of an object are analyzed to define it as either an inlier or outlier (Angiulli & Fassetti, 2007).

5 Results and Discussion

This chapter discusses the results of the devised fault diagnosis methodologies relying on vibration-based systems. Sections 5.1 discusses the main approach for fault diagnosis. Section 5.2 evaluates the adaptation to streaming data with a fixed sliding window, while Section 5.3 analyses an alternative query by content algorithm based on time series numeric distance. Section 5.4 assesses the parallelize implementation and automatic hyperparameter tuning process. Section 5.5 studies algorithms intrinsically prepared for data streaming environments.

5.1 Results of the main approach (A)

To offer a concise discussion of the results, these are illustrated for an exemplary train passage while examining the overall performance of each step. Firstly, the HMM was fitted to the shape of the gauge signal considering between 2 and 5 states of gaussian emissions, as depicted in Figure 18. All states overcame the partial confoundment of consecutive wheels with one another due to the closely connected wagons. However, only with 2 and 3 hidden states there was a consistent segmentation for all wheel passages. The other segmentations, despite adapting well to the shape of the monitoring signal, were not reliable under a high noise ratio and defective conditions.



(a) 2 states

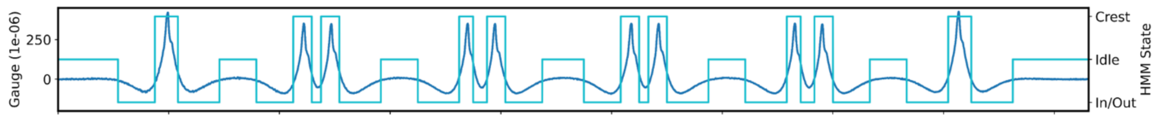
(b) 3 states

(c) 4 states

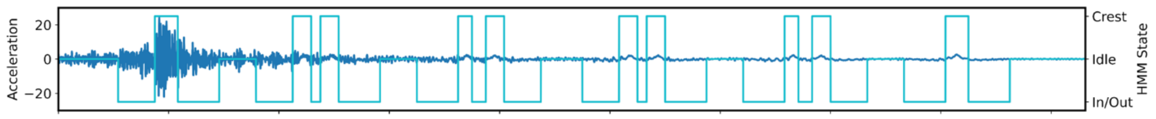
(d) 5 states

Figure 18 – Strain gauge for wheels between 2 and 5 gaussian emissions segmented into sub-crest, crest, idle, in/out and valley

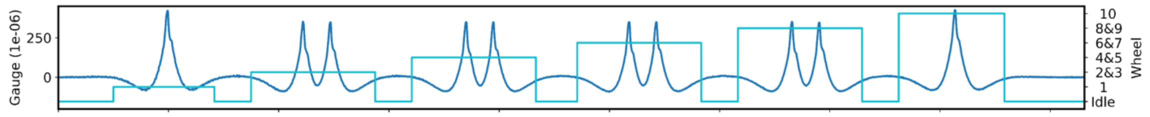
Consequently, for automatic segmentation, the HMM with 3 states was selected as it allowed to remove the portion of the wheel signal in between wheel passages. Furthermore, the obtained indexing was granular enough to capture the three stages of a wheel passage: the initial approach towards the sensor, the detection phase where the wheel is on top of the sensor and the departure, as it moves away from the sensor. This granularity proved useful in extracting the impact of an out-of-round wheel on the accelerometer signal, as shown in Figure 19. Leveraging on the state indexing, automatic segmentation could be performed for the attainment of meaningful subsequences in each passage. For groups of one and two axles, 3 and 5 segments were respectively engineered. These were then used to generate the window duration in STFT and for localized feature extraction.



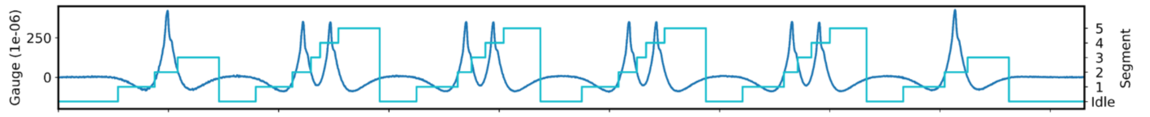
(a) Strain gauge segmentation



(b) Implicit acceleration indexing



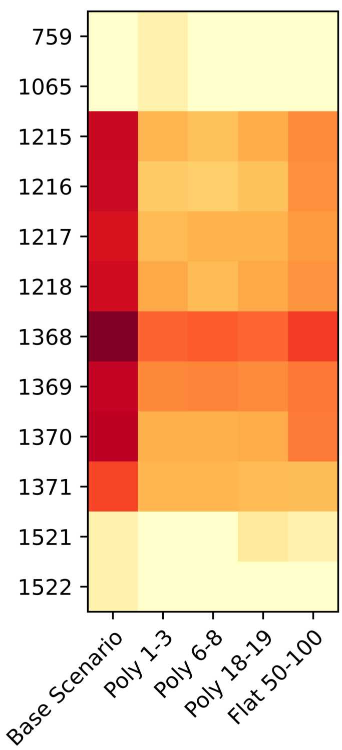
(c) Axle count



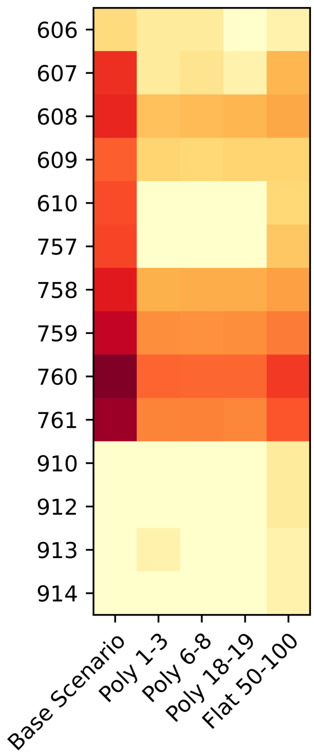
(d) Windowing technique for feature extraction

Figure 19 – Strain gauge for wheel flat with 3 gaussian emissions segmented, used to generate the window duration in STFT and for localized feature extraction on the implicit acceleration indexing

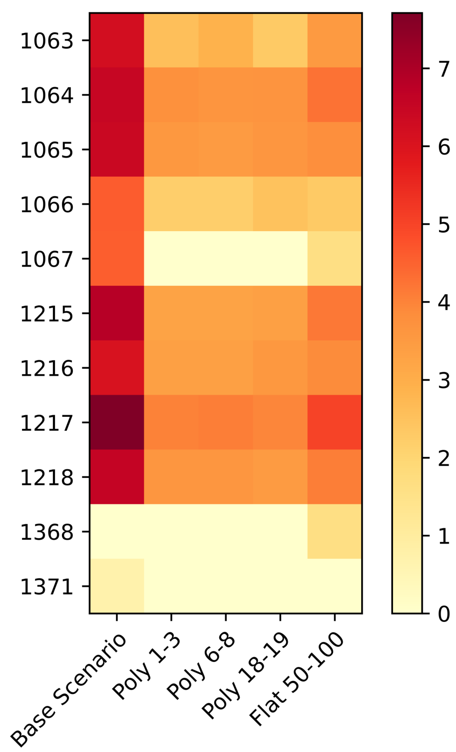
As illustrated in Figure 19, the shape of accelerometer signal is disrupted upon a damaged scenario, opposed to the strain gauge. Thus, leveraging on the unaltered shape of the strain gauge signal upon a damaged wheel, automatic segmentation could be performed for the attainment of meaningful subsequences in the accelerometer signal of each passage. For groups of one and two axles, 3 and 5 segments were respectively engineered as shown in Figure 20 and Figure 21. These were then used to generate the window duration in STFT and for localized feature extraction in the accelerometer signal, which is more sensitive to damage. To study the robustness of the segmentation, Figure 20 and Figure 21 provide the distribution of segment lengths in logarithmic scale for the different scenarios in groups of one and two axles.



(a) Wheel approaches



(b) Wheel on top

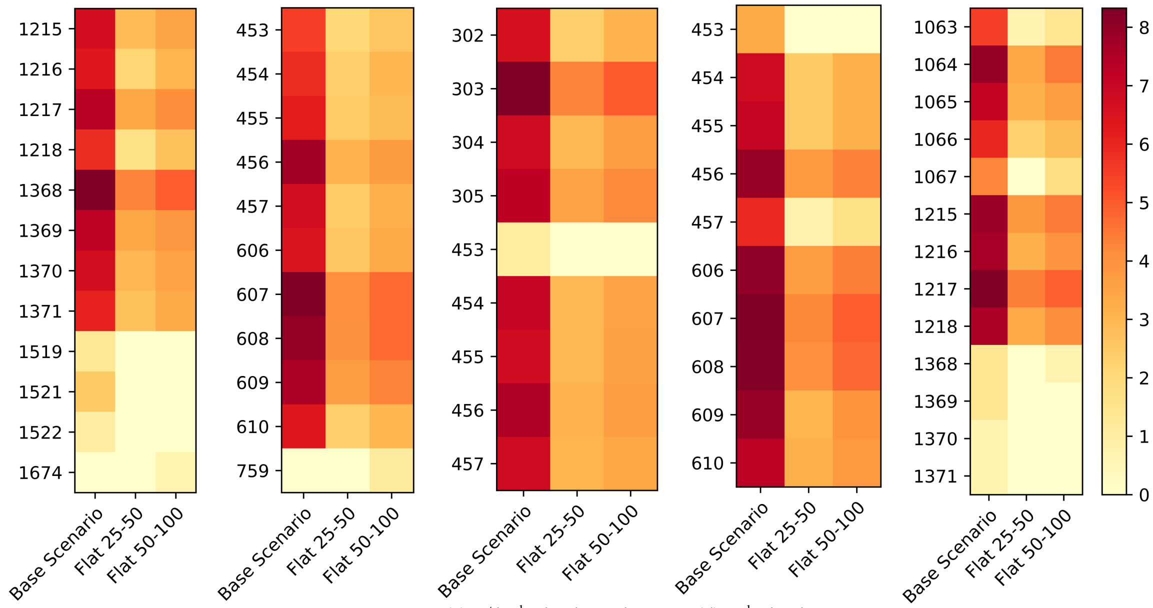


(c) Wheel moves away

Figure 20 – Distribution of segment lengths in log scale for groups of one axle: baseline, poly 1-3, 6-8 and 18-19 and flat 50-100 mm

It is important to note the color bar is in logarithmic scale. In Figure 20, for the three stages of the wheel passage in groups of one axle, there is a stable segmentation between within ranges 1215-1371, 606-761 and

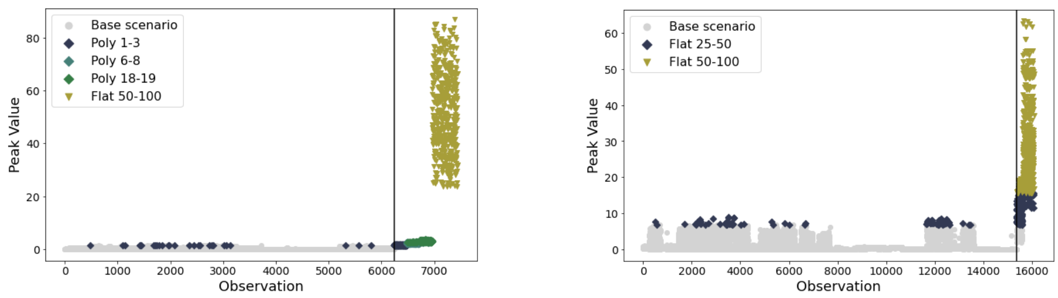
1063-1218 tenths of a millisecond, corresponding to approximately 10% deviation. Within these ranges, the interaction between the length of segments and the damaged scenarios is merely indicative of the imbalanced dataset. Apart from a few outliers, only in the detection phase with the wheel on top of the accelerometer, considerable larger segments of 910-914 were extracted. Regarding the groups of two axles in Figure 21, both the movement of the first wheel towards and of the second wheel away from the accelerometer exhibit the same pattern of the groups of one axle. However, the partial confoundment of the consecutive wheels results in smaller segments of 453-610, when the wheel is on top, and 302-457 in the overlap of the wheels.



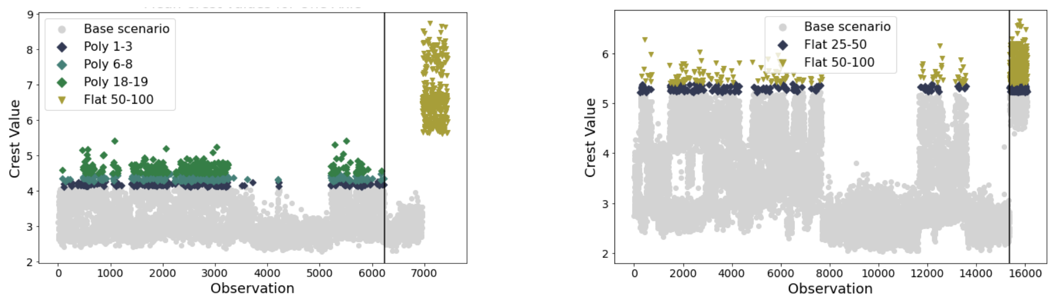
(a) 1st wheel reach (b) 1st wheel on top (c) 1st/2nd wheel overlap (d) 2nd wheel on top (e) 2nd wheel moves away

Figure 21 – Distribution of segment lengths in log scale for groups of two axles: baseline, flat 25-50 mm and flat 50-100 mm

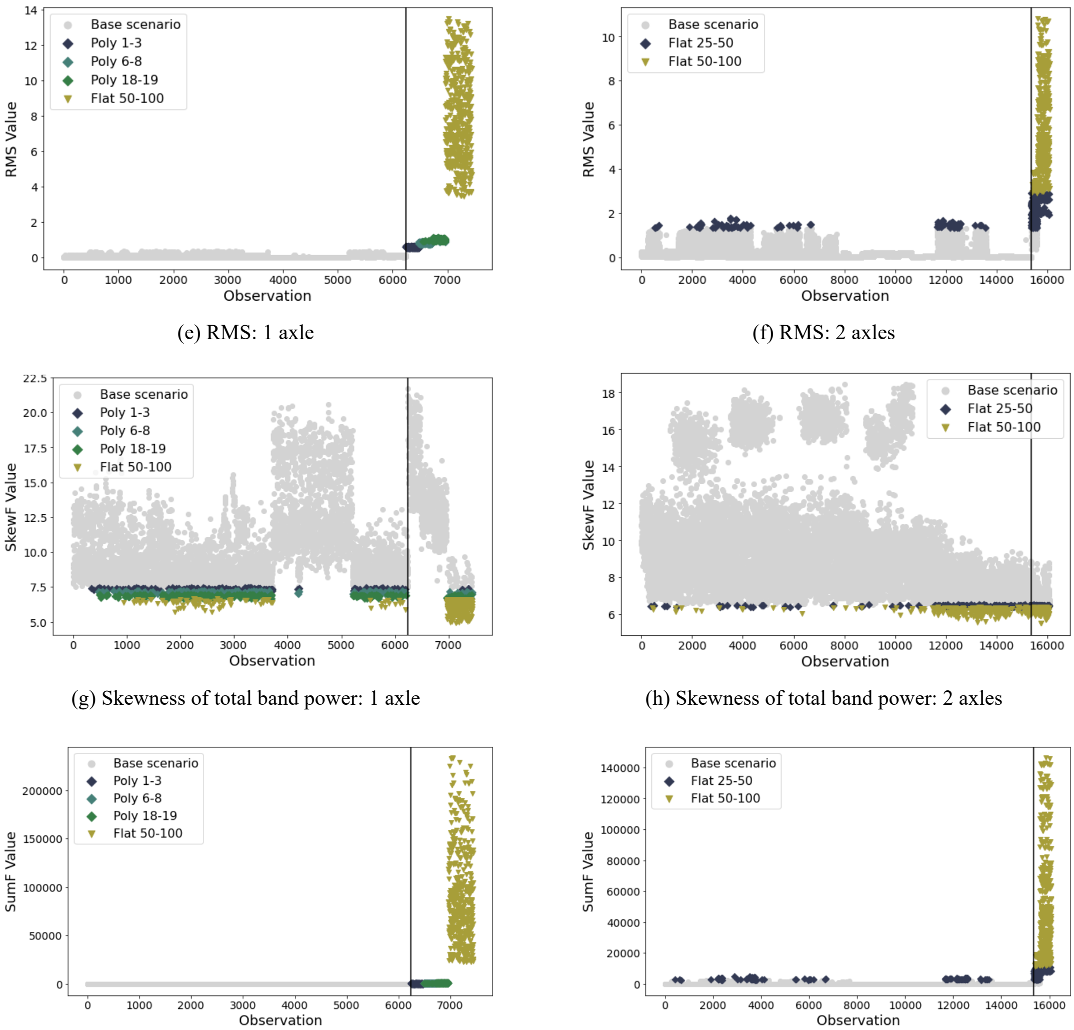
For each subsequence, a total of 14 time-domain and 7 frequency-domain features were extracted. Figure 22 depicts the sensitivity to damage of 5 out of the 21 features that were extracted from each subsequence. The features are divided into two groups based on the condition of the train's wheel, with the baseline scenarios shown before the black line and the damaged scenarios shown subsequently. In the figure, a set of symbols with the anticipated failure mode if the feature under consideration was used as a singular damage indicator.



(a) Peak: 1 axle (b) Peak: 2 axles



(c) Crest factor: 1 axle (d) Crest factor: 2 axles



(e) RMS: 1 axle

(f) RMS: 2 axles

(g) Skewness of total band power: 1 axle

(h) Skewness of total band power: 2 axles

(i) Sum of total band power: 1 axle

(j) Sum of total band power: 2 axles

Figure 22 – Individual feature values as anticipated damage indicators. In groups of one axle, the symbols, from left to right, denote the three types of polygonization, with wavelengths comprising 1-3, 6-8 and 18-19 harmonics, and the wheel flat range of 50-100 mm. For groups of two axles, the symbols, from left to right, denote the wheel flats of ranges 25-50 mm and 50-100 mm, respectively

As shown in the figure, certain features extracted from the data exhibit sensitivity to damage, with higher magnitudes and fluctuations compared to the baseline scenarios. For instance, Figure 22 demonstrates that the amplitude differences between healthy and damaged scenarios are more pronounced for certain features, e.g., peak, which remains relatively stable results. Conversely, it also presents features, e.g., skewness of total band power, whose amplitude variation due to wheels out-of-roundness is similar to that resulting from environmental and operational effects in the baseline scenarios. These findings suggest that each feature captures unique aspects of the damage condition, and that a comprehensive approach based on selected features is necessary to fully capture the diagnostic information. Both nvironmental and operational effects may obscure the damage signal and make it challenging to differentiate between baseline and damaged scenarios.

To identify the most informative and discriminative features that can best differentiate between different conditions or states, PFA was applied to the matrices of both groups of axle 7440-by-21-by-3 and 16080-by-21-by-5. For the first stage of Principal Component Analysis, as explained previously, a cumulative variance of 80% to streamline the analysis. In the second stage, 10000 runs of k-means clustering were performed, varying k between 2 and 10, to overcome the non-deterministic nature of the algorithm. This step ensured that the resulting clusters were stable and not dependent on the initial random seed value. The features that were most selected as representative of a cluster are presented in Table 13.

Table 13. Principal feature analysis for 10000 runs of k-means clustering were performed, varying *k* between 2 and 10 and using cumulative variance of 80% in Principal Component Analysis

Number of axles	Time-domain		Frequency domain	
	Peak	Root main square	Skewness of total band power	Sum of total band power
1	5952	1843	4889	4700
2	4577	2916	3227	4252

The most selected features were identified the peak and root main square in the time domain, and the sum and skewness of total band power in the frequency domain. To gain a deeper understanding of this selection process, illustrates the correlation analysis of the extracted features in each stage of a wheel passage, with features f1 corresponding to the initial approach towards the sensor, f2 to the detection phase where the wheel is on top of the sensor and f3 to the departure, as the wheel moves away from the sensor. The results in Figure 23 illustrate how these features represent clusters of highly correlated variables within each segment, except for the skewness of the total band power feature. In fact, as described earlier, this statistic was found to be irrelevant to the diagnose of defective wheels and thus removed from the analysis. The presence of the remaining features suggested the existence of underlying physical phenomena. Both the peak and root main square in the time domain were successful discriminators of the higher force exerted on the track by a flat wheel, while for other types of wheel tread defects, i.e., polygonization, there was a clear difference in frequency bands of the measurement, which was captured by the sum of total band power. Thus, these 3 features were selected for the anomaly detection stage.

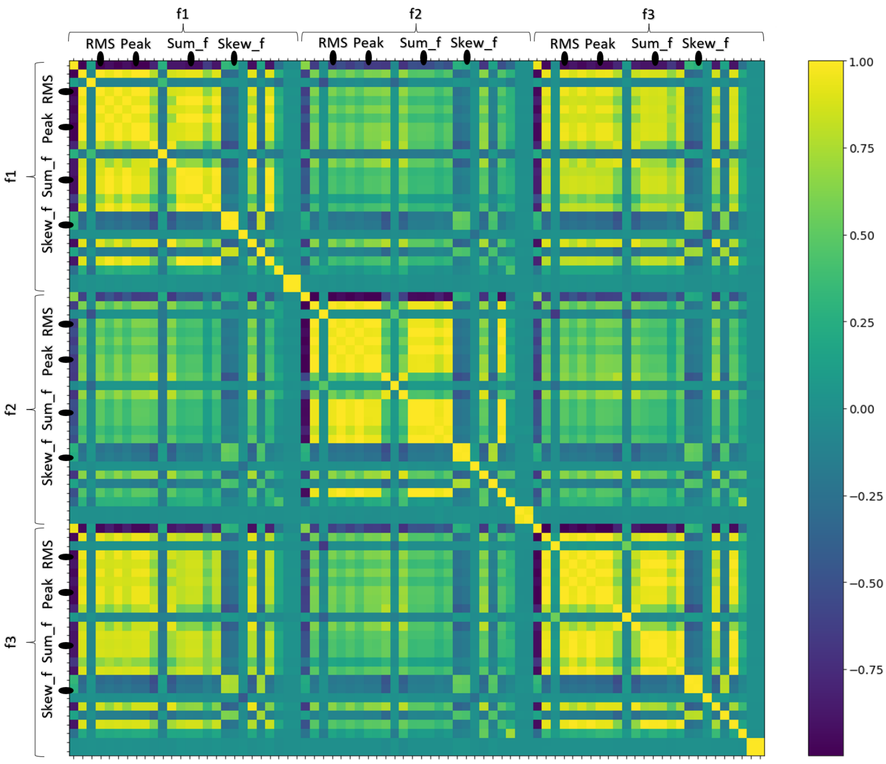


Figure 23 – Correlation matrix for groups of one axle, with features f1 corresponding to the initial approach towards the sensor, f2 to the detection phase where the wheel is on top of the sensor and f3 to the departure, as the wheel moves away from the sensor

In this study, two separate anomaly detection rankings were built for each matrix of 7440-by-3-by-3 and 16080-by-3-by-5. Building separate models for the two groups helped to deal with the differences in vibration patterns caused by the consecutive wheels in groups of two axles. Regarding hyperparameter tuning of the iForest, the findings of the analysis support the original paper's proposition that a broad range of values for the parameters of number of trees and samples required to split have little impact on detecting instances deemed to be anomalies, as shown in Table 14. In this research study, the parameters were set as 450 and

0.01, respectively, based on the F1-score of overall detection. For this calculation, the contamination percentage, i.e., the threshold to consider instances as anomalies, was defined with the actual value of 1200/7440 for the group of one axle and 720/16080 for two axles.

In addition, to allow for a fair performance benchmark of the iForest, the other algorithms were also submitted to hyperparameter tuning. In LOF, it was studied a range between 100 and 500 of nearest neighbors to consider when calculating the local density of each data point, as depicted in Table 15. A smaller number of neighbors makes the algorithm more sensitive to outliers, as it may classify points with a lower density as outliers if their nearest neighbors also have a low density. While for SVM, the gamma parameter that determines the influence of each training example on the model's decision boundary was varied between 0.1 and 0.9, as shown in Table 16. A smaller gamma value will result in a smoother decision boundary, while a larger gamma value will result in a more complex decision boundary that is more likely to overfit the data.

Table 14. iForest hyperparameter tuning: no. of trees and samples

No. of trees	Samples	F1-score	
		1 axle	2 axles
150	0.01	0.9975	0.8306
300	0.01	0.9925	0.8389
450	**0.01**	**1**	0.8389
150	0.21	0.9442	0.8667
300	0.21	0.9383	0.8653
450	0.21	0.9450	0.8708
150	0.41	0.9475	0.8722
300	0.41	0.9425	0.8708
450	0.41	0.9392	0.8764
150	**0.61**	0.9383	**0.8833**
300	0.61	0.9342	0.8792
450	0.61	0.9300	0.8819
150	0.81	0.9375	0.8750
300	0.81	0.9342	0.8736
450	0.81	0.9308	0.8778

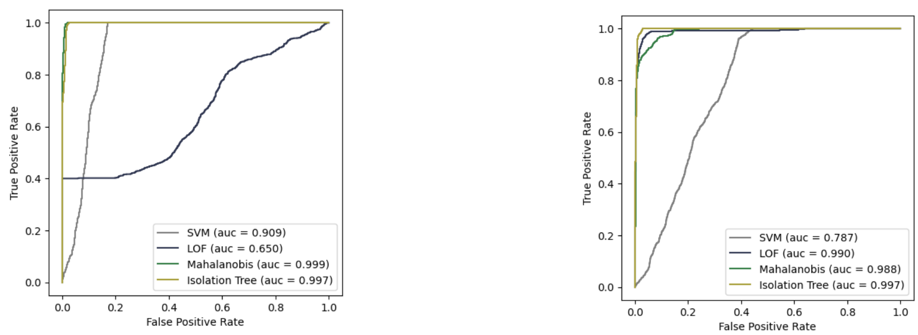
Table 15. LOF hyperparameter tuning: no. of neighbours

No. of neighbours	F1-score	
	1 axle	2 axles
100	0.0883	0.4139
200	0.0850	0.6417
300	0.1925	0.7722
400	0.3308	0.8139
500	**0.4017**	**0.8264**

Table 16. SVM hyperparameter tuning: gamma

Gamma	F1-score	
	1 axle	2 axles
0.1	0.4356	**0.1793**
0.3	**0.5830**	0.1372
0.5	0.4911	0.1535
0.7	0.5563	0.1338
0.9	0.5619	0.1320

The F1-score offers provides an indication of the overall success of the iForest in both fault detection and isolation, due to the individual treatment of groups of one or two axles. Furthermore, for the selected hyperparameters, it was further studied how the contamination percentage used as threshold affected the area under the receiver operating characteristic cure (Baeza-Yates & Ribeiro-Neto, 1999). In practice, this threshold might be limited by various factors, e.g., the number of defective wheels, stock of spare parts and availability of workforce. Figure 24 shows the AUC-ROC obtained for the various anomaly detection models.



(a) 1 axle (b) 2 axles

Figure 24 – Area under the receiver operating characteristic curve, by varying the threshold for the anomaly score

MD and LOF were successful in fault detection and isolation for groups of one and two axles, respectively. However, only iForest presented a stable performance in localized detection, with scores of 0.997 in both groups. The AUC-ROC and its corresponding score provide insights into the success of the fault identification task in terms of severity. A higher AUC indicates that the system is better at distinguishing between faulty and non-faulty instances. It implies that the ranking obtained from the anomaly scores is effective in capturing the severity of faults, as the system can differentiate between them with a higher degree of accuracy. As shown, MD and LOF were successful in fault identification for groups of one and two axles, respectively. However, only iForest presented a stable performance, with scores of 0.997 in both groups.

To provide a more detailed analysis of the success of the full fault diagnosis task with the iForest algorithm, Table 17 depicts the confusion matrix with contamination equal to the proportion of defective wheels. Out of 7440 wheels, only 22 (0.2%) polygonization treads of low harmonics were confounded with undamaged scenarios. As for wheels with flats of 50-100 mm, these were identified as most anomalous with perfect precision. Furthermore, the progression of false positives across only consecutive classes suggests that the model is successfully ranking the severities.

Table 17. Confusion matrix for groups of one axle for the iForest algorithm, with contamination equal to the proportion of defective wheels

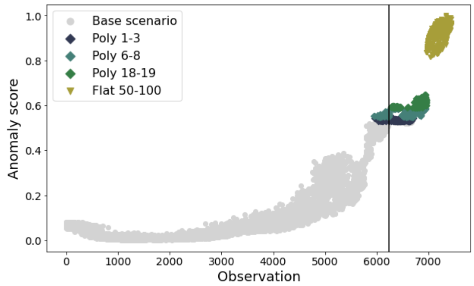
	Baseline	**Poly** 1-3	**Poly** 6-8	**Poly** 18-19	**Flat** 50-100 mm
Baseline	6218	22	0	0	0
Poly 1-3	22	205	13	0	0
Poly 6-8	0	13	187	40	0
Poly 18-19	0	0	40	200	0
Flat 50-100 mm	0	0	0	0	480

Regarding groups of two axles, results were also promising in terms of the most severe flats, as illustrated in Table 18. Only 105 out of 240 wheels with small flat depths were confounded as non-defective wheels. Notably, these wheels exhibited a flat length below the replacement threshold recommended by the General Contract for the Use of Waggons (GCU, 2022). For instance, a wheelset with a diameter larger than 840mm should only be replaced if the wheel flat length exceeds 60mm. Therefore, the proposed strategy not only demonstrated complete accuracy in detecting faults, but also had the potential to predict the emergence of flats before they reach a critical state.

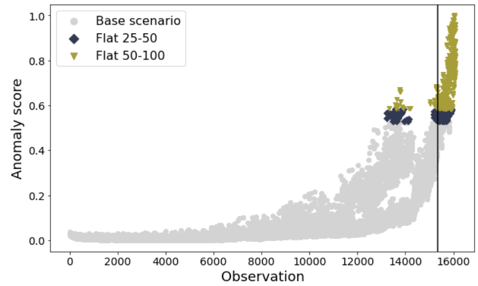
Table 18. Confusion matrix for groups of two axles for the iForest algorithm, with contamination equal to the proportion of defective wheels

	Healthy	**Flat** 25-50 mm	**Flat** 50-100 mm
Healthy	15255	105	0
Flat 25-55 mm	105	115	20
Flat 50-100 mm	0	20	460

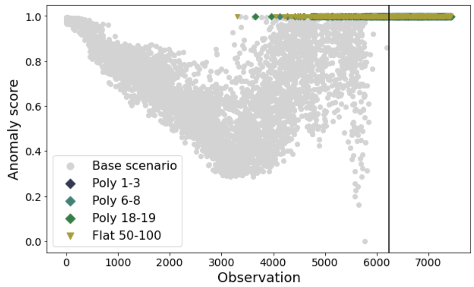
In practice, these findings suggest that when tackling the unsupervised context of the real-world application, if the maintenance technician follows the predicted rank and starts to be consistently advised to replace wheels then he can be confident that all wheels are in good conditions. Finally, Figure 25 summarizes the damage indicator for all 23520-wheel passages, with benchmark models being outperformed by the iForest. The scenarios were separated by a black line into two groups, as in Figure 22.



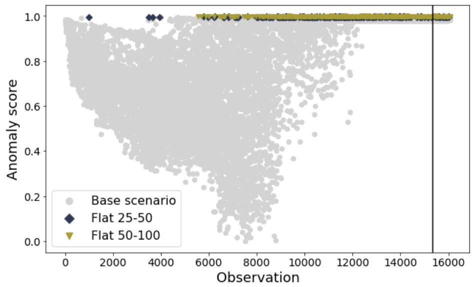
(a) iForest: 1 axle



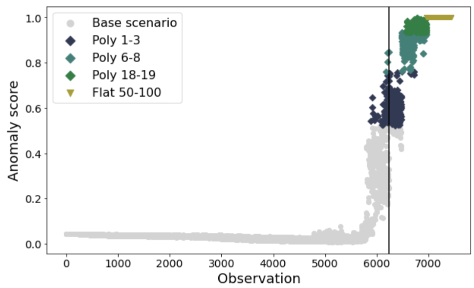
(b) iForest: 2 axles



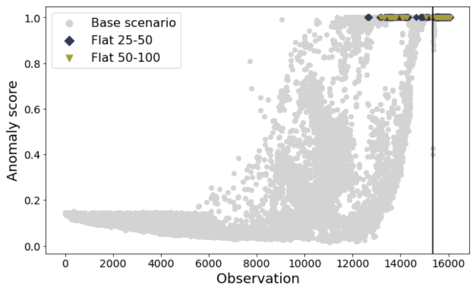
(c) SVM: 1 axle



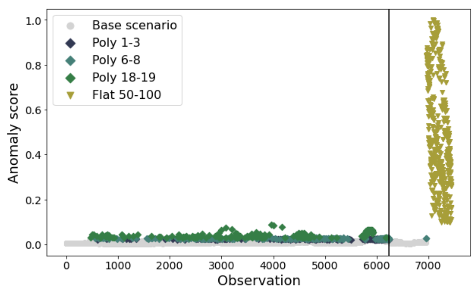
(d) SVM: 2 axles



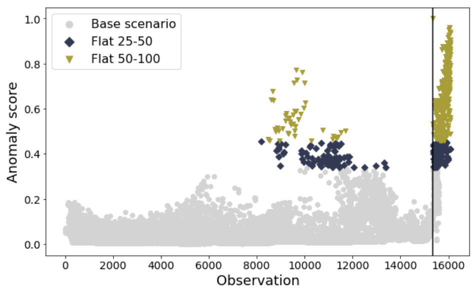
(e) Mahalanobis: 1 axle



(f) Mahalanobis: 2 axles



(g) LOF: 1 axle

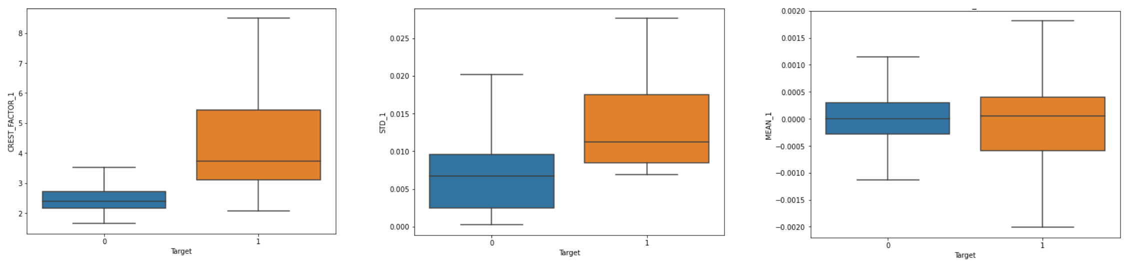


(h) LOF: 2 axles

Figure 25 – Damage indicator based on various anomaly detection algorithms. In groups of one axle, the symbols, from left to right, denote the three types of polygonization, with wavelengths comprising 1-3, 6-8 and 18-19 harmonics, and the wheel flat range of 50-100 mm. For groups of two axles, the symbols, from left to right, denote the wheel flats of ranges 25-50 mm and 50-100 mm, respectively

5.2 Results for sliding window feature extraction (B)

For each of the seven sensors, 21 features were extracted, being 14 of them relative to the time domain and 7 to the frequency domain. The middle sensor was not included in the analysis since its signal magnitude was insignificant. For dimensionality reduction, the predictive power of al features was assessed with boxplots. Several features were not adequate to distinguish between the base scenario and the anomalies since the distributions of both were similar. Figure 26 presents an illustration of boxplots for feature values extracted from the back left sensor.



(a) Crest factor (b) Standard deviation (c) Mean

Figure 26 – Boxplots for the observations of several back left accelerometer features: both crest factor and standard deviation yield strong predictive power while mean is irrelevant

After filtering out features incapable of distinguishing between base scenario observations and anomalies, a correlation matrix was computed for the remaining features, which results can be seen in Figure 27. The final features to be used on modelling were the crest factor, margin and sum of total band power.

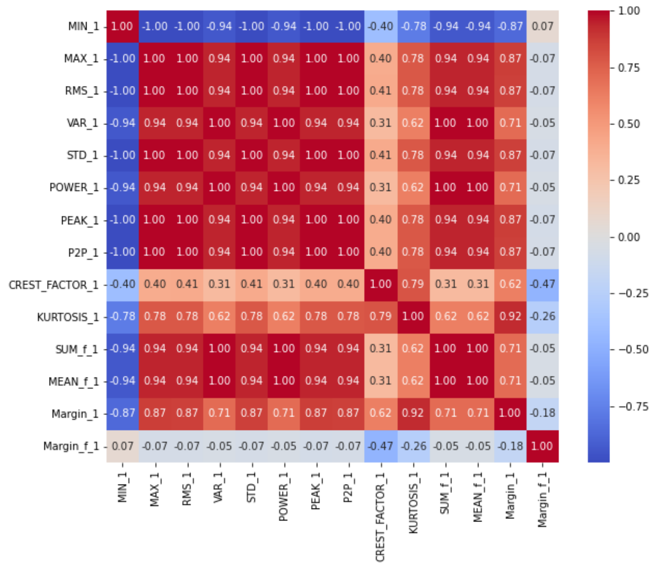


Figure 27 – Correlation matrix results for the back left sensor features

Hyperparameter tuning was performed for the iForest, LOF and SVM algorithms. Regarding iForest, the tuned parameters were the number of trees and the subsample range – Table 19. In terms of LOF, the number of neighbors was tuned – Table 20. To finish, in the case of the SVM, the gamma parameter was the tuned parameter – Table 21. The decision criteria for choosing the best combinations of hyperparameters for each algorithm was the F1 score.

Table 19. iForest hyperparameter tuning: no. of trees and samples

No. of trees	Samples	F1-score	No. of trees	Samples	F1-score
300	0.21	0.4547	300	0.61	0.4563
400	0.21	0.4547	400	0.61	0.4484
500	0.21	0.4504	500	0.61	0.4594
600	0.21	0.4563	600	0.61	0.4563
300	0.41	0.4469	300	0.81	0.4594
400	0.41	0.4438	400	0.81	0.4563
500	0.41	0.4469	**500**	**0.81**	**0.4656**
600	0.41	0.4516	600	0.81	0.4609

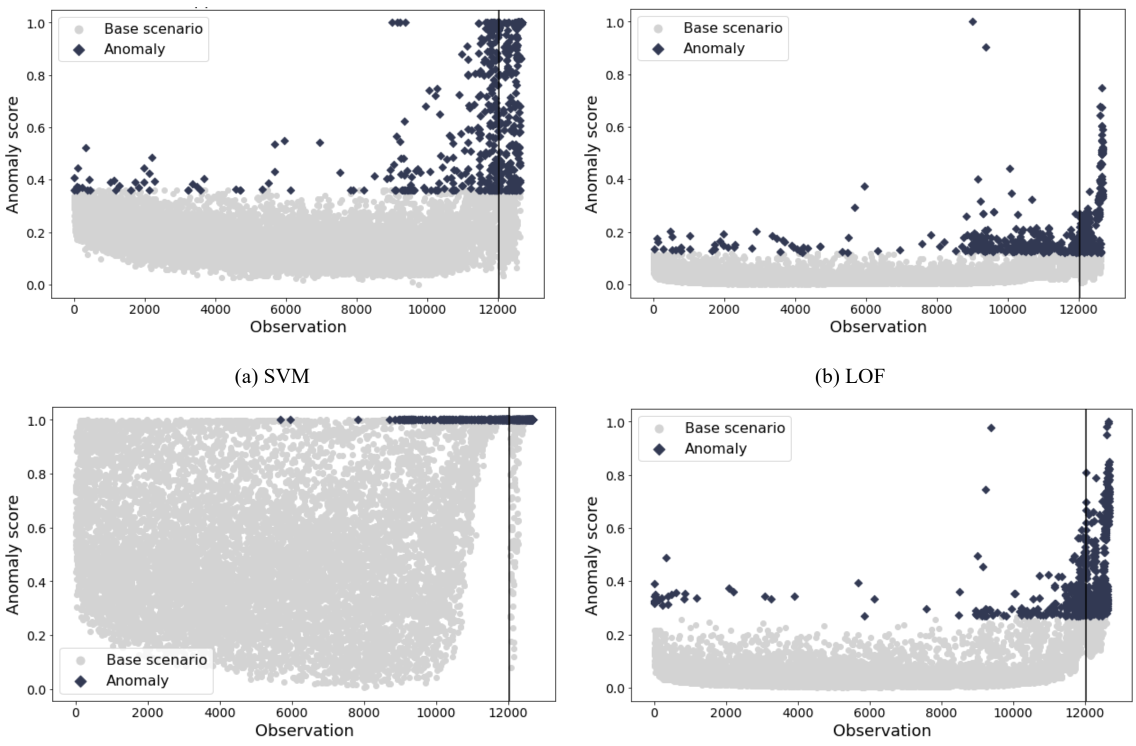
Table 20. LOF hyperparameter tuning: no. of neighbors

No. of neighbors	F1-score
100	0.3188
200	0.3281
300	0.3578
400	0.3984
500	**0.4250**

Table 21. SVM hyperparameter tuning: gamma

Gamma	F1-score
0.1	**0.4432**
0.3	0.4149
0.5	0.3580
0.7	0.3102
0.9	0.2801

After having the results for the hyperparameter tuning, all the algorithms were modelled with the optimal parameters. Mahalanobis Distance was also used in the analysis. Figure 28 presents a depiction of all the algorithms' detected anomalies. The black vertical line represents the real separation between the Baseline Scenario and the anomalies. The blue-coloured observations are the values predicted as anomalies by the algorithms. All the colored observations to the left of the vertical line and grey observations to the right of it correspond to misclassifications by the algorithm. The plots also make visible the anomaly scores attributed to each observation. Observations with values closer to one were evaluated as having higher anomaly scores by the models.

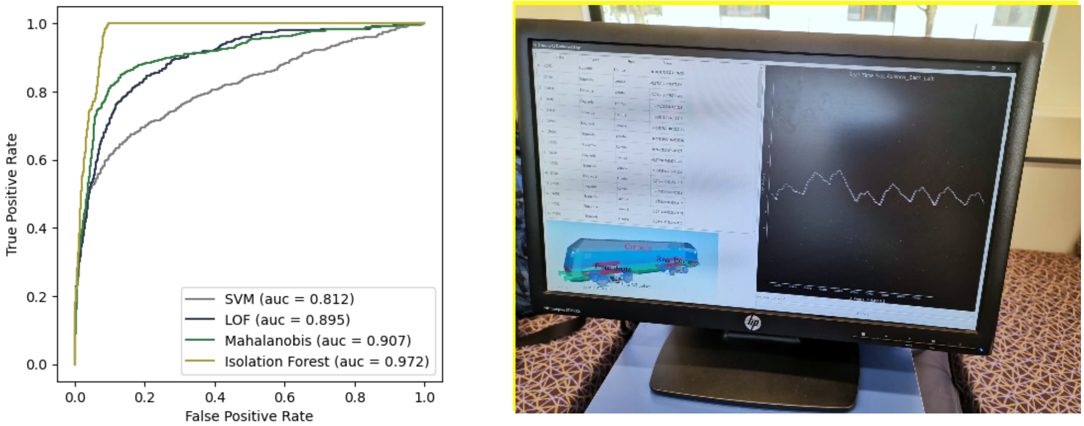


(a) SVM (b) LOF

(c) Mahalanobis (d) iForest

Figure 28 – Anomaly detection for each of the algorithms. All the colored observations to the left of the vertical line and grey observations to the right of it correspond to misclassifications by the algorithm

After analyzing the results, all of the models have attributed high anomaly scores to regular observations at times. This may have happened because of the nature of the dataset since each data row, correspondent to different windows and scenarios, includes as features characteristics of six sensors. This means that, for timeframes near anomalous windows, some features could have captured noise that led to misclassification. A possible way to circumvent this problem could be by applying noise reduction techniques to the dataset. The metric used for model evaluation was AUC-ROC, computed by varying the threshold for the anomaly score. Figure 29 contains the results for all the models. It is visible that the Isolation Forest was the best performing algorithm, with an AUC-ROC value of 0.972. In this dataset, the algorithm with the worst performance was the SVM, with a value of 0.812. However, in general, it can be observed that all algorithms have a reasonable performance in anomaly detection.

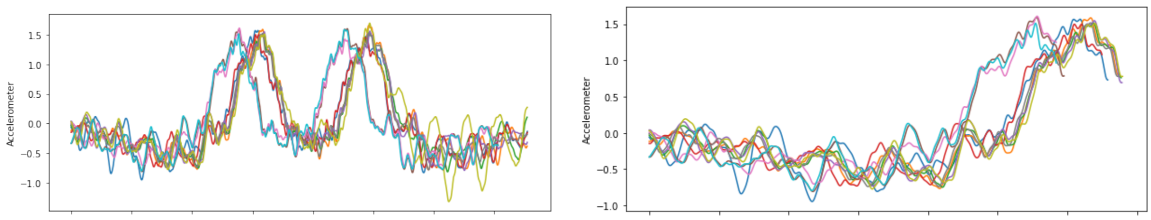


(a) AUC-ROC curve (b) Prototype

Figure 29 – Final AUC-ROC results for all algorithms, by varying the threshold for the anomaly score, and application of the selected iForest model with 500 trees and samples proportion required to split of 0.81 in a prototype

5.3 Results for numeric distance time series representation (C)

To assess the robustness of the model, 10 experimental runs were performed using baseline queries with different unevenness profiles. Figure 30 respectively represents the baselines for groups of two axles for S1 and the second wheel of wagons, except last, for S2.



(a) 2 axles (b) 1 axle

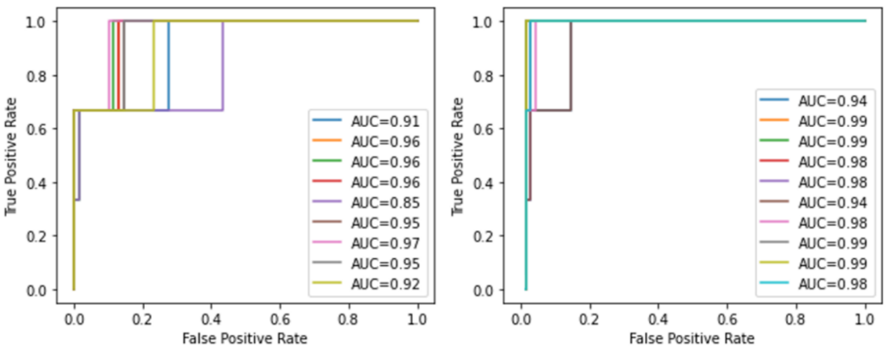
Figure 30 – Baselines with different unevenness profiles used as query

For each run, the similarity between the incoming individual passages and baseline queries of the same type was compared and dynamically added to an array. Regarding the groups of one axle, which are the 1st and last wheel, all runs yield perfect results with area under the curves equal to 1. An illustrative summary of DTW scores obtained at the end of one of the runs is shown in Table 22.

Table 22. Top 5 anomalous groups of 1 axle, with corresponding harmonic order, DTW score and time interval

Ranking	Harmonic order	DTW score	Time Interval
1st	17-20	593	3043
2nd	10-13	491	3164
3rd	5-7	382	2922
4th	None	349	3163
5th	None	313	2802

As can be seen, if maintenance interventions were performed by the given sequence, wheels would have the ideal replacement order. Regarding the polygonization damages simulated in three train passages, located in the 2nd left wheel of the 3rd wagon, different results were observed for each strategy, as illustrated by the ROC-AUC curve of 10 runs in Figure 31.



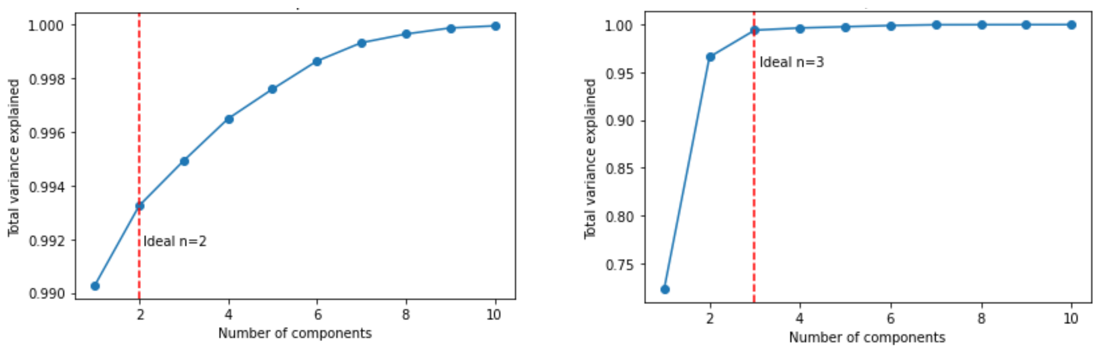
(a) 1 axle (b) 2 axles

Figure 31 – ROC-AUC for 10 runs with different baseline queries

Out of 10 runs in S1, the first ranked passage was always a true positive. Furthermore, second damage detection was ranked between 2nd and 3rd while the third was between 3rd and 22nd. On the other hand, S2 showed more promising results with third damage detections ranked within 3rd and 12th. These substantial drops in rank power show the confounding effect of measuring two wheels consecutively. Thereof, an interesting conclusion can be drawn to tackle the unsupervised context of a real-world application. As the maintenance technician starts to be consistently advised to replace wheels that turn out to be in good conditions, he can start to foresee a ROC behaviour like Figure 31: false positive rate increasing while true positive rate remains unaltered for a large number of inspections. In this case, the best option would be to completely disregard the assessment made on the remaining wheels, by dynamically defining the last checked DTW score as a permanent threshold for normality.

5.4 Results for parallelized implementation and automatic hyperparameter tuning (D)

Firstly, the HMM was applied exactly as depicted in Approach A. Then, for each subsequence, a total of 22 features were extracted obtaining matrices of n-by-66 and m-by-110, where n and m are number of passages with groups of one and two axles, respectively. To reduce the dimensionality and disregard the noise effect of EOVs, PCA was applied to both datasets. The scree plots for groups of one and two axles are presented in Figure 32, showing that the selected principal components explained well over 90% of the data variance. The obtained matrices were n-by-2 and m-by-3.



(a) One axle (b) Two axles

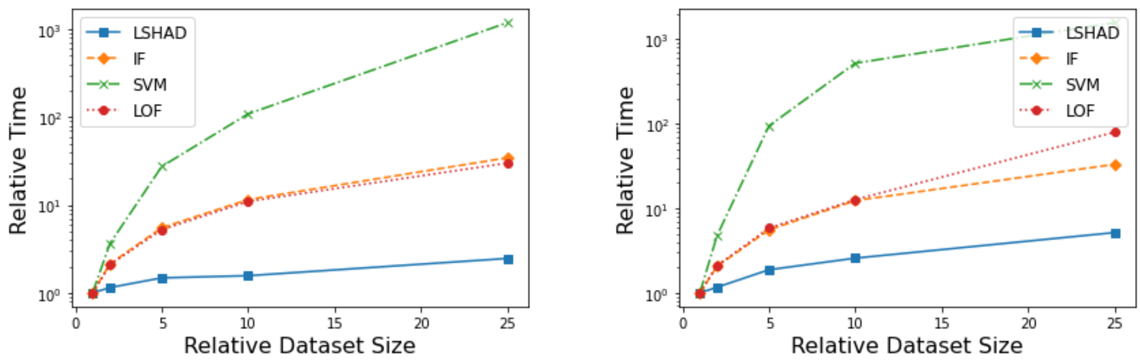
Figure 32 – Data fusion, with the threshold for cumulative variance explained in principal components marked by a red line

For anomaly detection, LSHAD was performed using the Apache Spark implementation, with two folds applied, while the remaining algorithms were implemented in scikit-learn python package. To allow for a fair performance benchmark of the LSHAD, the other algorithms were also submitted to hyperparameter tuning. In iForest, a grid search was performed varying the number of trees from 150 to 450 and samples proportion required to split from 0.01 to 0.81. In SVM, the gamma parameter that determines the influence of each training example on the model's decision boundary was varied between 0.1 and 0.9. In LOF, it was studied a range between 100 and 500 of nearest neighbours to consider when calculating the local density of each data point. Table 23 presents the AUC with the best hyperparameter set.

Table 23. AUC-ROC for the various algorithms, considering the best hyperparameter sets and a contamination equal to the proportion of defective wheels

	LSHAD		iForest		SVM		LOF	
	1 Axle	**2 Axles**	**1 Axle**	**2 Axles**	**1 Axle**	**2 Axles**	**1 Axle**	**2 Axles**
1x	0.759	0.885	0.835	0.873	0.513	0.316	0.861	0.873
2x	0.754	0.882	0.833	0.874	0.512	0.255	0.854	0.863
5x	0.764	0.912	0.864	0.880	0.513	0.503	0.779	0.721
10x	0.726	0.909	0.851	0.875	0.513	0.504	0.683	0.585
25x	0.769	0.909	0.869	0.877	0.513	0.504	0.555	0.526

In terms of time, the hyperparameter tuning process was excluded from the analysis. Figure 33 shows the duration of fitting and predicting operations. Experiments were performed in a laptop with 8GB of RAM and a 1.30GHz Intel Quad-Core i7 processor.



(a) One axle

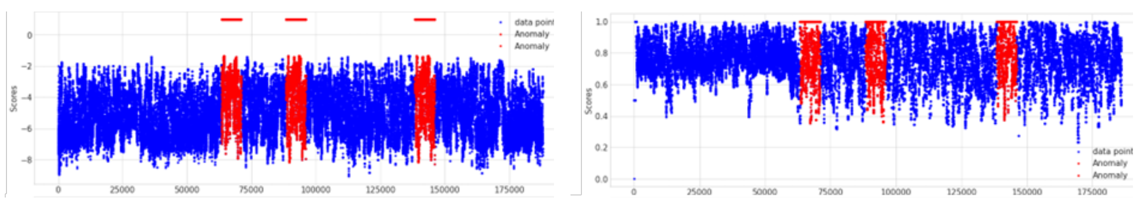
(b) Two axles

Figure 33 – Processing time, across all dataset sizes, relative to the 1x dataset

Across all dataset sizes, LSHAD has demonstrated comparable AUC results. Additionally, LSHAD has significantly outperformed the other algorithms in terms of processing times, with detection times ranging from 30 seconds to 3 minutes, faster than other algorithms with detection times up to 14 minutes for the largest dataset. These results demonstrate the superior performance of LSHAD in detecting anomalies efficiently and sustainably, making it more suitable for real-world applications.

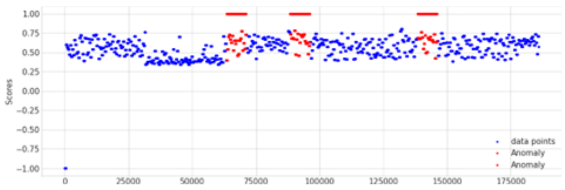
5.5 Results for data streaming algorithms (E)

To offer a concise discussion of the results, these are illustrated for the exemplary train passage depicted in Table 12. By observing Figure 34, the anomaly score provided by the different models can be assessed. It is important to note that these scores were incrementally computed after each instance or window.

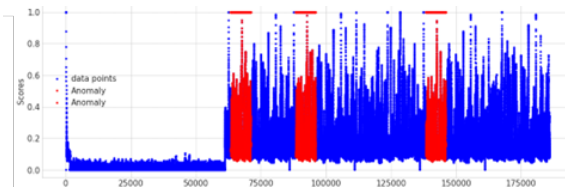


(a) XStream

(b) HST



(c) iForest-DS



(d) Exact Storm

Figure 34 – Anomaly scores obtained for the accelerometer in left wheel of back axle with the four data stream algorithms

The evaluation of the algorithms considered various performance metrics, depicted in Table 24. Accuracy, precision, and recall are commonly used indicators, but relying solely on these metrics would introduce bias in the interpretation of results. Thus, to provide a more comprehensive analysis, the F1 score, which combines precision and recall, was adopted as the main metric for selecting the best anomaly detection model in this dataset.

Table 24. Evaluation of the four data stream algorithms: F1-score, accuracy, precision, recall and run time

Algorithm	**F1 Score (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Run Time (min)**
XStream	88.6	96.8	79.5	100.0	11.04
Exact Storm	83.6	95.2	71.9	100.0	0.48
HST	78.2	64.2	64.2	100.0	1.07
iForest-DS	69.5	53.3	53.3	100.0	77.67

Analyzing the results, it becomes evident that XStream is the top-performing model, achieving an outstanding F1 score of 88.6%. Following closely behind is Exact Storm with an impressive F1 score of 83.6%. Furthermore, an examination of the confusion matrices reveals that only XStream and Exact Storm misclassified fewer than 10,000 observations. This reinforces their superiority over the other models in terms of accuracy and reliability. However, it is important to go beyond performance metrics when evaluating these algorithms. In real-world scenarios, the time required to obtain results becomes a critical factor, particularly in fault detection, where prompt interventions can significantly reduce costs and minimize disruptions to the railway system. Considering both performance and runtime, Exact Storm emerges as a compelling choice. It demonstrates a high level of accuracy, as indicated by its F1 score, while also delivering results swiftly, with a run time of only 0.48 minutes. This makes both XStream and Exact Storm practical options in situations where timely interventions are crucial.

6 Conclusion

The goals of this thesis were to investigate the challenges and opportunities of employing artificial intelligence for condition-based maintenance in railway systems, with a particular emphasis on the wheel-track interface.

Chapter 2 focused on the potential solutions to address these challenges, with a comprehensive literature review centered on six specific research questions:

(SQ1) What are the most common types of sensors used to collect time series data in railway systems, particularly for the wheel-track interface, and how do they differ in terms of accuracy and reliability?

(SQ2) What are the most commonly used machine learning algorithms for condition-based maintenance in railway systems, and how do they compare in terms of accuracy and computational efficiency?

(SQ3) What are the most effective pre-processing techniques for time series data in the context of railway maintenance, and how can they be optimized for different types of sensors and maintenance tasks?

(SQ4) How can artificial intelligence-based maintenance models be adapted to account for streaming data and automated maintenance decision making?

(SQ5) How can the computational expense, interpretability and cost considerations be incorporated into artificial intelligence-based maintenance models to improve trust and adoption by maintenance personnel and decision makers?

(SQ6) What are the key opportunities and challenges in scaling up the use of artificial intelligence for condition-based maintenance in railway systems, and what strategies can be employed to overcome these challenges?

Chapter 3 presented the data generation process, representing vibration-based systems with onboard and wayside measurement methods. With this data, it was possible to parameterize, interpret and validate devised fault diagnosis methodologies, since it would be too cost-prohibitive to induce them in a real railway system.

Chapter 4, leveraging on the identified research gaps, described novel efficient computer algorithms with characteristics that only a few works in the current literature have achieved. Firstly, it focused on an automatic fault diagnosis algorithm for wheel out-of-roundness, under two failure modes: wheel flats and wheel tread polygonization (approach A). The proposed methodology consisted in applying an HMM to guide signal preprocessing with variable STFT to ultimately extract localized features in the statistical and frequency domain, overcoming the limitations of addressing non-linear environmental and operational effects. Using PFA for a subset of the attained time series representation, it was possible to select damage-sensitive features with minimal redundancy. With the selected features a single severity indicator was computed based on the iForest anomaly score. This completed the fault diagnosis with (1) detection of aberrant train behavior; (2) isolation of specific defective wheels; (3) identification of the severity. The main achievements, including contributions from this study, can be summarized as follow:

- A novel method of signal segmentation, leveraging the sensitivity of each type of sensor. The stable temporal structure of the strain gauge signal was used to segment both signals, while feature extraction was focused on the most damage-sensitive accelerometer, less impacted by non-linear operational and environmental factors;
- Fault localization method proposed in this study offers a wheel-based approach with implicit indexing and axle count, thereby providing infrastructure managers and maintenance personnel with more comprehensive and valuable information;
- The methodology does not require the installment of a series of sensors on the rail, reducing the cost of installation and maintenance. By only using one accelerometer and strain gauge data quality was not compromised;
- The final anomaly detection stage not only provided a comprehensive context to the success of the fault localization method in distinguishing a healthy wheel from defective ones, but also provided a damage indicator for severity identification, regardless of different wheel profiles, track irregularities, train speeds, position of the sensors and artificial noise, associated to other environmental and operational variations.

In addition, a scalable and sustainable methodology was developed for condition-based maintenance of wheel out-of-roundness with large time series datasets generated by sensors in the railway network (approach D). The approach was employed in an Apache Spark framework for parallelized implementation, enabling efficient data processing across multiple clusters. Moreover, the model automates the hyperparameter tuning process, saving time, and reducing the potential for errors, while making the algorithm accessible to non-expert users in the machine learning field. Furthermore, three other methodologies were introduced for streaming environments that automatically update themselves and address issues such as concept drift (approaches B, C and E).

Chapter 5 discussed the results and demonstrated the high potential of these innovative applications of artificial intelligence in railway maintenance. As these methods were validated with physical-based modelling, the next efforts should lie in ensuring there are not too many unfulfilled assumptions and rough approximations. Thus, future work includes a field trial to further evaluate the proposed methodology based on on-site measurements. Furthermore, the robustness and efficiency of the methodology will be assessed under distinct track environments, particularly in the presence of bridges, tunnels and other under passing structures, will be evaluated. In addition, the various advantages of the proposed approach should be unified as a comprehensive methodology.

References

Ahmad, W., Khan, S. A., Islam, M. M. M., & Kim, J.-M. (2019). A reliable technique for remaining useful life estimation of rolling element bearings using dynamic regression models. Reliability Engineering & System Safety, 184, 67–76.

Al-Douri, Y. K., Tretten, P., & Karim, R. (2016). Improvement of railway performance: a study of Swedish railway infrastructure. Journal of Modern Transportation, 24, 22–37.

Alemi, A., Corman, F., & Lodewijks, G. (2017). Condition monitoring approaches for the detection of railway wheel defects. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 231(8), 961–981.

Alves, V., Cury, A., Roitman, N., Magluta, C., & Cremona, C. (2015). Novelty detection for SHM using raw acceleration measurements. Structural Control and Health Monitoring, 22(9), 1193–1207.

Amini, A., Entezami, M., Huang, Z., Rowshandel, H., & Papaelias, M. (2016). Wayside detection of faults in railway axle bearings using time spectral kurtosis analysis on high-frequency acoustic emission signals. Advances in Mechanical Engineering, 8(11).

Andrade, A. R., & Teixeira, P. F. (2013). Hierarchical Bayesian modelling of rail track geometry degradation. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 227(4), 364–375.

Angiulli, F., & Fassetti, F. (2007). Detecting Distance-Based Outliers in Streams of Data. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, 811–820.

Arasteh khouy, I., Larsson-Kråik, P.-O., Nissen, A., & Kumar, U. (2016). Cost-effective track geometry maintenance limits. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 230(2), 611–622.

Audley, M., & Andrews, J. D. (2013). The effects of tamping on railway track geometry degradation. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 227(4), 376–391.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval (Vol. 463). ACM press New York.

Bai, L., Liu, R., Sun, Q., Wang, F., & Xu, P. (2015a). Markov-based model for the prediction of railway track irregularities. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 229(2), 150–159.

Barke, D., & Chiu, W. K. (2005). Structural health monitoring in the railway industry: a review. Structural Health Monitoring, 4(1), 81–93.

Barke, D. W., & Chiu, W. K. (2005). A review of the effects of out-of-round wheels on track and vehicle components. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 219(3), 151–175.

Bergmeir, C., Sáinz, G., Martínez Bertrand, C., & Benítez, J. M. (2013). A study on the use of machine learning methods for incidence prediction in high-speed train tracks. Recent Trends in Applied Artificial Intelligence: 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2013, Amsterdam, The Netherlands, June 17-21, 2013. Proceedings 26, 674–683.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute, 4(510), 126.

Bollas, K., Papasalouros, D., Kourousis, D., & Anastasopoulos, A. (2013). Acoustic emission monitoring of wheel sets on moving trains. Construction and Building Materials, 48, 1266–1272.

Braga, J. A. P., & Andrade, A. R. (2021). Multivariate statistical aggregation and dimensionality reduction techniques to improve monitoring and maintenance in railways: The wheelset component. Reliability Engineering & System Safety, 216, 107932.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 93–104.

Brizuela, J., Fritsch, C., & Ibáñez, A. (2011). Railway wheel-flat detection and measurement by ultrasound. Transportation Research Part C: Emerging Technologies, 19(6), 975–984.

Brizuela, J., Ibañez, A., Nevado, P., & Fritsch, C. (2010). Railway wheels flat detector using Doppler effect. Physics Procedia, 3(1), 811–817.

Bukhsh, Z. A., Saeed, A., Stipanovic, I., & Doree, A. G. (2019). Predictive maintenance using tree-based classification techniques: A case of railway switches. Transportation Research Part C: Emerging Technologies, 101, 35–54.

Caetano, L. F., & Teixeira, P. F. (2013). Availability approach to optimizing railway track renewal operations. Journal of Transportation Engineering, 139(9), 941–948.

Cai, W., Chi, M., Tao, G., Wu, X., & Wen, Z. (2019a). Experimental and numerical investigation into formation of metro wheel polygonalization. Shock and Vibration.

Cárdenas-Gallo, I., Sarmiento, C. A., Morales, G. A., Bolivar, M. A., & Akhavan-Tabatabaei, R. (2017). An ensemble classifier to predict track geometry degradation. Reliability Engineering & System Safety, 161, 53–60.

Chaolong, J., Weixiang, X., Futian, W., & Hanning, W. (2012). Track irregularity time series analysis and trend forecasting. Discrete Dynamics in Nature and Society, 2012.

Chen, S., Wang, K., Zhou, Z., Yang, Y., Chen, Z., & Zhai, W. (2022). Quantitative detection of locomotive wheel polygonization under non-stationary conditions by adaptive chirp mode decomposition. Railway Engineering Science, 30(2), 129–147.

Chen, Z., Li, Y., Xia, T., & Pan, E. (2019). Hidden Markov model with auto-correlated observations for remaining useful life prediction and optimal maintenance policy. Reliability Engineering & System Safety, 184, 123–136.

Chenariyan Nakhaee, M., Hiemstra, D., Stoelinga, M., & van Noort, M. (2019a). The recent applications of machine learning in rail track maintenance: A survey. Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification: Third International Conference, RSSRail 2019, Lille, France, June 4–6, 2019, Proceedings 3, 91–105.

Chong, S. Y., Lee, J.-R., & Shin, H.-J. (2010). A review of health and operation monitoring technologies for trains. Smart Structures and Systems, 6(9), 1079–1105.

Cohen, L. (1995). Time-frequency analysis (Vol. 778). Prentice hall New Jersey.

Costa, J. N., Ambrósio, J., Frey, D., & Andrade, A. R. (2022). A multivariate statistical representation of railway track irregularities using ARMA models. Vehicle System Dynamics, 60(7), 2494–2510.

Davari, N., Veloso, B., Costa, G. de A., Pereira, P. M., Ribeiro, R. P., & Gama, J. (2021). A survey on data-driven predictive maintenance for the railway industry. Sensors, 21(17), 5739.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. Chemometrics and Intelligent Laboratory Systems, 50(1), 1–18.

de Oliveira Dias Prudente dos Santos, J. P., Crémona, C., da Silveira, A. P. C., & de Oliveira Martins, L. C. (2016). Real-time damage detection based on pattern recognition. Structural Concrete, 17(3), 338–354.

Du, H., Li, Z., Chen, R., Yin, Z., Fu, Z., Zhang, Q., Xiao, X., Luo, M., & Bao, F. (2019). Dynamic Time Warping and Spectral Clustering Based Fault Detection and Diagnosis of Railway Point Machines. 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 595–600.

Dukkipati, R. V, & Dong, R. (1999). Impact loads due to wheel flats and shells. Vehicle System Dynamics, 31(1), 1–22.

Dwyer-Joyce, R. S., Yao, C., Lewis, R., & Brunskill, H. (2013). An ultrasonic sensor for monitoring wheel flange/rail gauge corner contact. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 227(2), 188–195.

Esling, P., & Agon, C. (2012). Time-series data mining. ACM Computing Surveys (CSUR), 45(1), 1–34.

Falamarzi, A., Moridpour, S., & Nazem, M. (2019a). A review of rail track degradation prediction models. Australian Journal of Civil Engineering, 17(2), 152–166.

Falamarzi, A., Moridpour, S., & Nazem, M. (2019b). Development of a tram track degradation prediction model based on the acceleration data. Structure and Infrastructure Engineering, 15(10), 1308–1318.

Falamarzi, A., Moridpour, S., Nazem, M., & Cheraghi, S. (2018). Development of a random forests regression model to predict track degradation index: Melbourne case study. Australian Transport Research Forum, 12.

Falamarzi, A., Moridpour, S., Nazem, M., & Hesami, R. (2018). Rail degradation prediction models for tram system: Melbourne case study. Journal of Advanced Transportation.

Famurewa, S. M., Zhang, L., & Asplund, M. (2017). Maintenance analytics for railway infrastructure decision support. Journal of Quality in Maintenance Engineering, 23(3), 310–325.

Filograno, M. L., Corredera, P., Rodriguez-Plaza, M., Andres-Alguacil, A., & Gonzalez-Herraez, M. (2013). Wheel flat detection in high-speed railway systems using fiber Bragg gratings. IEEE Sensors Journal, 13(12), 4808–4816.

Filograno, M. L., Guillén, P. C., Rodríguez-Barrios, A., Martín-López, S., Rodríguez-Plaza, M., Andrés-Alguacil, Á., & González-Herráez, M. (2011). Real-time monitoring of railway traffic using fiber Bragg grating sensors. IEEE Sensors Journal, 12(1), 85–92.

Fröhling, R. D., & Hettasch, G. (2010). Wheel-rail interface management: a rolling stock perspective. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 224(5), 491–497.

Fumeo, E., Oneto, L., & Anguita, D. (2015). Condition based maintenance in railway transportation systems based on big data streaming analysis. Procedia Computer Science, 53, 437–446.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys (CSUR), 46(4), 1–37.

Gao, S., Szugs, T., & Ahlbrink, R. (2018). Use of combined railway inspection data sources for characterization of rolling contact fatigue. 12th European Conference on Non-Destructive Testing (ECNDT). Gothenburg.

Gao, X., Wang, Z., Peng, C., & Zhang, Y. (2011). Research on automatic defect localization for ultrasonic normal probe detection on railway wheel. Seventh International Symposium on Precision Engineering Measurements and Instrumentation, 8321, 747–755.

GCU. (2022). General contract of use for wagons – GCU (Edition dated. 1 January 2022).

Gerum, P. C. L., Altay, A., & Baykal-Gürsoy, M. (2019). Data-driven predictive maintenance scheduling policies for railways. Transportation Research Part C: Emerging Technologies, 107, 137–154.

Goldstein, M., & Uchida, S. (2014). Behavior Analysis Using Unsupervised Anomaly Detection.

Gómez, M. J., Corral, E., Castejon, C., & García-Prada, J. C. (2018). Effective crack detection in railway axles using vibration signals and WPT energy. Sensors, 18(5), 1603.

Gonçalves, V., Mosleh, A., Vale, C., & Montenegro, P. A. (2023). Wheel Out-of-Roundness Detection Using an Envelope Spectrum Analysis. Sensors, 23(4), 2138.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 236–243.

Guedes, A., Silva, R., Ribeiro, D., Vale, C., Mosleh, A., Montenegro, P., & Meixedo, A. (2023). Detection of Wheel Polygonization Based on Wayside Monitoring and Artificial Intelligence. Sensors, 23(4), 2188.

Guler, H. (2014). Prediction of railway track geometry deterioration using artificial neural networks: a case study for Turkish state railways. Structure and Infrastructure Engineering, 10(5), 614–626.

Gustavsson, E. (2015). Scheduling tamping operations on railway tracks using mixed integer linear programming. EURO Journal on Transportation and Logistics, 4(1), 97–112.

Härdle, W. K., & Simar, L. (2019). Applied multivariate statistical analysis. Springer Nature.

Hertz, H. (1882). Ueber die Berührung fester elastischer Körper, vol. 1882, no. 92, 1882, pp. 156-171.

Hidirov, S., & Guler, H. (2019). Reliability, availability and maintainability analyses for railway infrastructure management. Structure and Infrastructure Engineering, 15(9), 1221–1233.

Hodge, V. J., O'Keefe, S., Weeks, M., & Moulds, A. (2014). Wireless sensor networks for condition monitoring in the railway industry: A survey. IEEE Transactions on Intelligent Transportation Systems, 16(3), 1088–1106.

Hu, C., & Liu, X. (2016). Modeling track geometry degradation using support vector machine technique. ASME/IEEE Joint Rail Conference, 49675, V001T01A011.

Hu, H., Tang, B., Gong, X., Wei, W., & Wang, H. (2017). Intelligent Fault Diagnosis of the High-Speed Train With Big Data Based on Deep Neural Networks. IEEE Transactions on Industrial Informatics, 13(4), 2106–2116.

Huang, S., Zhang, F., Yu, R., Chen, W., Hu, F., & Dong, D. (2017). Turnout fault diagnosis through dynamic time warping and signal normalization. Journal of Advanced Transportation.

Jamshidi, A., Roohi, S. F., Núñez, A., Babuska, R., De Schutter, B., Dollevoet, R., & Li, Z. (2016). Probabilistic defect-based risk assessment approach for rail failures in railway infrastructure. IFAC-PapersOnLine, 49(3), 73–77.

Javed, K., Gouriveau, R., Zerhouni, N., & Nectoux, P. (2014). Enabling health monitoring approach based on vibration data for accurate prognostics. IEEE Transactions on Industrial Electronics, 62(1), 647–656.

Jayaswal, P., Verma, S. N., & Wadhwani, A. K. (2011). Development of EBP-Artificial neural network expert system for rolling element bearing fault diagnosis. Journal of Vibration and Control, 17(8), 1131–1148.

Jia, S., & Dhanasekar, M. (2007). Detection of rail wheel flats using wavelet approaches. Structural Health Monitoring, 6(2), 121–131.

Jiang, Y., Wang, H., Tian, G., Yi, Q., Zhao, J., & Zhen, K. (2019). Fast classification for rail defect depths using a hybrid intelligent method. Optik, 180, 455–468.

Johansson, A., & Andersson, C. (2005). Out-of-round railway wheels—a study of wheel polygonalization through simulation of three-dimensional wheel–rail interaction and wear. Vehicle System Dynamics, 43(8), 539–559.

Jolliffe, I. (2005). Principal component analysis. Encyclopedia of Statistics in Behavioral Science.

Kalker, J. J. (1996). Book of Tables for the Herzian Creep-force Law. Delft University of Technology, Faculty of Technical Mathematics and Informatics.

Kamlu, S., & Laxmi, V. (2019). Condition-based maintenance strategy for vehicles using hidden Markov models. Advances in Mechanical Engineering.

Kang, S., Sristi, S., Karachiwala, J., & Hu, Y.-C. (2018). Detection of anomaly in train speed for intelligent railway systems. 2018 International Conference on Control, Automation and Diagnosis (ICCAD), 1–6.

Karimpour, M., Hitihamillage, L., Elkhoury, N., Moridpour, S., & Hesami, R. (2018). Fuzzy approach in rail track degradation prediction. Journal of Advanced Transportation.

Khajehei, H., Ahmadi, A., Soleimanmeigouni, I., & Nissen, A. (2019). Allocation of effective maintenance limit for railway track geometry. Structure and Infrastructure Engineering, 15(12), 1597–1612.

Khouy, I. A., Schunnesson, H., Juntti, U., Nissen, A., & Larsson-Kråik, P.-O. (2014). Evaluation of track geometry maintenance for a heavy haul railroad in Sweden: A case study. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 228(5), 496–503.

Kim, H., Sa, J., Chung, Y., Park, D., & Yoon, S. (2016). Fault diagnosis of railway point machines using dynamic time warping. Electronics Letters, 52(10), 818–819.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1–26.

Krummenacher, G., Ong, C. S., Koller, S., Kobayashi, S., & Buhmann, J. M. (2017). Wheel defect detection with machine learning. IEEE Transactions on Intelligent Transportation Systems, 19(4), 1176–1187.

Kwon, S. J., Seo, J. W., Lee, D. H., & You, W. J. (2011). Detection of sub-surface crack in railway wheel using a new sensing system. Nondestructive Characterization for Composite Materials, Aerospace Engineering, Civil Infrastructure, and Homeland Security, 7983, 909–913.

Lagnebäck, R. (2007). Evaluation of wayside condition monitoring technologies for condition-based maintenance of railway vehicles. PhD Thesis. Luleå tekniska universitet.

Lai, C. C., Kam, J. C. P., Leung, D. C. C., Lee, T. K. Y., Tam, A. Y. M., Ho, S. L., Tam, H.-Y., & Liu, M. S. Y. (2012). Development of a fiber-optic sensing system for train vibration and train weight measurements in Hong Kong. Journal of Sensors.

Lam, H. F., Yang, J. H., Hu, Q., & Ng, C. T. (2018). Railway ballast damage detection by Markov chain Monte Carlo-based Bayesian method. Structural Health Monitoring, 17(3), 706–724.

Lasisi, A., & Attoh-Okine, N. (2018). Principal components analysis and track quality index: A machine learning approach. Transportation Research Part C: Emerging Technologies, 91, 230–248.

Lasisi, A., & Attoh-Okine, N. (2019). Machine learning ensembles and rail defects prediction: Multilayer stacking methodology. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 5(4).

Lederman, G., Chen, S., Garrett, J., Kovačević, J., Noh, H. Y., & Bielak, J. (2017). Track-monitoring from the dynamic response of an operational train. Mechanical Systems and Signal Processing, 87, 1–16.

Lee, J. S., Hwang, S. H., Choi, I. Y., & Choi, Y. (2020). Deterioration prediction of track geometry using periodic measurement data and incremental support vector regression model. Journal of Transportation Engineering, Part A: Systems, 146(1).

Lee, J. S., Hwang, S. H., Choi, I. Y., & Kim, I. K. (2018). Prediction of track deterioration using maintenance data and machine learning schemes. Journal of Transportation Engineering, Part A: Systems, 144(9).

Lesouple, J., Baudoin, C., Spigai, M., & Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. Pattern Recognition Letters, 149, 109–119.

Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., & Hampapur, A. (2014). Improving rail network velocity: A machine learning approach to predictive maintenance. Transportation Research Part C: Emerging Technologies, 45, 17–26.

Li, H., Qian, B., Parikh, D., & Hampapur, A. (2013). Alarm prediction in large-scale sensor networks—A case study in railroad. 2013 IEEE International Conference on Big Data, 7–14.

Li, Q., Peng, Q., Liu, R., Liu, L., & Bai, L. (2019). Track grid health index for grid-based, data-driven railway track health evaluation. Advances in Mechanical Engineering, 11(11).

Li, Z., & He, Q. (2015). Prediction of railcar remaining useful life by multiple data source fusion. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2226–2235.

Liang, B., Iwnicki, S., Ball, A., & Young, A. E. (2015). Adaptive noise cancelling and time–frequency techniques for rail surface defect detection. Mechanical Systems and Signal Processing, 54, 41–51.

Liang, B., Iwnicki, S. D., Zhao, Y., & Crosbee, D. (2013). Railway wheel-flat and rail surface defect modelling and analysis by time–frequency techniques. Vehicle System Dynamics, 51(9), 1403–1421.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 1–39.

Liu, S., Wang, Q., & Luo, Y. (2019). A review of applications of visual inspection technology based on image processing in the railway industry. Transportation Safety and Environment, 1(3), 185–204.

Lourenço, A., Fernandes, M., Canito, A., Almeida, A., & Marreiros, G. (2022). Using an Explainable Machine Learning Approach to Minimize Opportunistic Maintenance Interventions. In A. González-Briones, A. Almeida, A. Fernandez, A. El Bolock, D. Durães, J. Jordán, & F. Lopes (Eds.), Highlights in Practical Applications of Agents, Multi-Agent Systems, and Complex Systems Simulation. The PAAMS Collection (pp. 41–54). Springer International Publishing.

Lourenço, A., Fernandes, M., Marreiros, G., & Corchado, J. M. (2022). Using simulation to evaluate a concept drift detector for condition based maintenance. IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, 1–7.

Lourenço, A., Meira, J., & Marreiros, G. (2023). Online adaptive learning for out-of-round railway wheels detection. Proceedings of ACM SAC Conference.

Lu, Y., Cohen, I., Zhou, X. S., & Tian, Q. (2007). Feature selection using principal feature analysis. Proceedings of the 15th ACM International Conference on Multimedia, 301–304.

Luo, H., Bo, L., Peng, C., & Hou, D. (2020). Fault diagnosis for high-speed train axle-box bearing using simplified shallow information fusion convolutional neural network. Sensors, 20(17), 4930.

Ma, S., Gao, L., Liu, X., & Lin, J. (2019). Deep learning for track quality evaluation of high-speed railway based on vehicle-body vibration prediction. IEEE Access, 7, 185099–185107.

Manzoor, E., Lamba, H., & Akoglu, L. (2018). XStream: Outlier Detection in Feature-Evolving Data Streams. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1963–1972.

Matsumoto, A., Sato, Y., Ohno, H., Shimizu, M., Kurihara, J., Saitou, T., Michitsuji, Y., Matsui, R., Tanimoto, M., & Mizuno, M. (2014). Actual states of wheel/rail contact forces and friction on sharp curves–continuous monitoring from in-service trains and numerical simulations. Wear, 314(1–2), 189–197.

Meira, J., Eiras-Franco, C., Bolón-Canedo, V., Marreiros, G., & Alonso-Betanzos, A. (2022). Fast anomaly detection with locality-sensitive hashing and hyperparameter autotuning. Information Sciences, 607, 1245–1264.

Meixedo, A., Santos, J., Ribeiro, D., Calçada, R., & Todd, M. (2021a). Damage detection in railway bridges using traffic-induced dynamic responses. Engineering Structures, 238.

Meixedo, A., Santos, J., Ribeiro, D., Calçada, R., & Todd, M. D. (2022). Online unsupervised detection of structural changes using train–induced dynamic responses. Mechanical Systems and Signal Processing, 165, 108268.

Mercier, S., Meier-Hirmer, C., & Roussignol, M. (2012). Bivariate Gamma wear processes for track geometry modelling, with application to intervention scheduling. Structure and Infrastructure Engineering, 8(4), 357–366.

Mohammadi, M., Mosleh, A., Vale, C., Ribeiro, D., Montenegro, P., & Meixedo, A. (2023). An unsupervised learning approach for wayside train wheel flat detection. Sensors, 23(4), 1910.

Molodova, M., Li, Z., Núñez, A., & Dollevoet, R. (2014). Automatic detection of squats in railway infrastructure. IEEE Transactions on Intelligent Transportation Systems, 15(5), 1980–1990.

Molodova, M., Oregui, M., Núñez, A., Li, Z., & Dollevoet, R. (2016). Health condition monitoring of insulated joints based on axle box acceleration measurements. Engineering Structures, 123, 225–235.

Montenegro, P. A., Neves, S. G. M., Calçada, R., Tanabe, M., & Sogabe, M. (2015). Wheel–rail contact formulation for analyzing the lateral train–structure dynamic interaction. Computers & Structures, 152, 200–214.

Mosleh, A., Costa, P. A., & Calçada, R. (2020). A new strategy to estimate static loads for the dynamic weighing in motion of railway vehicles. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 234(2), 183–200.

Mosleh, A., Meixedo, A., Ribeiro, D., Montenegro, P., & Calçada, R. (2022b). Automatic clustering-based approach for train wheels condition monitoring. International Journal of Rail Transportation, 1–26.

Mosleh, A., Meixedo, A., Ribeiro, D., Montenegro, P., & Calçada, R. (2022c). Early wheel flat detection: an automatic data-driven wavelet-based approach for railways. Vehicle System Dynamics, 1–30.

Mosleh, A., Montenegro, P. A., Costa, P. A., & Calçada, R. (2021). Railway vehicle wheel flat detection with multiple records using spectral kurtosis analysis. Applied Sciences, 11(9), 4002.

Mosleh, A., Montenegro, P., Alves Costa, P., & Calçada, R. (2021). An approach for wheel flat detection of railway train wheels using envelope spectrum analysis. Structure and Infrastructure Engineering, 17(12), 1710–1729.

Nadarajah, N., Shamdani, A., Hardie, G., Chiu, W. K., & Widyastuti, H. (2018). Prediction of railway vehicles' dynamic behavior with machine learning algorithms. Electronic Journal of Structural Engineering, 18(1), 38–46.

Ni, Y.-Q., & Zhang, Q.-H. (2018). A Bayesian machine learning approach for online wheel condition detection using track-side monitoring. 2018 International Conference on Intelligent Rail Transportation (ICIRT), 1–6.

Nicolai, R. P., & Dekker, R. (2008). Optimal Maintenance of Multi-component Systems: A Review. In K. A. H. Kobbacy & D. N. P. Murthy (Eds.), Complex System Maintenance Handbook (pp. 263–286). Springer London.

Nielsen, J. C. O., & Johansson, A. (2000). Out-of-round railway wheels-a literature survey. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 214(2), 79–91.

Nielsen, J. C. O., Lundén, R., Johansson, A., & Vernersson, T. (2003). Train-track interaction and mechanisms of irregular wear on wheel and rail surfaces. Vehicle System Dynamics, 40(1–3), 3–54.

Oukhellou, L., Come, E., Bouillaut, L., & Aknin, P. (2008). Combined use of sensor data and structural knowledge processed by Bayesian network: Application to a railway diagnosis aid scheme. Transportation Research Part C: Emerging Technologies, 16(6), 755–767.

Palo, M., Galar, D., Nordmark, T., Asplund, M., & Larsson, D. (2014). Condition monitoring at the wheel/rail interface for decision-making support. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 228(6), 705–715.

Pan, J., Li, W., & Dai, X. (2013). Train overload and unbalanced load detection based on FBG gauge. Fourth Asia Pacific Optical Sensors Conference, 8924, 347–350.

Pargar, F., Kauppila, O., & Kujala, J. (2017). Integrated scheduling of preventive maintenance and renewal projects for multi-unit systems with grouping and balancing. Computers & Industrial Engineering, 110, 43–58.

Peng, C., Gao, X., Wang, L., Wang, Z., Zhao, Q., Zhang, Y., Peng, J., & Yang, K. (2011). Automatic railway wheelset inspection system by using ultrasonic technique. Seventh International Symposium on Precision Engineering Measurements and Instrumentation, 8321, 596–601.

Pintão, B., Mosleh, A., Vale, C., Montenegro, P., & Costa, P. (2022). Development and validation of a weigh-in-motion methodology for railway tracks. Sensors, 22(5), 1976.

Podder, P., Khan, T., Khan, M., & Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. International Journal of Computer Applications, 96, 1–7.

Puggini, L., & McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. Engineering Applications of Artificial Intelligence, 67, 126–135.

Quiroga, L., & Schnieder, E. (2010). Modelling high speed railroad geometry ageing as a discrete-continuous process. Proceedings of the Stochastic Modeling Techniques and Data Analysis International Conference, SMTDA, Chania Crete Greece, 8(11).

Rabatel, J., Bringay, S., & Poncelet, P. (2011). Anomaly detection in monitoring sensor data for preventive maintenance. Expert Systems with Applications, 38(6), 7003–7015.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257–286.

Sadeghi, J., & Askarinejad, H. (2010). Development of improved railway track degradation models. Structure and Infrastructure Engineering, 6(6), 675–688.

Sadeghi, J., & Askarinejad, H. (2012). Application of neural networks in evaluation of railway track quality condition. Journal of Mechanical Science and Technology, 26, 113–122.

Salvador, P., Naranjo, V., Insa, R., & Teixeira, P. (2016). Axlebox accelerations: Their acquisition and time–frequency characterisation for railway track monitoring purposes. Measurement, 82, 301–312.

Salvador, S., & Chan, P. (2007). Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5), 561–580.

Sammouri, W., Côme, E., Oukhellou, L., Aknin, P., & Fonlladosa, C.-E. (2014). Pattern recognition approach for the prediction of infrequent target events in floating train data sequences within a predictive maintenance framework. 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), 918–923.

Schalk, R., Vicencio, A. N., Zoeteman, A., & Wolfert, R. (2017). Data analytics for the of RCF damages on the Dutch high speed line. Proceedings of the 1st International Conference on Rail Transportation, Chengdu, China, 10–12.

Schenkendorf, R., Dutschk, B., Lüddecke, K., & Groos, J. C. (2016). Improved Railway Track Irregularities Classification by aModel Inversion Approach. PHM Society European Conference, 3(1).

Shan, W., Wu, P., Wu, X., Zhang, F., & Shi, H. (2019). Effect of wheel polygonization on the axle box vibrating and bolt self-loosening of high-speed trains. Journal of Physics: Conference Series, 1213(5).

Sharma, S., Cui, Y., He, Q., Mohammadi, R., & Li, Z. (2018). Data-driven optimization of railway maintenance for track geometry. Transportation Research Part C: Emerging Technologies, 90, 34–58.

Silva, R., Guedes, A., Ribeiro, D., Vale, C., Meixedo, A., Mosleh, A., & Montenegro, P. (2023). Early Identification of Unbalanced Freight Traffic Loads Based on Wayside Monitoring and Artificial Intelligence. Sensors, 23(3), 1544.

Soleimanmeigouni, I., Ahmadi, A., Arasteh Khouy, I., & Letot, C. (2018). Evaluation of the effect of tamping on the track geometry condition: A case study. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 232(2), 408–420.

Sotiris, V. A., Peter, W. T., & Pecht, M. G. (2010). Anomaly detection through a bayesian support vector machine. IEEE Transactions on Reliability, 59(2), 277–286.

Stamp, M. (2004). A revealing introduction to hidden Markov models. Department of Computer Science San Jose State University, 26–56.

Sun, Y., Liu, Y., & Yang, C. (2019). Railway joint detection using deep convolutional neural networks. 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 235–240.

Tan, S. C., Ting, K. M., & Liu, T. F. (2011). Fast Anomaly Detection for Streaming Data. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, 1511–1516.

Tao, G., Wen, Z., Jin, X., & Yang, X. (2020). Polygonisation of railway wheels: a critical review. Railway Engineering Science, 28, 317–345.

Thakkar, N. A., Steel, J. A., & Reuben, R. L. (2012). Rail–wheel contact stress assessment using acoustic emission: a laboratory study of the effects of wheel flats. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 226(1), 3–13.

Ting, K. M., Tan, S. C., & Liu, F. T. (2009). Mass: A new ranking measure for anomaly detection. Gippsland School of Information Technology, Monash University.

Togbe, M. U., Barry, M., Boly, A., Chabchoub, Y., Chiky, R., Montiel, J., & Tran, V.-T. (2020). Anomaly Detection for Data Streams Based on Isolation Forest Using Scikit-Multiflow. In O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre, & Y. Karaca (Eds.), Computational Science and Its Applications – ICCSA 2020 (pp. 15–30). Springer International Publishing.

Trinh, H., Haas, N., Li, Y., Otto, C., & Pankanti, S. (2012). Enhanced rail component detection and consolidation for rail track inspection. 2012 IEEE Workshop on the Applications of Computer Vision (WACV), 289–295.

Tsai, C.-F., Lin, W.-Y., Hong, Z.-F., & Hsieh, C.-Y. (2011). Distance-based features in pattern classification. EURASIP Journal on Advances in Signal Processing, 2011(1).

Tsunashima, H. (2019). Condition monitoring of railway tracks from car-body vibration using a machine learning technique. Applied Sciences, 9(13), 2734.

Verbert, K., De Schutter, B., & Babuška, R. (2017). Timely condition-based maintenance planning for multi-component systems. Reliability Engineering & System Safety, 159, 310–321.

Vileiniskis, M., Remenyte-Prescott, R., & Rama, D. (2016). A fault detection method for railway point systems. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 230(3), 852–865.

Vrignat, P., Avila, M., Duculty, F., & Kratz, F. (2015). Failure event prediction using hidden markov model approaches. IEEE Transactions on Reliability, 64(3), 1038–1048.

Wang, Q.-A., & Ni, Y.-Q. (2019). Measurement and forecasting of high-speed rail track slab deformation under uncertain SHM data using variational heteroscedastic gaussian process. Sensors, 19(15), 3311.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery, 26(2), 275–309.

Ward, C. P., Weston, P. F., Stewart, E. J. C., Li, H., Goodall, R. M., Roberts, C., Mei, T. X., Charles, G., & Dixon, R. (2011). Condition monitoring opportunities using vehicle-based sensors. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 225(2), 202–218.

Wei, J., Liu, C., Ren, T., Liu, H., & Zhou, W. (2017). Online condition monitoring of a rail fastening system on high-speed railways based on wavelet packet analysis. Sensors, 17(2), 318.

Weston, P., Roberts, C., Yeo, G., & Stewart, E. (2015). Perspectives on railway track geometry condition monitoring from in-service railway vehicles. Vehicle System Dynamics, 53(7), 1063–1091.

Xia, Y., Xie, F., & Jiang, Z. (2010). Broken railway fastener detection based on adaboost algorithm. 2010 International Conference on Optoelectronics and Image Processing, 1, 313–316.

Xie, J., Huang, J., Zeng, C., Jiang, S.-H., & Podlich, N. (2020). Systematic literature review on data-driven models for predictive maintenance of railway track: Implications in geotechnical engineering. Geosciences, 10(11), 425.

Xin, G., Li, Z., Jia, L., Zhong, Q., Dong, H., Hamzaoui, N., & Antoni, J. (2021). Fault diagnosis of wheelset bearings in high-speed trains using logarithmic short-time Fourier transform and modified self-calibrated residual network. IEEE Transactions on Industrial Informatics, 18(10), 7285–7295.

Yang, K., Ma, L., Gao, X., & Wang, L. (2012). Profile parameters of wheelset detection for high speed freight train. Fourth International Conference on Digital Image Processing (ICDIP), 8334, 382–387.

Yen, G. G., & Lin, K.-C. (2000). Wavelet packet feature extraction for vibration monitoring. IEEE Transactions on Industrial Electronics, 47(3), 650–667.

Yin, J., & Zhao, W. (2016). Fault diagnosis network design for vehicle on-board equipments of high-speed railway: A deep learning approach. Engineering Applications of Artificial Intelligence, 56, 250–259.

Zhai, W. M., Wang, Q. C., Lu, Z. W., & Wu, X. S. (2001). Dynamic effects of vehicles on tracks in the case of raising train speeds. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 215(2), 125–135.

Zhang, C., Mousavi, A. A., Masri, S. F., Gholipour, G., Yan, K., & Li, X. (2022). Vibration feature extraction using signal processing techniques for structural health monitoring: A review. Mechanical Systems and Signal Processing, 177, 109175.

Zhou, F., Xia, L., Dong, W., Sun, X., Yan, X., & Zhao, Q. (2016). Fault diagnosis of high-speed railway turnout based on support vector machine. IEEE International Conference on Industrial Technology (ICIT), 1539–1544.