INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Artificial Intelligence Applied to Railway Disruption Management: From Mathematical Optimization to Transformer Architectures in the Portuguese Operational Context

Luís Carlos Lima Marques

PhD in Information Science and Technology

Supervisors:
Doctor, Sérgio Moro, Full Professor at ISCTE-IUL
Doctor, Pedro Nogueira Ramos, Full Professor at ISCTE-IUL

November, 2025

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Artificial Intelligence Applied to Railway Disruption Management: From Mathematical Optimization to Transformer Architectures in the Portuguese Operational Context

Luís Carlos Lima Marques

PhD in Information Science and Technology

Jury:
Doctor Joana Martinho Costa, Associate Professo at Iscte - Instituto Universitário de Lisboa.
Doctor Luís Miguel Pacheco Mendes Gomes, Assistant Professor at Universidade dos Açores
Doctor Mariana Sofia Barreira Cavique Santos, Assistant Professor at Universidade europeia
Doctor António Jorge Filipe Fonseca, Assistant Professor at Iscte - Instituto Universitário de Lisboa.

November, 2025

Acknowledgment

This doctoral research has reached its final station, although I believe I have left sufficient material for other researchers to continue this theme.

There is no point in embellishing the process. It is a hard journey, often solitary, with dead ends and frequent feelings of inadequacy. Throughout this journey, several disruptive events hindered progress: the pandemic and the passing of my father. I remember his last days, when we exchanged a few words about the doctorate. I will never know if he truly understood what my research was about.

I am grateful to my supervisors for believing in the project and for always guiding me with their valuable contributions. Without them, this research would not exist. I thank my colleagues at CP for their availability and for sharing their railway knowledge.

A word of affection to all my family and, especially, to Diana, for her patience and understanding of the countless hours at the computer.

Resumo

Esta tese investiga como a inteligência artificial pode transformar a gestão de perturbações operacionais nos sistemas ferroviários de passageiros, focando-se no contexto português da Comboios de Portugal (CP). O estudo parte da constatação de que a literatura é fragmentada e excessivamente dependente de métodos de otimização linear, pouco escaláveis e adaptáveis a cenários reais, verificando-se ainda uma subutilização de abordagens baseadas em machine learning.

Desenvolvida segundo o paradigma de Design Science Research, a investigação organiza-se em três dimensões complementares: analítica, preditiva e prescritiva. A primeira identifica lacunas metodológicas e geográficas em 28 estudos científicos, revelando a predominância de trabalhos centrados nos Países Baixos e a ausência de modelos que integrem múltiplos recursos operacionais. A dimensão preditiva aplica técnicas avançadas de machine learning (Random Forest, CNN e RNN) a 89.338 registos operacionais da CP, permitindo prever simultaneamente o número de comboios afetados, total de minutos de atraso e passageiros impactados, alcançando ganhos significativos de precisão.

A terceira dimensão introduz uma arquitetura Transformer adaptada ao domínio ferroviário, utilizando 371.668 registos. A inovação principal reside na tokenização semântica que preserva a estrutura hierárquica das tarefas operacionais, permitindo prever sequências completas de tarefas de material circulante após eventos disruptivos. Este modelo superou as arquiteturas LSTM tradicionais, reduzindo o erro de validação em 75% e demonstrando desempenho computacional quatro vezes superior.

Os resultados evidenciam que abordagens baseadas em deep learning são mais eficazes e escaláveis do que métodos determinísticos clássicos. As aplicações práticas incluem sistemas de apoio à decisão em tempo real capazes de antecipar perturbações, otimizar material circulante e melhorar a pontualidade. A tese estabelece novas bases metodológicas para integração de IA na gestão ferroviária, reforçando a resiliência e eficiência do transporte.

Palavras-chave: Gestão de perturbações ferroviárias, Aprendizagem automática, Arquitetura Transformer, Modelos preditivos, Alocação de material circulante

Abstract

This thesis investigates how artificial intelligence can transform the management of operational disruptions in passenger railway systems, focusing on the Portuguese context of Comboios de Portugal (CP). The study begins by noting that literature is fragmented and overly dependent on linear optimization methods, which are poorly scalable and difficult to adapt to real scenarios, and that machine learning approaches remain underused.

Developed under the Design Science Research paradigm, the investigation is organized into three complementary dimensions: analytic, predictive, and prescriptive. The first identifies methodological and geographical gaps in 28 scientific studies, revealing a predominance of work centered in the Netherlands and the absence of models that integrate multiple operational resources. The predictive dimension applies advanced machine learning techniques (Random Forest, CNN, and RNN) to 89,338 operational records from CP, enabling the simultaneous prediction of the number of affected trains, total minutes of delay, and impacted passengers, achieving significant accuracy gains.

The third dimension introduces a Transformer architecture adapted to the railway domain, using 371,668 records. The main innovation lies in semantic tokenization that preserves the hierarchical structure of operational tasks, enabling the prediction of complete sequences of rolling-stock tasks after disruptive events. This model outperformed traditional LSTM architecture, reducing validation errors by 75% and demonstrating computational performance four times higher.

The results show that deep learning approaches are more effective and scalable than classical deterministic methods. Practical applications include real-time decision-support systems capable of anticipating disruptions, optimizing rolling stock, and improving punctuality. The thesis establishes new methodological foundations for integrating AI into railway management, strengthening the resilience and efficiency of transport.

Keywords: Railway disruption management, Machine learning, Transformer architecture, Predictive models, Rolling stock scheduling

Contents

List of Figures

List of Tables

List of Acronyms

AdamW	Adaptive Moment Estimation Weight Decay
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Networks
COVID	Coronavirus Disease
CP	Comboios de Portugal
CRISP-DM	Cross Industry Standard Process for Data Mining
DSR	Design Science Research
EEC	European Economic Community
EU	European Union
GPU	Graphics Processing Unit
ID	Identifier
IoT	Internet of things
IQR	Interquartile Range
KL	Kullback–Leibler
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MILP	Mixed Integer Linear Programming
MLP	Multilayer Perceptron
MSE	Mean Squared Error
OHE	One-hot encoding
PSO	Public Service Obligation
RF	Random Forest
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SNCF	Société Nationale des Chemins de fer Français
SVR	Support Vector Regression
XGBoost	Extreme Gradient Boosting

Chapter 1: Introduction

This chapter outlines the research framework, the motivation behind the investigation, the objectives, the scientific and practical relevance, and the structure of this document.

1.1 Framework and Motivation

Railway transportation constitutes a fundamental pillar of the economic and social development of nations. It plays a crucial role in promoting sustainable mobility and territorial cohesion (Kroon et al., 2015). In the European Union, the progressive liberalization of the railway market began with Directive 91/440/EEC and was later consolidated through Directives 2012/34/EU and 2016/2370/EU. These reforms profoundly transformed the sector's operational paradigm by introducing competition mechanisms that demand increasing levels of efficiency and operational resilience (Schipper & Gerrits, 2018).

The management of passenger railway systems is hierarchically structured into three distinct levels, as illustrated in Figure 1. The strategic level encompasses network and line planning, where operators determine which infrastructure to use and define the core service structure. The tactical level involves timetable generation, rolling stock allocation, and crew scheduling, establishing medium-term operational plans. Finally, the operational level focuses on real-time management, addressing disruptive events that may compromise established plans. It is within this operational level that the present research is positioned.

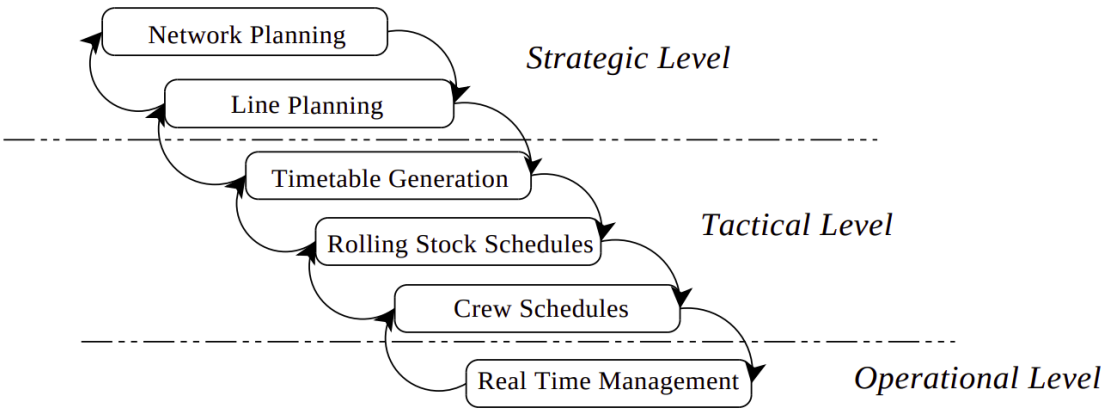

Figure 1: Hierarchical levels of railway management adapted from Yue et al. (2017)

This research specifically focuses on the operational level, concentrating on the real-time management of disruptive events. As demonstrated by Kohl et al. (2007) and illustrated in Figure 2, the operational process for managing disruptions follows a systematic cycle. The cycle begins with continuous monitoring of operations (Monitor operations). When a significant deviation is detected, the system identifies whether a conflict exists (Conflict?). The system then proceeds to identify possible options (Identify possible options), which are evaluated (Evaluate options)

according to operational criteria and passenger impact. This evaluation culminates in a decision (Make decision) that is subsequently implemented (Implement decision). The cycle closes by returning to monitoring, thus creating a continuous process of operational adjustment.

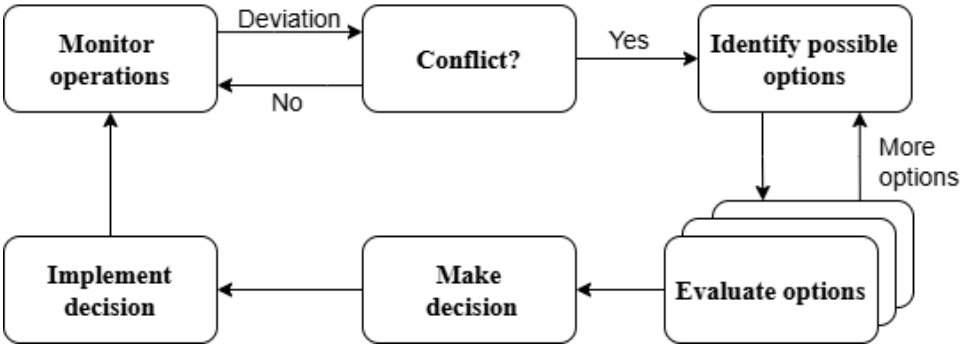

Figure 2: High-level view of the operational process based on Kohl et al. (2007)

In the Portuguese context, Comboios de Portugal (CP) is the national public railway operator of reference. CP faces the complex daily challenge of managing disruptive events that compromise service regularity and affect hundreds of thousands of passengers each year. Figure 3 systematizes the main operational challenges, their direct impact on the system, and the resulting consequences for both the operator and passengers.

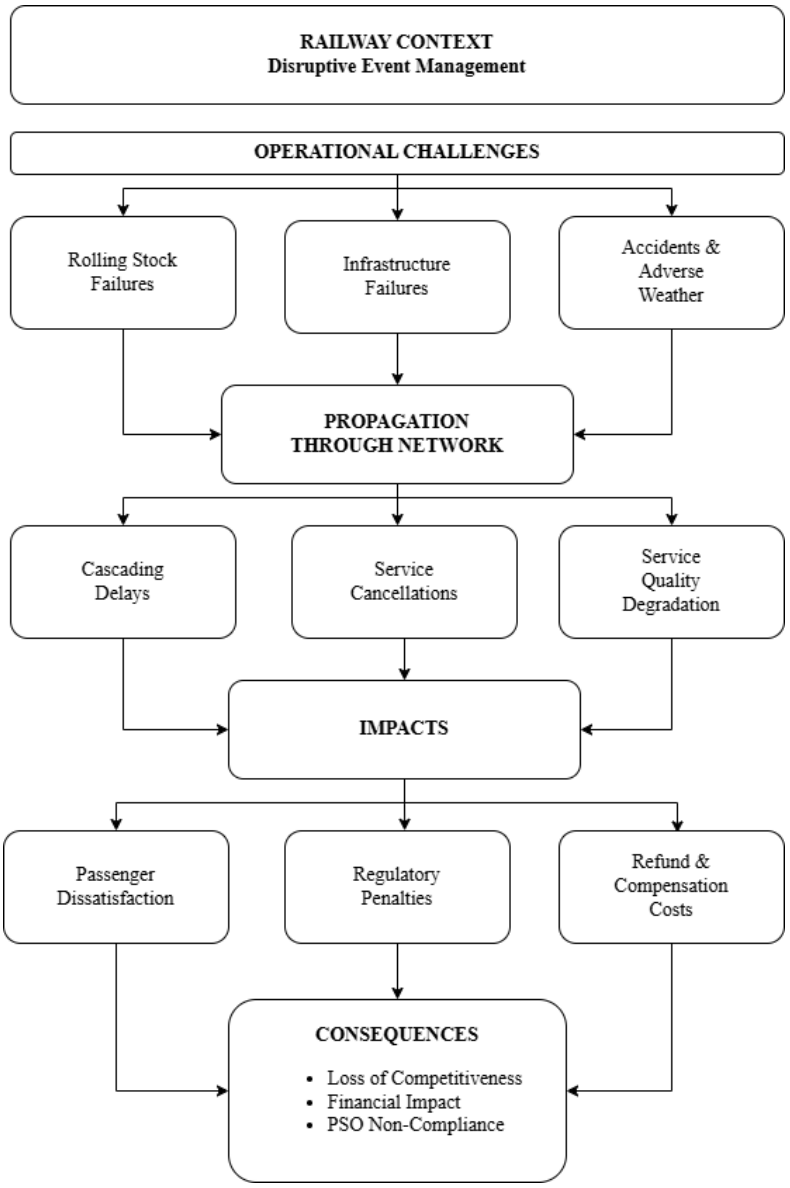

Figure 3: High-level schematic of disruptive event management in the railway context

This doctoral research has generated two peer-reviewed journal articles, both indexed in Scopus and Web of Science. The first, titled "Literature Review on Problem Models and Solution Approaches for Managing Real-Time Passenger Train Operations: The Perspective of Train Operating Companies," appeared in Transportation Research Record, a Q2 journal in Transportation. The second article, "Data-driven Insights to Reduce Uncertainty from Disruptive Events in Passenger Railways," was published in Public Transport, a Q2 journal in Transportation. These publications demonstrate the scientific rigor and international relevance of the research, contributing to the body of knowledge in railway operations and transportation systems.

1.2 Research Problem

Disruption management in passenger railway systems constitutes one of the most complex and critical operational challenges faced by contemporary operators (Kroon et al., 2015). Previous

studies have identified significant limitations in existing approaches to address this problem. Dollevoet et al. (2017) observed that the literature focuses predominantly on isolated rescheduling of individual resources, neglecting the inherently holistic nature of the problem. This observation was reinforced by Corman & Meng (2015), who demonstrated that few approaches can effectively integrate multiple problems, such as simultaneous delay management, rolling stock rescheduling, and crew reorganization.

The fragmentation of proposed solutions contrasts sharply with operational reality, where decisions about different resources are highly interdependent (Nielsen et al., 2012). A change in the rolling stock plan necessarily implies changes in the crew plan, creating cascading effects that propagate throughout the railway network. This systemic interdependence suggests a critical need for more integrated and adaptive approaches. However, the literature reveals an excessive concentration on linear programming methods, despite the known scalability limitations of these approaches when applied to large-scale problems (Louwerse & Huisman, 2014). The underutilization of emerging artificial intelligence (AI) technologies is particularly concerning. Systematic analysis of the literature revealed that few studies explore machine learning techniques, contrasting markedly with the significant advances of these technologies in similar domains of complex systems management. This technological gap is exacerbated by the extreme geographical concentration of research. Most studies focus exclusively on the operational context of the Netherlands, which severely limits the validation of solutions across different operational environments.

Furthermore, the management of these events depends predominantly on the accumulated experience of dispatchers and operational protocols that, although effective in routine situations, demonstrate limitations when facing complex disruptions. The absence of decision support systems based on historical data and capable of learning complex patterns of disruption propagation constitutes a critical operational vulnerability that compromises system efficiency and resilience (Shires et al., 2019).

This study proposes to address these fundamental gaps through systematic research that combines critical literature analysis, development of predictive models based on machine learning, and implementation of advanced deep learning architectures, contributing to a more integrated and technologically advanced approach to railway disruption management.

1.3 Research Objectives

This research proposes an innovative and integrated approach to managing disruptive events in passenger railway transportation, specifically applied to CP's operational context. The main

objectives are structured across three complementary dimensions: Analytical, Predictive, and Prescriptive, as detailed in Table 1.

The Analytical Dimension (Objective 1) involves conducting a systematic and critical review of existing literature, identifying methodological gaps and opportunities for innovation in the domain of real-time railway operations management, with particular emphasis on the perspective of passenger transport operators.

The Predictive Dimension (Objective 2) focuses on developing predictive models based on advanced machine learning techniques capable of accurately anticipating the multidimensional impact of disruptive events. These impacts include the number of affected trains, total delay minutes, and the number of impacted passengers. The models utilize real operational data from CP spanning 2015 to 2024.

The Prescriptive Dimension (Objective 3) addresses the implementation of state-of-the-art deep learning architectures, specifically Transformer models, to predict complete sequences of rolling stock operational tasks following disruptive events. This provides actionable recommendations for rapid and efficient operational replanning.

Table 1 presents the detailed structure of these three objectives, outlining their focus, key research questions, methodological approaches, expected results, and the corresponding chapters where each is developed.

Table 1: Structure of Research Objectives

Dimension	Focus	Key Question	Methods	Expected Results	Chapter
ANALYTICAL (Objective 1)	Identify methodological gaps	What is the state of the art and what are the existing gaps in railway disruption management?	Systematic literature review; Critical analysis of 28 scientific articles; Railway transport operators' perspective	Identification of methodological and technological gaps; Mapping of geographical trends; Theoretical foundation for artifact development.	Chapter 3: Review and Critical Analysis to Support the Artifacts
PREDICTIVE (Objective 2)	Anticipate impacts	What will be the multidimensional impact of the disruption?	Machine learning models (Random Forest, CNN, RNN, MLP); CRISP-DM methodology; 89,338 operational records (CP 2015-2022); Topological variables (betweenness centrality)	Simultaneous estimates of: number of affected trains, total delay minutes, number of impacted passengers; Integrated multidimensional approach; Comparison among ML architectures.	Chapter 5: Development and Experimental Evaluation of Predictive Models
PRESCRIPTIVE (Objective 3)	Suggest actions	What should we do to address this impact?	Deep learning with Transformer architecture; Encoder-decoder model; Specialized semantic tokenization; 371,668 operational records (2016-2024); Attention mechanisms	Complete sequences of operational tasks; Recommendations for: schedule rescheduling, rolling stock redistribution, operational recovery strategies; Statistical plausibility evaluation.	Chapter 6: Prediction of Rolling Stock Tasks with Transformer

1.4 Research Contribution

This research provides a coherent response to the methodological and technological gaps identified in the state of the art, addressing the fragmentation, scalability, and limited use of artificial intelligence that constrain current approaches to railway disruption management.

(1) Systematic literature review for methodological integration (first published article). Current research remains fragmented, treating timetable, rolling stock, and crew as isolated problems. The first study, corresponding to the first published article, systematically reviewed 28 peer-reviewed papers from the past decade, delivering the first operator-centered synthesis of international work. It reveals how the lack of cross-resource integration and geographical bias limits generalization. The main contribution is a unified analytical framework that consolidates dispersed findings into actionable guidance for integrated, data-driven decision support systems.

(2) Multidimensional predictive modeling to close the performance gap (second published article). Existing models optimize single objectives and rely on limited simulations. The second study, corresponding to the second published article, develops predictive models that jointly estimate affected trains, total delay, and impacted passengers, marking the first multidimensional framework in passenger railway operations. Using real CP data (2015–2022) and incorporating topological variables such as betweenness centrality, the models show how machine learning can capture complex operational interactions. The key contribution is a validated predictive tool that provides operators with multidimensional foresight for faster and more balanced disruption recovery.

(3) Transformer-based prescriptive modeling to achieve scalability. Traditional optimization methods fail at large, real-time scales due to their inability to learn sequential dependencies. The third study introduces a Transformer architecture tailored to railway operations, with a semantic tokenization scheme that preserves operational hierarchies. By capturing long-range temporal and spatial dependencies, it generates complete and coherent rolling stock task sequences. This contribution establishes a scalable, data-driven alternative to deterministic optimization, reducing validation error by 75% compared with Long Short-Term Memory (LSTM) models and setting a new methodological benchmark for prescriptive decision support.

1.5 Thesis Structure

Figure 4 represents the methodological architecture of this investigation, showing a structural organization based on rigorous scientific progression that evolves from theoretical conceptualization to the development of technological artifacts. This sequential structure reflects the implementation of the Design Science Research (DSR) paradigm, which is executed across seven interconnected chapters that ensure methodological coherence and the incremental construction of scientific knowledge.

Chapter 1 establishes the epistemological framework of the investigation by defining the central problem of disruptive event management in railway systems and contextualizing the scientific and practical relevance of the study.

Chapter 2 develops the theoretical foundation by conducting an extensive state-of-the-art review that documents the methodological evolution from traditional mathematical optimization approaches to emerging artificial intelligence applications in the railway sector.

Chapter 3 implements a systematic critical analysis of the literature through a structured review process that identifies and classifies relevant studies, highlighting methodological gaps and research opportunities.

Chapter 4 presents the global research methodology that structures the entire investigation. Grounded in the DSR paradigm, it defines the iterative cycles of problem identification, artifact design, demonstration, and evaluation that guide the subsequent chapters.

Chapters 5 and 6 describe the development and validation of the technological artifacts. Chapter 5 focuses on the creation of multidimensional predictive models capable of simultaneously forecasting different impact metrics of disruptive events. This work applies to the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology and implements multiple machine learning architectures. Chapter 6 constitutes the technological core of the investigation through the design of a Transformer architecture specifically adapted to the railway domain, incorporating specialized semantic tokenization strategies to preserve the structural integrity of operational data.

Chapter 7 concludes the thesis by systematizing the main scientific and technological contributions, evaluating the fulfillment of the established objectives, and proposing strategic directions for future research.

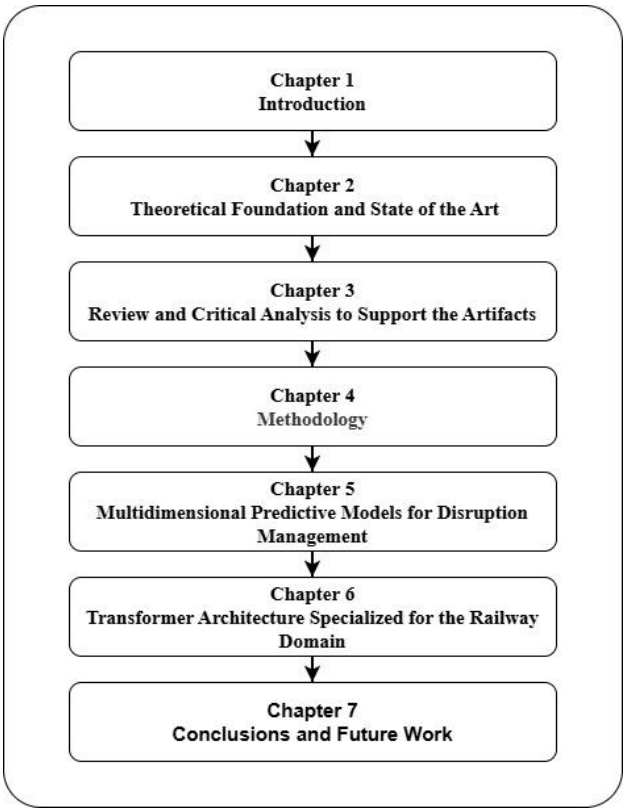

Figure 4: Structure and organization of the thesis chapters

This sequential organization ensures that each structural component contributes to validating the proposed artifacts, demonstrating how the convergence between railway domain knowledge and advanced artificial intelligence technologies can generate innovative solutions for operational management. Thus, the first chapter not only frames the motivation and objectives of the investigation but also paves the way for Chapter 2, which then deepens the theoretical background and conceptual foundations upon which the entire research is built.

Chapter 2: Theoretical Foundation and State of the Art

Building on the research problem, objectives, and contributions laid out in Chapter 1, this chapter establishes the theoretical foundation that frames railway disruption management as a multidimensional, system-wide challenge. The chapter synthesizes the dominant methodological paradigms, their assumptions, and known limitations, while identifying where emerging AI approaches begin to reshape the field.

2.1 Dominant Methodological Paradigms and Systemic Limitations

The scientific literature on railway disruption management demonstrates a pronounced dependence on mixed-integer linear programming methods, identified in seventeen of the twenty-eight studies reviewed. This predominance reflects the historical preference for methodologies that ensure optimal solutions for specific problem formulations, while also exposing critical limitations in scalability and adaptability.

Dollevoet et al. (2017) observed that most research focuses on the isolated rescheduling of individual resources, neglecting the holistic nature of railway disruption management problems. Their study, based on an iterative framework combining multiple heuristics, demonstrated the ability to obtain feasible solutions in under five minutes for networks with up to five hundred trains. However, it also revealed substantial limitations when applied to integrated multi-resource problems. Cacchiani et al. (2014) corroborated these findings in their comprehensive review of recovery models for real-time railway rescheduling, identifying that linear programming methods encounter significant computational constraints in large-scale scenarios.

Scalability remains a major challenge in large-scale railway problems. Fang et al. (2015) reported that only a limited number of studies address large-scale settings, noting that optimal or near-optimal solutions can be achieved within reasonable computational time only when the railway network is relatively small. König and Schön (2020) reinforced this view, arguing that incorporating stochastic aspects into delay management for large space–time networks pose a computationally prohibitive challenge.

Corman et al. (2012) established a seminal reference in the domain through a bi-objective heuristic solution for conflict detection and resolution in railway traffic management. However, Corman and Meng (2015) observed in their review of online dynamic models that few approaches could effectively integrate multiple simultaneous

problems, underlining the persistence of methodological fragmentation even in more recent studies.

2.2 Geographical and Contextual Fragmentation

Extreme geographical concentration is a critical characteristic of the current state of the art, with nineteen of the twenty-eight articles focusing exclusively on the Netherlands, representing 67.9% of the analyzed literature. Kroon et al. (2015) exemplify this trend through their study on rolling stock rescheduling with dynamic passenger flows, based exclusively on data from Nederlandse Spoorwegen, the main railway company in the Netherlands. This concentration, although reflecting the maturity of this country's railway system and substantial research investment in this domain, severely limits the applicability of developed solutions to other contexts with distinct structural characteristics. Schipper and Gerrits (2018) analyzed the differences and similarities in European disruption management practices, identifying substantial variations between countries that fundamentally question the generalization of solutions developed for specific contexts.

The Netherlands railway system presents characteristics that may not be representative of other operational environments: a relatively compact network of 3,223 kilometers, high service density, modern infrastructure, and straightforward topology. The Portuguese network, in contrast, presents distinct operational challenges characterized by elongated north-south topology, significant presence of single-track sections, and pronounced variations in traffic density between urban and rural regions. These fundamental structural differences suggest that solutions optimized for the Dutch context may demonstrate limited effectiveness when applied to other operational environments with divergent topographical, demographic, and infrastructural characteristics.

2.3 Evolution of Artificial Intelligence Approaches

The application of artificial intelligence techniques represents a significant gap identified in literature, with only 3.6% of studies exploring these technologies. This limited adoption stands in stark contrast to the advances demonstrated in related domains and represents a critical opportunity for methodological innovation. Marković et al. (2015) pioneered the use of Support Vector Regression (SVR) to analyze passenger train delays in the Serbian network, achieving accuracies exceeding 85% for 30-minute prediction horizons. Their work established a methodological precedent for applying machine learning in the railway domain.

Li et al. (2021) extended this research by applying the Random Forest (RF) algorithm to data from the Netherlands railway network for short-term delay prediction. Their analysis incorporated historical rail operation data, including scheduled and actual times, crew changes, rolling stock circulation, and weather conditions. The RF model was optimized and compared with other algorithms such as Artificial Neural

Network (ANN), XGBoost, and gradient boosting decision trees. Results showed high accuracy and demonstrated the ability of these techniques to handle heterogeneous variables, while acknowledging limitations arising from dependence on specific historical data. Nabian et al. (2019) applied a bi-level RF approach to ten million data points from the Netherlands network over thirteen weeks, demonstrating the scalability of these methods for massive volumes of operational data, though without achieving the contextual generalization needed for universal application.

Grandhi et al. (2021) developed an estimation framework to quantify railway disruption parameters using neural networks applied to the Danish network, identifying meteorological variables as critical factors while facing challenges related to the accuracy of manually entered data. Huang et al. (2020) introduced a Bayesian network model to predict disruption effects on Chinese railway operations, demonstrating high accuracy in delay propagation prediction despite facing limitations due to algorithmic complexity and the need for extensive specialized knowledge.

Chen et al. (2022) conducted an analysis of unplanned disruptions in Hong Kong urban railway systems using quantile regression models. Their findings showed that temporal factors and meteorological conditions exert significant influence, although they acknowledged limitations in the availability of information about disruption procedures and infrastructure data. Klumpenhouwer and Shalaby (2022) applied machine learning techniques, including Random Forest and Elastic Net, to support passenger railway operations on the GO Rail network in Ontario. Their approach proved effective in identifying signal failures and incidents as influential attributes, though with recognized limitations in modeling delay propagation through the network.

Wang and Zhang (2019) developed a gradient boosting tree model for predicting delay times in Chinese passenger trains, integrating meteorological factors, traffic volume per station, and operational history. Their results demonstrated that traffic volume characteristics and train conditions exert a substantial impact on delays, and that these effects propagate systematically to subsequent trains.

2.4 Advanced Methodological Approaches and Emerging Paradigms

Recent literature demonstrates a gradual emergence of hybrid methodologies aiming to overcome the limitations of traditional paradigms. Veelenturf et al. (2017) presented a passenger-oriented rescheduling strategy based on heuristics applied to operational data from the Société Nationale des Chemins de fer Français (SNCF), showing effectiveness in structured scenarios but revealing scalability and generalization issues when

faced with high operational variability. Zhou et al. (2022) introduced a novel framework for the joint optimization of train timetabling and rolling stock circulation planning with flexible train composition modes, validated in operational contexts but requiring extensive parameter calibration and displaying high sensitivity to input data quality.

Zhong et al. (2019) implemented a hierarchical rolling stock scheduling system incorporating maintenance requirements for the Chinese High-Speed Railway, utilizing a progressive coarse-to-fine segmentation approach. While this demonstrated computational efficiency in structured planning tasks, it exhibited high computational costs and limited flexibility under unforeseen disruption scenarios. Yin et al. (2023) proposed an integrated architecture based on Bender's decomposition for the joint optimization of rolling stock allocation and timetables in urban transit networks, offering computational efficiency but demanding intensive calibration and showing sensitivity to variations in operational data quality.

Nishi et al. (2017) developed an innovative combination of column generation and heuristic methods that integrate historical data with mixed-integer linear programming models for short-term rolling stock planning optimization, incorporating regular inspection constraints. Liu et al. (2024) combined business rules with real-time adaptive scheduling adjustments for coupling–decoupling operations, achieving robust operational outcomes while still facing adaptability challenges in unforeseen operational scenarios.

2.5 Emergence of Advanced Machine Learning Techniques

Recent literature reveals a gradual emergence of hybrid methodologies designed to overcome the limitations of traditional paradigms. Veelenturf et al. (2017) presented a passenger-oriented rescheduling strategy using heuristics applied to operational data from SNCF. Their approach proved effective in structured scenarios, though it exhibited scalability and generalization limitations when faced with high operational variability. Zhou et al. (2017) introduced multi-population genetic algorithms for robust rolling stock planning under uncertainty. The validation was conducted exclusively in simulated environments, and the method required extensive parameter calibration, exhibiting high sensitivity to input data quality.

Borndörfer et al. (2017) implemented a rolling hierarchical planning system for Deutsche Bahn, Germany's primary railway operator, employing a coarse-to-fine progressive segmentation approach. While the system demonstrated efficiency in structured planning contexts, it faced substantial computational demands and operational rigidity during unforeseen disruption scenarios. Yin et al. (2023) proposed an integrated architecture based on Bender's decomposition for the joint optimization of rolling stock allocation and scheduling in urban transit networks. Although the approach provided

computational advantages, it required intensive calibration and remained sensitive to variations in operational data quality.

Nishi et al. (2017) developed an innovative combination of column generation and heuristic techniques that integrate historical data with mixed-integer linear programming models for short-term rolling stock planning optimization, incorporating regular inspection constraints.

2.6 Critical Analysis of Systemic Limitations

A critical analysis of the literature reveals four fundamental categories of limitations that simultaneously act as obstacles to scientific progress and as structural opportunities for innovative research. Problem fragmentation manifests through a dominant tendency toward decomposing issues into isolated subproblems: approximately 75% of studies focus exclusively on specific components without accounting for systemic interdependencies. This approach overlooks the operational reality identified by Lusby et al. (2017), in which decisions regarding different railway resources are highly interdependent and generate cascading effects that propagate throughout the network.

Nielsen et al. (2012) provides a notable methodological exception through their rolling-horizon approach to rolling stock disruption management, employing operational data and demonstrating superior performance compared to methods that do not incorporate historical information. This empirical evidence underscores the untapped potential of operational data to capture complex disruption propagation patterns, sharply contrasting with the predominance of studies based on simulated scenarios or isolated case studies.

Fang et al. (2015) conducted a comprehensive survey of problem models and solution approaches for rescheduling in railway networks, confirming methodological fragmentation and revealing that most solutions focus on specific subproblems without considering systemic interactions. Schipper et al. (2015) supported this perspective through a dynamic network analysis of information flows in railway disruption management, exposing the complexity of interdependencies that prevailing fragmented approaches systematically overlook.

The technological gap in artificial intelligence applications is particularly evident: AI accounts for only 3.6% of the analyzed studies, in stark contrast with the advances achieved in comparable domains. Zhan et al. (2024) identified uncertainty management in railway rescheduling as a critical area requiring targeted methodological innovation, confirming the need for more sophisticated approaches. Fabella and Szymczak (2021) examined the vulnerability of the German railway network to natural disasters using negative binomial regression models, demonstrating that floods significantly reduce the

number of operating trains and emphasizing the need for more comprehensive datasets to analyze concurrent events. Golightly and Dadashi (2017), through interviews with British railway personnel, identified 26 relevant attributes for transport interruptions; however, the translation of this tacit knowledge into automated systems remains unexplored.

2.7 Theoretical Convergence and Innovation Opportunities

A critical synthesis of literature reveals three fundamental tensions that define the boundaries of current knowledge and guide opportunities for future research. The tension between methodological fragmentation and the need for holistic integration is evident in the predominance of studies that treat operationally interdependent components in isolation, thereby ignoring the systemic effects identified by König (2020) in his review on railway delay management. König argues that holistic models are conceptually superior because they simultaneously satisfy passenger and operator needs, yet he identifies a critical scarcity of stochastic models that adequately consider operational uncertainty.

The tension between traditional optimization methods and emerging artificial intelligence technologies becomes evident in the resistance to adopting data-driven adaptive paradigms. This gap is particularly significant given that comparable domains have already demonstrated substantial benefits with machine learning and deep learning approaches. The research by Huang et al. (2020), who developed models based on Bayesian networks for predicting disruption effects, achieved high accuracy but faced limitations due to algorithmic complexity and the need for extensive specialized knowledge.

2.8 Prospective Synthesis and Research Directions

The convergence of identified limitations provides theoretical support for a fundamentally different approach that moves beyond traditional dichotomies, including exact versus heuristic optimization, fragmented versus integrated methodologies, and deterministic versus adaptive approaches. The shift from conventional optimization paradigms to adaptive methodologies based on machine learning represents more than a mere technical evolution; it signals a fundamental conceptual change in how we understand railway systems as complex, dynamic, and interdependent entities.

Recent empirical evidence supports this shift. The work of Gao et al. (2023), which achieved coefficients of determination of 0.87 in large-scale datasets, demonstrates that careful application of artificial intelligence techniques can substantially overcome the limitations of conventional methods. Nevertheless, the literature also highlights the need for specific methodological development tailored to the railway domain, one that

accounts for the operational, regulatory, and contextual particularities distinguishing this sector from other application areas.

Future research should integrate methodological rigor with practical applicability, developing approaches that leverage emerging artificial intelligence technologies to create more adaptive, integrated, and effective decision support systems. Such research can establish new scientific frontiers and enable effective translation of academic advances into tangible improvements in railway operations management and passenger experience.

The theoretical insights and documented gaps identified here motivate a rigorous, evidence-based inquiry in Chapter 3, where a systematic literature review and critical analysis are conducted to operationalize and quantify the trends, limitations, and opportunities mapped in this chapter.

Chapter 3: Review and Critical Analysis for Artifact Support

While Chapter 2 provided a conceptual synthesis of paradigms, assumptions, and systemic limitations, this chapter translates those insights into a structured, systematic literature review and critical analysis. The chapter applies explicit search, selection, and appraisal criteria to identify trends, quantify fragmentation, and identify gaps that directly inform artifact design.

3.1 Introduction

Based on the theoretical framework established in Chapter 2 and following the DSR methodology defined in Chapter 4, this chapter operationalizes the first phase of the research through a systematic literature review. The specific objectives identified in Section 1.3, particularly the need to identify methodological gaps and innovation opportunities, underpin the analytical strategy that follows.

The systematic literature review conducted in the first scientific article adopted a specific and differentiated methodological perspective, focusing exclusively on the viewpoint of railway passenger transport operators. This conceptual delimitation addresses a critical gap in the existing literature, where most studies emphasize the perspective of railway infrastructure managers.

The study established four specific objectives that systematically guided the investigation and analysis of the literature. The first objective was to present a literature review on disruptive events, specifically from the perspective of railway transport operators. This approach acknowledges that train operators have distinct motivations and interests compared to infrastructure managers, as they oversee different operational variables such as crew scheduling and rolling stock allocation.

The second objective focused on identifying the variables and problem classifications considered in the analyzed studies. This analytical dimension aimed to establish a comprehensive taxonomy of existing methodological approaches and to determine the critical factors influencing railway disruption management.

The third objective consisted of identifying and synthesizing the main ideas proposed in literature. This systematic examination of existing contributions made it possible to understand dominant methodological trends and the conceptual gaps that define the state of the art.

The fourth objective sought to identify a set of significant variables explicitly from the perspective of passenger railway operators. This prospective dimension laid the foundation for the subsequent development of methodological solutions that directly

address the operational needs of railway operators, thereby distinguishing this work from conventional approaches centered on infrastructure management.

3.2 Methods

This section outlines the methods employed in the systematic literature review, including the inclusion and exclusion criteria, the databases consulted, and the search strategies applied, as well as the procedures for screening and assessing the quality of the selected studies.

3.2.1 Search Strategy

The systematic review followed a structured protocol in four phases, described in the subsequent subsections. For literature collection, we consulted the Scopus and Web of Science databases, selected for their comprehensiveness and quality in indexing scientific publications in this field.

The initial search employed specific railway terminology based on the taxonomy of disruptive events established by the European Railway Agency (document ERA-PRG-004-TD-002). The search terms included combinations of keywords such as “railway disruption,” “train delay management,” “real-time rescheduling,” “rolling stock recovery,” and “crew rescheduling,” all with a focus on the railway operator perspective.

3.2.2 Selection Criteria

The inclusion criteria were carefully defined to ensure the relevance and quality of the analyzed studies. Articles published in peer-reviewed scientific journals or indexed conference proceedings were considered eligible for analysis. Additionally, studies specifically focused on real-time railway passenger operations management were included. Publications presenting models or algorithms for resolving disruption problems were also incorporated, with preference given to articles adopting a railway operator perspective.

For the exclusion criteria, studies focused exclusively on railway infrastructure management were eliminated from the analysis, along with articles dedicated to railway freight transport. Publications lacking a clear methodological component were excluded, as were purely theoretical studies without demonstrable operational applicability. This methodological approach ensured that the systematic literature review considered only works directly related to the problem under investigation and with practical relevance to the sector.

3.2.3 Selection Process and Data Extraction

The initial search identified 131 candidate articles through the structured search strategy. After rigorous application of exclusion criteria and comprehensive content analysis, 28 articles were

selected for detailed analysis, representing a selection rate of 21.4% that ensures high scientific quality. The selection process was operationalized through four sequential methodological phases:

1. Identification: Systematic execution of the search strategy in bibliographic databases, applying predefined search terms and specific Boolean operators, resulting in the identification of 131 articles potentially relevant to the research scope.
2. Coding: Development and validation of a structured coding system for systematic meta-information extraction, covering critical analytical dimensions such as central theoretical contributions, adopted methodological paradigms, implemented modeling techniques, considered study variables, academic profile of researchers, and potential for extrapolation to future developments.
3. Extraction: Operationalization of systematic and structured collection of previously categorized meta-information, using Mendeley software for integrated bibliographic management and NVivo for computer-assisted qualitative analysis, ensuring methodological rigor in the systematization of extracted data.
4. Synthesis: Analytical compilation and structured presentation of results through interpretive formats and graphical representations, facilitating the identification of emerging patterns, methodological trends, and knowledge gaps in the investigated scientific domain.

3.2.4 Quality Analysis

The methodological quality assessment of studies included in this systematic review was conducted by analyzing the quartile classification of publications. This approach assumes that the hierarchical position of journals in bibliometric indices serves as a reliable indicator of scientific excellence and rigorous peer review.

The quartile distribution of studies analyzed shows a significant concentration in high-impact publications. Specifically, the analysis reveals that most studies were published in first-quartile journals, with five articles indexed in Web of Science and eleven in Scopus. This predominance of publications in prestigious journals reflects both the methodological robustness of the research and the thematic relevance and contemporary importance of the topics addressed.

3.3 Results

This section presents the results of the systematic literature review, organized to highlight the main trends, methodological approaches, and limitations identified in the analyzed studies. The discussion addresses temporal evolution patterns, geographical distribution, research objectives, and the frequency of different methodological approaches.

3.3.1 Temporal Characterization and Publication Impact

The temporal analysis of the 28 scientific publications examined in the systematic review revealed significant patterns of research evolution in railway disruption management from 2010 to 2020.

The most cited article was the heuristic proposal by Corman et al. (2012), which accumulated 135 citations in Scopus and 124 in Web of Science. This work established itself as a seminal contribution for addressing delay problems through two-phase approaches. This fundamental study demonstrated the viability of computationally efficient solutions for real-time disruption management scenarios, significantly influencing the subsequent development of methodologies in the field.

The geographical distribution of scientific contributions revealed a notable concentration, with 68% of studies originating from the Netherlands. This reflects the advanced maturity of the country's railway system. However, this geographical concentration contrasts markedly with the limited representation of other European and worldwide operational contexts. This constitutes a critical methodological gap that compromises the generalizability of proposed solutions to railway systems with distinct topological, operational, and regulatory characteristics.

A critical finding is the substantial under exploration of artificial intelligence techniques, with only 3.6% of studies incorporating machine learning methodologies. This is surprising given the transformative potential demonstrated by these technologies in related operational domains. The predominance of traditional linear programming methods, which represent 60.7% of methodological approaches, suggests either resistance to or lack of awareness of the advanced capabilities offered by machine learning algorithms to capture complex and non-linear patterns characteristic of contemporary railway systems.

3.3.2 Problem Model Classification

Table 2 summarizes the results of the bibliometric analysis. This methodological classification shows the distribution of research approaches in railway disruption management, revealing a clear predominance of traditional mathematical programming methods in contrast to the limited application of emerging artificial intelligence techniques. The table presents four fundamental analytical dimensions: method identification, absolute number of publications, percentage representation in the total sample, and technical characterization of each methodological approach. This framework provides an essential tool for understanding dominant paradigms in scientific literature and for identifying methodological gaps that can guide future research directions in the field of complex railway systems management.

Table 2: Methodological paradigms in railway disruption management research

Method	Publications	Percentage	Description
Mixed Integer Linear Programming (MILP)	10	35.7%	Most popular approach due to flexibility in admitting continuous and discrete decision variables, suitable for modeling the complexity of railway operations where times (continuous) and binary decisions (cancel/not cancel) coexist.
Integer Linear Programming	7	25.0%	Used primarily when all decision variables are discrete, as in determining train service suppressions or binary resource allocation.
Simulation Models	5	17.9%	Allow the creation of digital prototypes of the railway system and testing of hypothetical scenarios, being particularly useful for evaluating the impact of different recovery strategies.
Heuristic Models	4	14.3%	Essential approaches when uncertainty and problem dimension make obtaining optimal solutions in useful time impractical.
Machine Learning	1	3.6%	Machine learning techniques, specifically Support Vector Machines and Kalman filters for delay prediction. This scarcity contrasts with the proliferation of AI in other domains and represents a significant research opportunity.

3.3.3 Decision Variable Analysis

Figure 5 illustrates the 56 variables employed in the proposed solutions for disruptive events. These variables are classified into two decision groups: binary variables, which take Boolean states, and numerical variables, which capture measurable values associated with events.

Figure 5: Map of decision variables

Binary Variables (21 variables): The most frequent binary variable determines whether a train is suppressed, appearing in 18 of the 28 analyzed articles. This predominance reflects how central this decision is in disruption management, often serving as the primary response when maintaining the original schedule becomes impractical (Lusby et al., 2017). Decisions about maintaining connections between services constitute another fundamental category, particularly relevant in systems with a high density of transfers. The use of alternative routes emerges as a third critical dimension, especially in contexts with flexible infrastructure that allows strategic rerouting during disruptive events.

The activation of contingency plans represents a fourth significant category, reflecting the transition from normal operations to exceptional protocols. These protocols may include substantial alterations in circulation patterns or implementation of alternative services. The analysis also reveals variables related to specific resource allocation, including decisions about rolling stock repositioning and crew reassignment. This demonstrates the multi-resource nature of railway management problems (Corman & Meng, 2015).

Numerical Variables (35 variables): Delays dominated this category. They were modeled in various ways, with conceptualizations that distinguish between absolute delays, accumulated delays, delay propagation, and recovery times. The modeling of delay propagation constitutes a particularly sophisticated dimension, seeking to capture how localized disruptions amplify through the railway network due to the interconnected nature of operations (Huang et al., 2020).

Variables related to discrete real-time traffic events, such as departure and arrival schedules at stations, were identified as critical for effective disruption management. The literature shows that the capacity to predict and optimize these events is a determining factor for minimizing the global impact of disruptions. This capability enables strategic synchronizations that reduce waiting times and maximize the utilization of available capacity.

The fragmentation identified in the literature between approaches that treat different aspects of the problem in isolation contrasts with operational reality. In practice, decisions about apparently distinct variables are highly interdependent. An alteration in the rolling stock plan may require coordinated changes in the crew plan, while train suppressions may necessitate simultaneous reconfiguration of connections and activation of contingency plans (Lusby et al., 2017). This systemic interdependence suggests that future investigations should adopt more integrated approaches capable of capturing synergies between different categories of decision variables.

3.3.4 Optimization Objectives

The systematic analysis of models identified nine main optimization objectives that reflect differentiated priorities in railway disruption management, demonstrating a hierarchy of operational concerns that does not always align with predominant theoretical paradigms in literature. These objectives include maximizing the number of journeys, maximizing the use of train crews, maximizing passenger flow, maximizing passenger satisfaction, maximizing the use of rolling stock, and minimizing the number of delayed journeys, alongside various multi-objective combinations (Cacchiani et al., 2014). This distribution of objectives reveals fundamental tensions between internal operational efficiency and service quality as perceived by users, suggesting a potential disconnection between organizational metrics and passenger expectations.

Minimizing the number of delayed trains emerges as the most prevalent objective, identified in 32% of analyzed studies, reflecting the central importance of punctuality as a key performance indicator in railway systems. This predominance demonstrates the persistence of an organizational culture focused on regularity metrics that, although operationally relevant, may not adequately capture the multidimensional complexity of disruption impact on passenger experience. The complementarity of this objective with minimizing total delay minutes, present in 25% of studies, suggests a dual approach that seeks to balance frequency and magnitude of temporal deviations, recognizing that delays can vary substantially in terms of operational impact and user perception.

The under-representation of passenger satisfaction maximization is particularly revealing, as it was identified in only 18% of studies. This is a surprising result considering that the present review specifically adopts the perspective of train operating companies. This apparent neglect of user-centered objectives contrasts markedly with contemporary trends in public transport management, where passenger expectations regarding service reliability and transparency have progressively intensified (Cats et al., 2016; Gkiotsalitis & Cats, 2018). This discrepancy suggests a potential gap between customer-focused organizational rhetoric and the optimization practices implemented, indicating opportunities for more balanced approaches that systematically integrate both operational and user perspectives.

Real-time railway operational management involves balancing multiple objectives that frequently conflict with one another. Railway operators seek to minimize operational costs while maximizing service quality for passengers. This management challenge is complicated by the substantial costs involved and the train operating company's need to identify optimal solutions from a comprehensive operational perspective. Poor management decisions can prove highly detrimental, producing lasting effects on schedule adherence. According to Dollevoet et al. (2017), reliability represents one of the

essential criteria for passenger satisfaction. Transfer connections constitute another critical factor for passengers, particularly for those who must use multiple trains to reach their destinations. In these circumstances, effective operational management becomes essential to ensure passengers reach their final destinations, which may occasionally necessitate arranging alternative transport options.

Objectives related to maximizing the number of circulations and maximizing passenger flow, each present in 7% of studies, represent more sophisticated approaches that seek to optimize the utilization of available capacity under disruption conditions. Maximizing the number of circulations assumes relevance in situations of partial infrastructure blockage, where maintaining the maximum possible number of services can minimize the global impact of disruptions. Complementarily, maximizing passenger flow focuses on preserving critical connections and minimizing impact on individual journeys, recognizing that network connectivity constitutes a determining factor in railway service quality.

The distribution of these objectives reveals conceptual fragmentation that reflects structural limitations in existing optimization approaches, where different studies adopt divergent priorities without systematic consideration of interdependencies between potentially conflicting objectives (Veelenturf et al., 2017). This observation suggests the need for methodological developments that transcend single-objective optimizations by adopting multi-objective frameworks capable of capturing complex trade-offs between operational efficiency, service quality, and economic sustainability. Integrating these diverse objectives into a holistic optimization paradigm constitutes a fundamental methodological challenge that future investigations should address to overcome the fragmented limitations identified in the current state of the art.

3.3.5 Disruption Management Phases

The application of the conceptual “bathtub” model proposed by Ghaemi et al. (2018) to the analysis of identified studies reveals a highly asymmetric distribution of investigative focus across the three fundamental phases of railway disruption management. This theoretical model conceptualizes the management of disruptive events as a tripartite process comprising the transition to an alternative plan, operation under disruption, and return to normality, as illustrated in Figure 6.

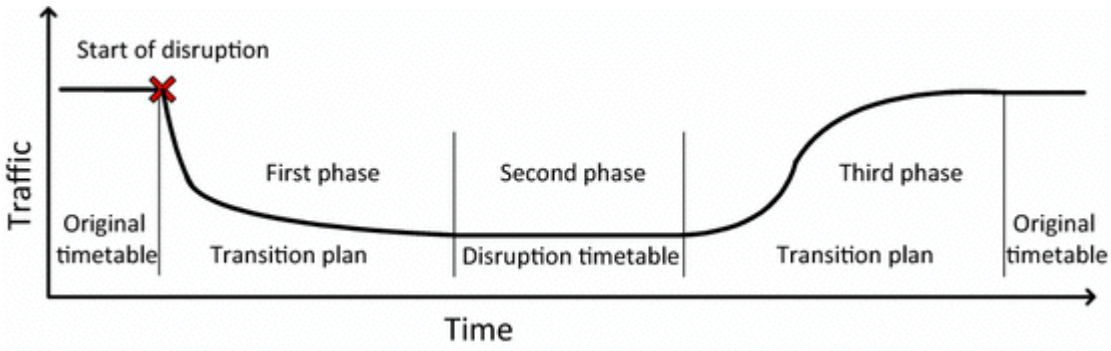

Figure 6: Bathtub model illustrating the traffic levels during a disruption (Ghaemi et al., 2018)

The disproportionate concentration of studies in the second operational phase constitutes a particularly revealing finding. Twenty-three articles, representing 82.1% of the analyzed corpus, focus exclusively on management during active disruption. This preponderance reflects an investigative tendency that privileges tactical resource optimization in already established disruption scenarios while neglecting the strategic dimensions inherent in transition and recovery phases. The predominance of this approach suggests a reductionist conceptualization of disruption management as a point-in-time optimization problem, ignoring the dynamic and evolutionary nature of disruptive events in complex railway systems.

The analysis of the first phase, addressed by only eight articles representing 28.6% of studies, demonstrates a critical gap in understanding initial transition processes. This phase holds fundamental importance in determining the overall success of disruption response by establishing initial conditions that shape all subsequent decisions. The underrepresentation of this dimension in literature potentially reflects methodological limitations of traditional optimization paradigms, which tend to favor pre-established scenarios at the expense of modeling decision-making processes under conditions of high uncertainty (Cacchiani et al., 2014).

Even more pronounced is the neglect of the third phase of return to normalcy, addressed in only three articles corresponding to 10.7% of the sample. This gap reveals an investigative perspective that treats disruptions as isolated events without considering systemic implications for long-term operational resilience. Effective management of the return to normalcy is a determining factor for minimizing residual effects and preparing the system for future disruptions, both critical dimensions frequently ignored by short-term optimization paradigms (Nielsen et al., 2012).

The identification of only one article representing 3.6% of the total that presents a truly integrated approach to the three phases constitutes perhaps the most significant finding of this analysis. This scarcity of holistic perspectives reveals fundamental

methodological fragmentation that contradicts the systemic nature of railway disruptive events. The interdependencies between phases are particularly pronounced in railway systems. Decisions made during the initial transition can have ramifications that extend far beyond immediate disruption resolution, influencing operational patterns and the capacity to respond to future disruptions (Huang et al., 2020).

This asymmetric distribution reveals profound structural limitations in predominant methodological approaches, suggesting that the reliance on integer and mixed integer linear programming methods may be partially responsible for the identified fragmentation. These paradigms, although effective for optimization under well-defined conditions, demonstrate inherent limitations in capturing the complex temporal dynamics and systemic interdependencies that characterize complete cycles of disruption and recovery (Fang et al., 2015). The need for more integrated approaches becomes evident when considering that the overall effectiveness of disruption management depends fundamentally on coherent coordination across all phases of the process. This requires methodologies capable of capturing synergies and trade-offs between temporally distributed but operationally interdependent decisions.

The findings of this review, particularly the documented fragmentation across phases, the scalability limits of optimization methods, and the underuse of machine learning, establish concrete design requirements that are formalized in Chapter 4's methodology and research framework.

Chapter 4: Methodology

In Chapter 2, we established the theoretical landscape and its limitations. In Chapter 3, we systematically validated those insights through a structured review and critical analysis. Building on these foundations, this chapter formalizes the research methodology, which is grounded in the Design Science Research paradigm. This approach transforms the identified gaps into actionable design cycles, evaluation criteria, and validation procedures for the proposed artifacts.

4.1 Investigation Framework through the Design Science Research Paradigm

The choice of DSR is justified by the multidisciplinary nature of the problem, which requires integrating knowledge from computer science, operational research, and railway management. This methodology provides a systematic approach that spans problem identification, solution implementation, and evaluation. It ensures that the developed artifacts maintain practical relevance while meeting rigorous scientific standards.

4.2 Detailed Application of the DSR Cycle

The Design Science Research cycle was implemented through three investigations that followed methodological progression from theoretical understanding to applied technological innovation (Hevner et al., 2004).

4.2.1 Problem Identification and Motivation

The management of disruptive events in passenger railway systems represents a widely recognized multidimensional operational challenge that significantly affects punctuality, service regularity, and the quality of user experience (Bešinović, 2020; Ghaemi et al. 2018). The literature shows that managing these situations effectively requires the capacity to quickly anticipate their multidimensional impact, including the total number of affected trains, the magnitude of resulting delays, and the number of impacted passengers (Huang et al., 2020).

The first investigation, a systematic review, identified 28 scientific articles and revealed three fundamental structural limitations:

- Most studies focus exclusively on the operational context of the Netherlands, which limits the generalization of developed solutions to other railway contexts.
- Most studies employed traditional linear programming methods, which encounter computational limitations when dealing with large-scale problems.
- Only 3.6% of studies systematically explored machine learning techniques, despite the transformative potential demonstrated in other domains.

4.2.2 Definition of Solution Objectives

The objectives follow a methodological progression that addresses the identified limitations across three fundamental dimensions:

- The first dimension identifies critical predictive attributes, including topological variables (betweenness centrality), operational variables (line typology, traffic density), and contextual variables (environmental and human factors).
- The second dimension develops multidimensional models to predict three critical metrics simultaneously: the number of affected trains, the extent of delays, and the volume of impacted passengers.
- The third dimension implements advanced deep learning architectures to predict complete operational sequences and provide recommendations for efficient replanning.

4.2.3 Design and Development

The methodology development followed an evolutionary approach, structured through three complementary investigations:

The first investigation consisted of a systematic literature review, involving critical analysis of 28 scientific articles to identify methodological gaps and establish theoretical foundations. This investigation adopted the perspective of railway transport operators, distinguishing themselves from conventional approaches that focus primarily on infrastructure management.

The second investigation focused on multi-target predictive models. Using the CRISP-DM methodology, models based on classical machine learning techniques were developed. The study analyzed 89,338 operational records from CP spanning 2015 to 2022, implementing multiple architectures including Random Forest (RF), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Multilayer Perceptron (MLP). A key innovation in this phase was the multidimensional approach for simultaneous prediction of three impact metrics, incorporating previously neglected variables such as the betweenness centrality indicator.

The third investigation involved developing a specialized Transformer architecture. This phase created an encoder-decoder model specifically adapted to the railway domain, utilizing 371,668 operational records from 2016 to 2024. The main architectural innovation was a semantic tokenization strategy employing structural delimiters to preserve the hierarchical integrity of railway operational data.

4.2.4 Demonstration

The artifact demonstration illustrated how each solution directly addresses the problems identified in railway disruption management, validating their practical applicability through implementation in real operational scenarios at CP.

The first investigation employed a systematic review to identify methodological gaps in the state of the art. This investigation systematically mapped existing paradigm limitations and established the theoretical foundations necessary to guide the development of subsequent solutions.

The second investigation developed multidimensional predictive models to address the integrated forecasting of disruptive impacts. These machine learning-based artifacts processed historical operational data and generated simultaneous predictions for three critical metrics, thereby providing railway managers with practical decision-support tools.

The third investigation created a specialized Transformer architecture to predict complete operational sequences after disruptions. This artifact captured long-term dependencies through attention mechanisms, while semantic tokenization preserved the hierarchical integrity of railway data. The multidimensional plausibility framework provided actionable recommendations for efficient operational replanning.

4.2.5 Evaluation

The artifact evaluation followed a multidimensional approach that integrated multiple analytical perspectives. This approach ranged from bibliometric quality criteria to innovative statistical plausibility metrics. Each investigation applied specific validation methods appropriate to the nature of the developed artifacts, ensuring scientific rigor and practical applicability of the proposed solutions.

The first investigation was evaluated through rigorous methodological quality criteria based on quartile classification of the periodical publications where the studies were disseminated. Studies published predominantly in first-quartile journals provided a substantial indicator of the scientific maturity of the investigative domain. This prevalence lent credibility to the conclusions and validated the methodological robustness of the systematic review.

The second investigation used classical regression and classification metrics. We implemented a multi-scale experimental strategy with datasets of increasing dimensions to ensure robust validation. The primary metrics included Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics allowed objective comparison between different architectures (RF, CNN, RNN, MLP)

and preprocessing techniques. The evaluation revealed differentiated performance patterns across dataset scales. Neural networks showed superior performance on smaller sets, while tree-based methods performed better on extensive volumes. These findings empirically validated the adequacy of different approaches for specific operational contexts.

The third investigation implemented an innovative multidimensional evaluation framework. This framework complemented traditional metrics with a statistical plausibility metric composed of six components derived from historical operational patterns: station sequence, point-to-point transitions, position-specific values, temporal intervals, task durations, and global route pattern. The methodology incorporated statistical robustness metrics including Kullback-Leibler divergence between component distributions, perplexity evaluation, and plausibility analysis through contextualized z-scores and percentiles.

4.2.6 Communication

The communication of results from the three investigations followed specific approaches suited to each type of contribution. The dissemination ensured that the research findings reached both the academic community and railway sector professionals.

The first investigation was communicated through a scientific publication accepted in the Transportation Research Record journal, establishing solid conceptual foundations for subsequent research development. The systematic literature review served as a fundamental theoretical foundation, identifying and characterizing critical methodological gaps in the international state of the art. This initial communication established the necessary scientific framework to justify the innovations proposed in subsequent investigations, creating a theoretical reference that demonstrates the need for a paradigm shift from traditional methods to artificial intelligence-based approaches.

The second investigation resulted in a scientific publication accepted in Public Transport (Springer) a prestigious international journal indexed in reference databases. This formal communication to the academic community documents the pioneering application of multidimensional predictive models in the Portuguese railway context, demonstrating the viability and superiority of classical machine learning techniques over traditional paradigms. The publication establishes a scientific milestone that empirically validates the effectiveness of the proposed approaches, providing a methodological reference for future investigations in the domain.

The third investigation was communicated through detailed documentation of specialized Transformer architectural innovations, complemented by structured workshops with experienced CP dispatchers. These workshops revealed critical aspects concerning

the practical applicability of the proposed methods and provided fundamental insights for operational implementation of the developed solutions. The transfer to operational practice was achieved through collaborative validation of the proposed functionalities, ensuring that technological innovations effectively address real operational needs. This bidirectional communication enabled the refinement of the solutions developed based on domain expert feedback, creating a virtuous cycle of continuous improvement that integrates academic knowledge and specialized operational experience.

This structured communication strategy lays the groundwork for the systematic presentation of the three investigations that follow, beginning with the critical literature review in Chapter 3.

The methodological structure defined here underpins the development and experimental evaluation presented in Chapter 5, where predictive models are built and compared. It also provides the foundation for the prescriptive sequence-generation architecture explored in Chapter 6.

Chapter 5: Development and Experimental Evaluation of Predictive Models

Anchored in the methodological framework of Chapter 4 and the research needs identified in Chapters 2 and 3, this chapter develops and evaluates multidimensional predictive models that quantify disruption impacts (affected trains, delay minutes, impacted passengers) using real operational data.

5.1 Introduction

The results of the systematic review presented in Chapter 3 identified three critical limitations: first, extreme geographical fragmentation; second, predominance of traditional linear programming methods; and third, exploration of artificial intelligence techniques (only 3.6% of the studies). Directly responding to these gaps, this chapter develops the second investigation, focusing on the application of advanced machine learning techniques to the Portuguese railway context.

This section seeks to provide a first contribution to addressing some of the gaps evidenced in the systematic literature review by developing predictive models using artificial intelligence techniques. The second article of this research, "Data-driven insights to reduce uncertainty from disruptive events in passenger railways", published in Public Transport (Springer), responds to the under exploration of machine learning techniques identified in the state of the art, where only 3.6% of the analyzed studies used these technologies.

The study uses 89,338 records of disruptive events from CP (2015-2022), constituting one of the most extensive databases applied to this domain in the Iberian context. This breadth allows the exploration of complex operational patterns previously unexplored in international literature. The main methodological innovation consists of the multidimensional approach that simultaneously predicts three impact metrics: number of affected trains, total delay minutes, and number of impacted passengers. This integrated perspective overcomes the fragmentation of existing literature, where studies traditionally addressed isolated metrics.

The need to develop multidimensional predictive models for quantifying the impacts of disruptive railway events is grounded in specific limitations identified in the literature on operational disruption forecasting. Bešinović (2020) documents that the occurrence of disruptive events in passenger railway systems constitutes a critical challenge that affects punctuality, regularity, and overall passenger experience, emphasizing the need for robust predictive tools for proactive management of these disruptions. The

literature shows that conventional approaches face fundamental methodological constraints related to the integrated forecasting capability of multiple impact dimensions.

Yaghini et al. (2013) demonstrated critical limitations in the application of artificial neural networks for railway forecasting, identifying dependence on extensive training data volumes as a fundamental obstacle to methodological scalability. Huang et al. (2020) documented that models based on Bayesian networks, although achieving high accuracy, face difficulties related to algorithmic complexity and extensive specialized knowledge requirements that compromise operational implementation. Grandhi et al. (2021) evidenced significant challenges in the accuracy of data manually entered predictive systems applied to the Danish network, highlighting the critical need for automated approaches based on structured historical data. Chen et al. (2022) identified structural gaps in the availability of information on operational procedures and infrastructure data in models applied to Asian systems, limiting predictive capability in complex disruption contexts.

The literature also documents limitations in the methodological generalization capability across different railway contexts. Li et al. (2020) recognized that the dependence on specific historical data can compromise the transferability of XGBoost and SVR models between different railway lines, while Li et al. (2021) identified similar constraints in the application of Random Forest models to the Dutch network. Marković et al. (2015) highlighted the critical importance of infrastructure as a determining factor in delay forecasting but limited empirical validation to restricted samples that compromise statistical robustness.

The research incorporates previously neglected predictive variables, with emphasis on the betweenness centrality indicator, which quantifies the strategic importance of different network sections in terms of connectivity and disruption propagation. Infrastructure characteristics, environmental, temporal, and operational factors are also integrated. The methodological development is based on the systematic integration of operational variable categories documented in specialized literature. The first investigation of this thesis identified that railway operators control differentiated variables such as crew management and rolling stock (Corman & Meng, 2015), while the literature documents 35 numerical variables related to delay modeling, including absolute delays, accumulated delays, and disruption propagation (Huang et al., 2020). Grandhi et al. (2021) evidenced the critical importance of climatic variables such as temperature and winter weather conditions in total delay forecasting, while Chen et al. (2022) identified temporal factors such as time of day and weather conditions as statistically significant.

The methodological development encompasses multiple machine learning architectures: Multilayer Perceptron Neural Networks, Convolutional and Recurrent Networks,

Random Forest, and Extra-Trees. The comparative analysis reveals that neural networks demonstrate superior effectiveness in smaller datasets, while tree-based methods stand out in extensive data volumes. The results establish new reference standards for disruptive impact forecasting, demonstrating substantial improvements in predictive accuracy. The validation confirms the superiority of winsorization based on the interquartile range over conventional methods, establishing practical guidelines for future implementations.

This convergence of limitations evidences the need to develop predictive approaches that simultaneously integrate multiple impact metrics, incorporate previously neglected infrastructure and network topology variables, and use extensive operational datasets that allow the capture of complex disruption patterns through advanced machine learning techniques specifically adapted to the railway domain.

5.2 Methodology

This section outlines the methodology used to develop predictive models. It covers the stages of data preparation, the selection of relevant attributes, and the modeling procedures that were applied. The rationale behind the choice of machine learning techniques is explained, along with the validation strategies employed to ensure robust results.

5.2.1 Comprehensive Methodological Analysis

The use of CRISP-DM in this study on predictive modeling of the impact of disruptive events in passenger railway systems is justified by its robust and flexible structure, covering all critical phases of the research. Starting with business and data understanding, the modeling phase allows the testing and optimization of different algorithms, such as neural networks and decision trees, while the evaluation phase ensures rigorous validation of the chosen models. The final implementation phase facilitates the integration of the models into Comboios de Portugal operations, ensuring practical and continuous application. This proven methodology provides a systematic and iterative approach, crucial for dealing with the complexity and variability of real data used in the study, ensuring robust and applicable results (Martinez-Plumed et al. 2021).

The six phases of CRISP-DM are illustrated in Figure 7 and detailed below:

1. Business understanding: Understanding the challenges faced by the railway and public transport sectors, particularly due to disruptive events. The objective is to develop predictive models to mitigate these impacts, improving the resilience and robustness of the railway system operated by CP. Creating models that predict and manage the impacts of disruptions with greater accuracy is essential.

2. Data understanding: The real data provided by CP includes information on railway infrastructure, train schedules, disruptive event records, and operational data. Weather data were collected from the Visual Crossing platform. Exploratory analysis involved the identification and analysis of key features, such as the betweenness centrality indicator, railway tracks, time of day, and train service group, to understand their relevance and impact on railway operations.

3. Data preparation: Data cleaning involved the removal of inconsistent, missing, or duplicated data to ensure the quality and accuracy of model inputs. Techniques such as outlier removal using the Interquartile Range (IQR) and Standard Deviation (SD) were employed to handle extreme values. Data transformation included the conversion of data into a format suitable for modeling, including normalization and encoding of categorical variables. Feature engineering was applied to create new features, such as centrality metrics, which enhance the performance of predictive models. Techniques such as the Chi-square test, LASSO, and Recursive Feature Elimination (RFE) were used for dimensionality reduction and selection of the most relevant features.

4. Modeling: Multilayer perceptrons (MLP) are particularly effective in modeling complex nonlinear relationships and can capture complex patterns and subtle interactions between input and output variables, making them suitable for nonlinear predictions. CNN, although traditionally used for image data, can also be applied to time series and spatially structured data. In the context of CP data, CNN can detect spatial and temporal patterns in passenger data, including the number of trains affected by delays, by learning feature hierarchies. RNN are designed to handle sequential and temporal data, such as those provided by CP, capturing long-term dependencies and patterns in time series, which are essential for predictions based on historical data. Random Forest combines multiple decision trees to reduce the risk of overfitting, capturing complex interactions between variables. Extra-Trees uses a more random splitting process than Random Forest, increasing diversity among trees and potentially improving performance on certain datasets. This technique is computationally efficient and can handle variability and complexity (Müller 2016).

5. Evaluation: Metrics such as precision, recall, F1-score, MSE, and RMSE are used to evaluate the predictive effectiveness of the models. OHE (One-hot) encoding is applied to the 'Geographic Area' variable to convert its categories into binary values, allowing compatibility with different machine learning models. A

comparison was made between the performance of neural networks and decision trees in different data scenarios, identifying the most suitable models for smaller and larger datasets, as well as their relationship with existing literature.

6. Implementation: Operational integration involves the implementation of the developed models in CP's daily operations to predict the impacts of disruptive events.

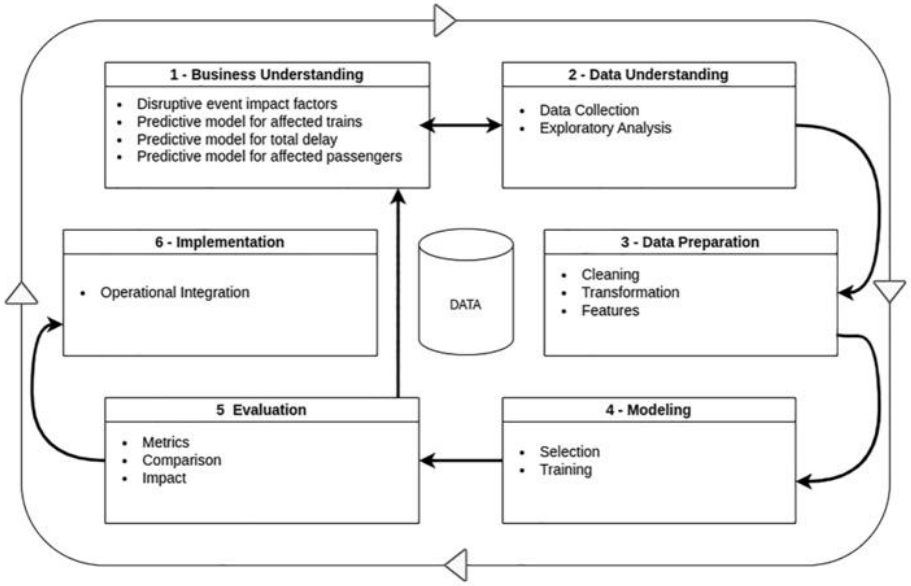

Figure 7: The six phases of CRISP-DM, adapted from Martinez-Plumed et al. (2021)

The methodological flow applied in this research illustrates the practical implementation of the study. Figure 8 presents the complete workflow developed for the construction and evaluation of predictive models used in railway disruption management

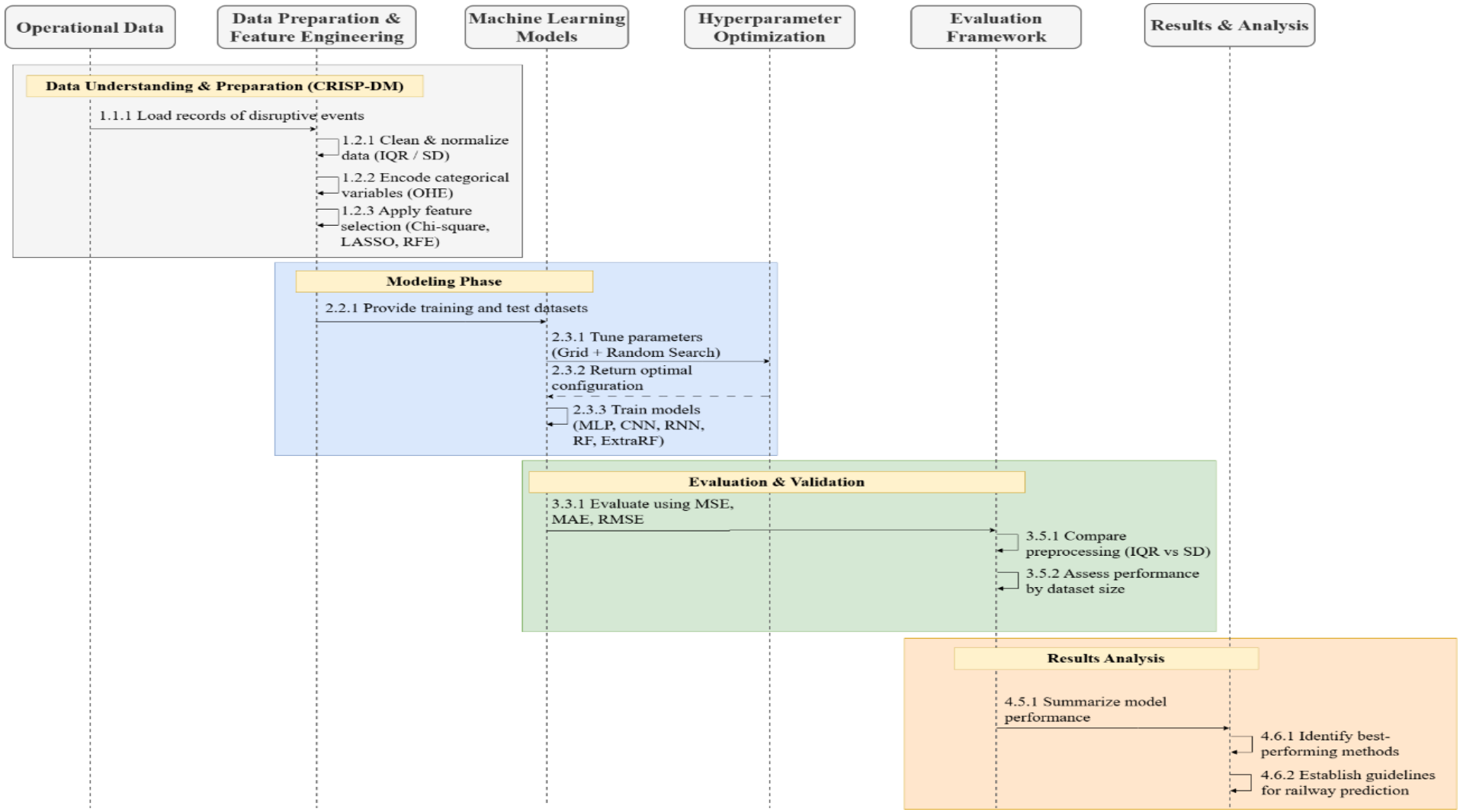

Figure 8: Sequence diagram of the practical implementation of the Machine Learning model

As shown in the figure, it depicts the practical sequence followed during the research, from data collection and preparation to model training, evaluation, and result analysis. Table 3 complements the figure by describing each stage and information flow represented in the diagram, clarifying the specific function of every step in the practical workflow.

Table 3: Description of the stages and processes of the workflow shown in Figure 8

Section / Step	Code	Description (American English)
1. Data Understanding & Preparation (CRISP-DM)	1.1.1	Load records of disruptive events: imports real operational data on disruptive events from CP.
	1.2.1	Clean & normalize data (IQR / SD): removes extreme values and normalizes variables based on the interquartile range or standard deviation.
	1.2.2	Encode categorical variables (OHE): convert categorical variables (such as geographic area or line type) into numerical format using one-hot encoding.
	1.2.3	Apply feature selection (Chi-square, LASSO, RFE): selects the most relevant attributes to reduce dimensionality and improve model performance.
2. Modeling Phase	2.2.1	Provide training and test datasets split the data into training and testing subsets.
	2.3.1	Tune parameters (Grid + Random Search): performs hyperparameter optimization using grid and random search strategies.
	2.3.2	Return optimal configuration selects the combination of parameters that yields the best performance.
	2.3.3	Train models (MLP, CNN, RNN, RF, ExtraRF): trains different machine learning models — neural networks and ensemble methods — using the prepared data.
3. Evaluation & Validation	3.3.1	Evaluate using MSE, MAE, RMSE: measures prediction errors using standard evaluation metrics such as mean squared error, mean absolute error, and root mean squared error.
	3.5.1	Compare preprocessing (IQR vs SD): evaluates how different outlier-handling methods affect model performance.
	3.5.2	Assess performance by dataset size tests the models on datasets of varying sizes to assess robustness and scalability.

4. Results Analysis	4.5.1	Summarizing model performance synthesizes the results obtained and identifies the best-performing models.
	4.6.1	Identify best performing methods determines the most effective techniques (e.g., Random Forest with IQR preprocessing).
	4.6.2	Establish guidelines for railway prediction defines practical recommendations for future railway disruption prediction applications.

5.2.2 Data Description

This study analyzed the period from 2015 to 2022. Due to the COVID-19 pandemic, reliable data on daily passenger numbers at the checkpoint were unavailable for 2021 and 2022. Therefore, missing values were replaced with the average from the remaining years in the dataset. Lockdown restrictions during this period compromised CP's capacity to provide reliable measurements for this variable. The selected timeframe yielded 166,471 disruptive event records. For the analysis, only records documenting affected trains and assigned delay minutes were retained, producing a final dataset of 89,338 disruptive event records.

Figure 9 illustrates the betweenness and centrality of the railway network under examination. This visualization helps identify critical points and areas with greater influence on the network. Nodes approaching 1 (lighter color) indicate higher influence, while those closer to 0 (darker color) represent lower influence on network operations.

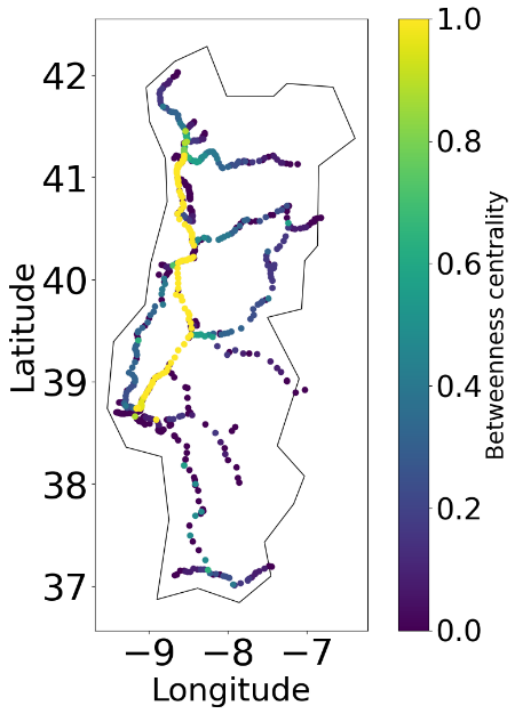

Figure 9: Betweenness centrality of the railway network under analysis

Table 4 presents the variables from the dataset collected for the experiments in this study. Variables 1 through 26 are independent variables, while variables 27 through 29 are dependent variables.

Table 4: Variable description

#	Abbreviated form	Description	Data type	Data range	Number of unique values
1	AltModes	Number of alternative transport modes	Num	0 to 87	32
2	Suppressions	Total number of suppressions occurred	Num	0 to 1,498	36
3	Un-plannedTrains	Total number of unplanned trains	Num	0 to 39	19
4	DailyTrains	Average daily number of trains at facility	Num	2 to 419	202
5	BetwCent	Betweenness centrality indicator	Num	0 to 0.44	32
6	CloseCent	Closeness centrality indicator	Num	0 to 0.42	28
7	Headway	Time interval between trains (headway)	Num	4 to 145	95

8	Track	Railway track	Cat	0 to 3	4
9	InfraDam	Infrastructure damage	Cat	0 to 1	2
10	RollStockDam	Rolling stock damage	Cat	0 to 1	2
11	AssistReq	Assistance request	Cat	0 to 1	2
12	MinorInj	Number of minor injuries	Num	0 to 34	6
13	SeriousInj	Number of serious injuries	Num	0 to 3	3
14	Deaths	Number of recorded deaths	Num	0 to 3	4
15	PaxPerDay	Number of passengers per day at checkpoint	Num	28 to 164,359	339
16	Km	Number of kilometers (of the train that originated the disruptive event)	int64	4,265 to 631,816	102
17	Area	Geographic area	Cat	1 to 16	16
18	CauseID	Incident cause group identity (ID)	Cat	110 to 999	37
19	GeoCtrl	Geographic checkpoint	Cat	1,008 to 95,125	416
20	SerialID	Rolling stock serial identification	Cat	350 to 9,630	12
21	Temp	Temperature in degrees Celsius	Num	-3.0 to 35.3	352
22	WindSpd	Wind speed (km/h)	Num	0.3 to 200.4	558
23	Hour	Time of day	Cat	0 to 23	24
24	DayWeek	Day of week	Cat	0 to 6	7
25	Month	Month of year	Cat	1 to 12	12
26	ServGroup	Train service group	Cat	1 to 39	29
27	TrainsDelay	Number of trains affected by delay	Num	1 to 3,839	129
28	TotDelay	Total number of delay minutes	Num	1 to 25,001	792
29	PaxAf	Number of affected passengers	Num	1 to 615,226	4,967

5.2.3 Data Processing

The analysis focused on three aspects: the number of delayed trains, total delays in minutes, and the number of affected passengers. Passenger numbers were classified into discrete categories to

simplify the prediction task, while the other variables were analyzed using regression methods. The first category ranged from 1 to 300 passengers, where 300 corresponds to the minimum occupancy of the smallest train in operation.

This classification approach creates a direct link between the predicted category and the rolling stock capacity. Figure 10 illustrates the distribution of the 29 variables, revealing diverse data distributions, skewness, and kurtosis patterns. Variables 1 through 3 exhibited limited variation and lower median values, whereas variables 19 through 21 displayed greater variation and more extreme values, indicating a broader data range. Variables 22 through 24 demonstrated more uniform distributions with fewer outliers. The whiskers in the boxplots extended primarily to 1.5 times the interquartile range, representing typical data ranges. This visual analysis revealed notable differences in centrality and dispersion across the variables.

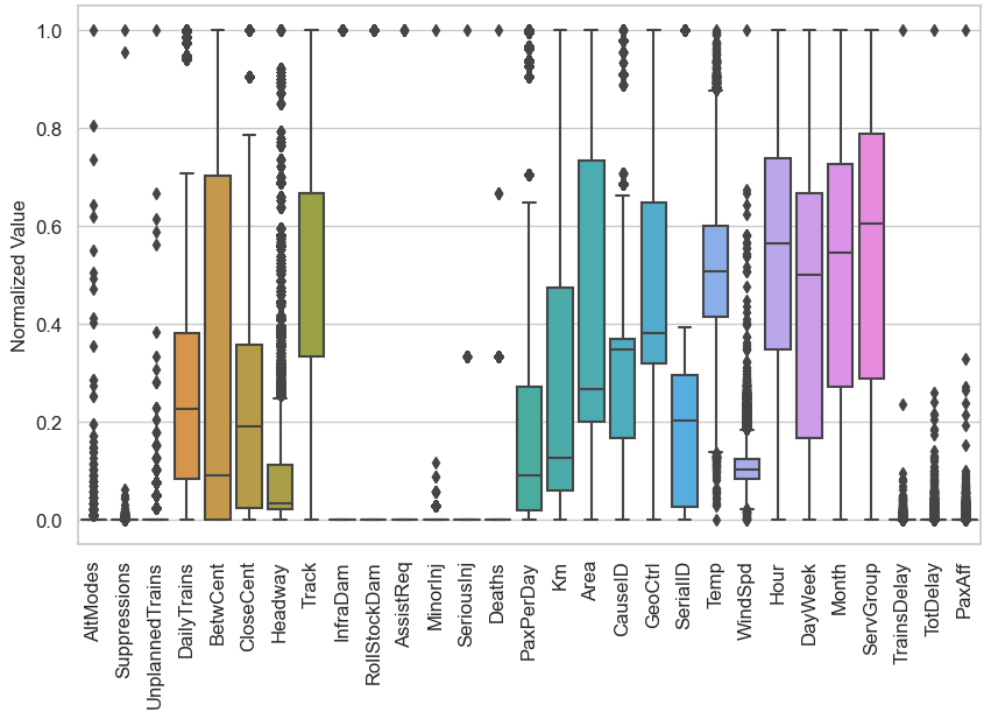

Figure 10: Original dataset (normalized data)

5.2.4 Outlier Treatment

Two commonly used techniques were employed for outlier removal: the interquartile range (IQR) and the standard deviation (SD). The IQR, a measure of statistical dispersion, is defined as the difference between the third and first quartiles (Q3 and Q1, respectively) in a data distribution. This measure indicates the median variation of the data and is particularly useful in datasets with non-normal distributions (Smiti, 2020). The IQR is frequently used to identify outliers, which are defined as values that fall below Q1 - 1.5 IQR or above Q3 + 1.5 IQR.

5.2.5 Dimensionality Reduction

Dimensionality reduction helps identify redundant variables and stabilize models. A common technique is the Chi-square test, which assesses the association or independence between two categorical variables. It compares observed and expected frequencies to determine whether the variables are independent. The test statistics are compared with a critical value from the Chi-square distribution (Yaghini et al. 2013). If the test statistic exceeds the critical value, the null hypothesis of independence is rejected. Another technique for dimensionality reduction is the least absolute shrinkage and selection operator (LASSO). LASSO regularization penalizes the absolute value of regression coefficients, forcing some coefficients to shrink to exactly zero. This property indicates that features associated with zero coefficients have no impact on the model, effectively reducing the number of features used (Klumpenhouwer and Shalaby 2022).

Similarly, recursive feature elimination (RFE) is a feature selection method that systematically removes the least important features from a model. The approach begins by fitting the model with all available features, which are then ranked according to the absolute values of their coefficients. The feature with the smallest absolute coefficient value is removed, and the model is refit with the remaining features. This iterative process continues until the desired number of features is achieved (Tiong et al. 2023). Both LASSO and RFE prove effective in simplifying models while retaining the most significant features, thereby enhancing model performance and interpretability.

5.2.6 Datasets

This section demonstrates how data were processed for each independent attribute under study. The analysis examined several key independent attributes, including the number of alternative transport modes, total suppressions, and unplanned trains. The study evaluated the average daily number of trains at each facility, betweenness and closeness centrality indicators, and intervals between trains.

Railway track conditions, infrastructure and rolling stock damage, and assistance requests were all considered. Injury data included both minor and serious injuries as well as fatalities. Additional variables comprised passenger numbers at checkpoints, kilometers traveled by disrupted trains, geographic areas of incidents, cause group identities, checkpoint identifiers, and rolling stock IDs. Environmental factors such as temperature and wind speed were incorporated alongside temporal elements (time of day, day of week, and month) and train service groups.

Three dependent variables were analyzed: the number of trains affected by delays, total delay minutes, and the number of affected passengers. These variables underwent dimensionality reduction.

5.2.6.1 Number of Trains Affected by Delay

Figure 11 illustrates the removal of outliers. To facilitate graphical visualization, the values were logarithmically transformed. The left graph uses standard deviation, while the right graph employs the IQR method for delay minutes prediction. Only attributes with statistically significant differences (p-value < 0.05) are presented. Both graphs include the average daily train count, centralities, interval between trains, presence of tracks, total kilometers, geographic area, incident cause, checkpoint, rolling stock, day of week, and service group. The standard deviation method additionally incorporates alternative transport, total suppressions, temperature, and month. In contrast, the IQR method uniquely includes passenger count at the checkpoint and time of day.

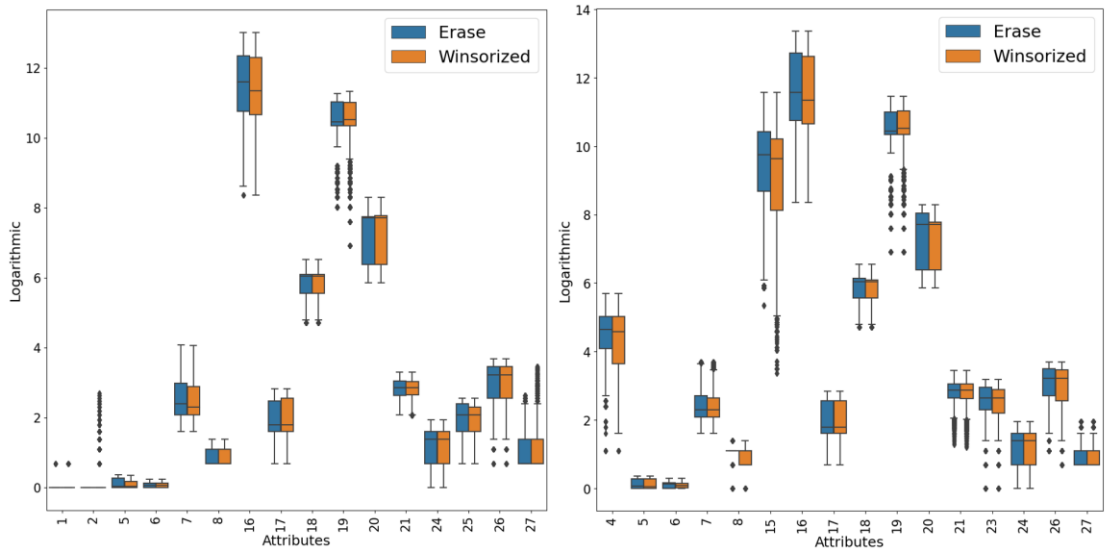

Figure 11: Attributes with statistically significant differences (Number of trains affected by delay). Variables are numbered as per Table 4

5.2.6.2 Total Number of Delay Minutes

Figure 12 illustrates the removal of outliers using standard deviation (left) and IQR (right) for predicting train delays. Both methods share several common attributes: average daily train count, centralities, interval between trains, track presence, total kilometers, geographic area, incident cause ID, checkpoint, rolling stock ID, day of week, and service group classification. The standard deviation method uniquely incorporates alternative transport, total suppressions, temperature, and month. The IQR method uniquely includes daily passenger count at the checkpoint and time of day.

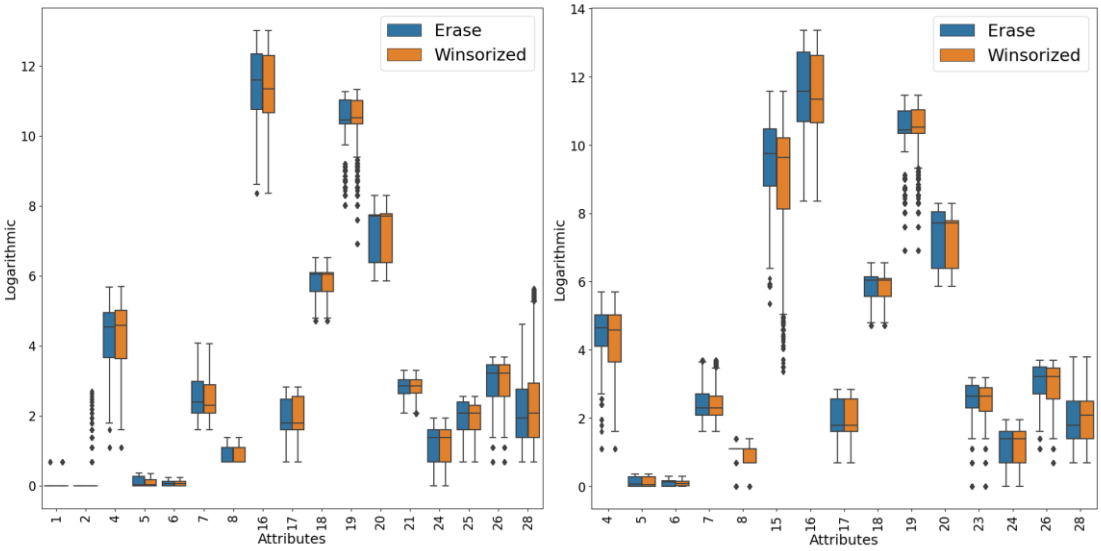

Figure 12: Attributes with statistically significant differences (Total number of minutes of delay). Variables are numbered as per Table 4

5.2.6.3 Number of Affected Passengers

Figure 13 illustrates the removal of outliers for the number of passengers, using standard deviation (left) and IQR (right), with statistically significant attributes (p-value < 0.05). Common to both methods are betweenness and closeness centralities, interval between trains, presence of tracks, train kilometers, covered area, incident cause ID, checkpoint ID, rolling stock ID, day of week, and service group classification. Exclusively in the standard deviation method are alternative transport, total suppressions, temperature, and month. Exclusively in the IQR method are average daily train count, daily passenger count at checkpoints, and time of day.

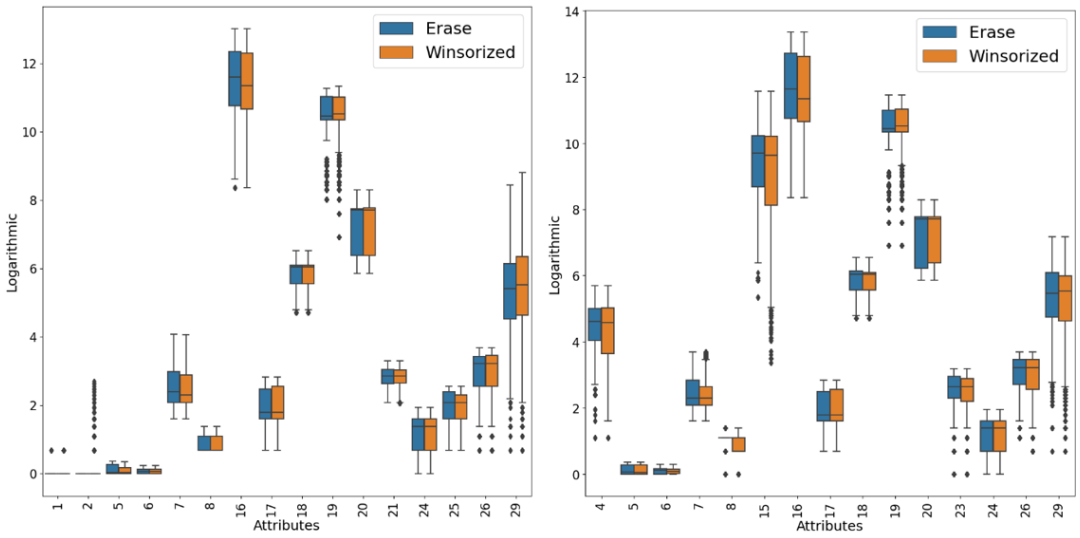

Figure 13: Attributes with statistically significant differences (Number of passengers affected). Variables are numbered as per Table 4

5.2.7 Hyperparameters

A hybrid approach was used for hyperparameter optimization, combining grid search and random search (Belete and Huchaiah 2022). First, grid search provides a comprehensive analysis using a predefined matrix of values, including learning rate and neuron count. Then, having identified a promising range, we shifted to random search for faster and less resource-intensive exploration. This combination produces an optimal combination of hyperparameters under certain assumptions, balancing computational cost and efficiency, resulting in a model with good fit to the training data and effective generalization to new data. For our grid strategy, we used the values described in Table 5 and Table 6.

Table 5: Combined hyperparameters in neural networks

Parameter	Values
Hidden layer sizes	[(128,), (64,), (32,), (128, 64), (128, 32), (64, 32), (128, 64, 32), (128, 64, 16), (128, 32, 16), (64, 32, 16), (128, 64, 32, 16)]
Alpha	[0.0001, 0.001, 0.01]
Initial learning rate	[0.001, 0.01]
Batch size	[32, 64]
Activation	['sigmoid', 'tanh', 'relu', 'adam', 'softmax']

Table 6: Combined Hyperparameters in random forests

Parameter	Values
n_estimators	[50, 100, 150, 200]
Max depth	[None, 10, 20, 30]
Min samples split	[2, 5, 10]
Min samples leaf	[1, 2, 4]
Criterion	['gini', 'entropy', 'Poisson']

5.2.8 Models Used

5.2.8.1 Multilayer Perceptron (MLP)

A Multilayer Perceptron (MLP) network is composed of multiple layers of neurons, where each neuron in a layer is connected to all neurons in the following layer. The output of a neuron is given by:

$$ai = f(\sum_{j} w_{ij} x_j + b_i) \quad (1)$$

Where f is the activation function, w_{ij} is the weight of the connection between neurons i and j, x_j is the input from the previous layer, and b_i is the bias term (Su et al. 2022).

5.2.8.2 Recurrent Neural Network (RNN)

An RNN processes input sequences X = $(x_1, x_2, \ldots, x_t)$ sequentially over time. At each time step t, the hidden state h_t is updated based on the previous hidden state h_{t-1} and the current input:

$$X_t : h_t = f(W_{hh} h_{t-1} + W_{xh} x_t + b_h) \quad (2)$$

where f is a nonlinear activation function, W_{hh} and W_{xh} are weight matrices, and b_h is a bias vector (Su et al. 2022).

5.2.8.3 Convolutional Neural Network (CNN)

CNNs use compact filters to create feature maps by identifying specific attributes in the input data, such as edges and textures. In a Conv1D layer, for an input $X(dimension(W, 1))$ and a filter $F(dimension(3, 1))$, the convolution is:

$$Y[i] = \sum_{j=0}^{2} X[i + j] \cdot F[j] \quad (3)$$

Where Y is the output of the convolutional layer (Boateng and Yang 2023).

5.2.8.4 Random Forests (RF)

In classification tasks using Random Forest, each tree in the ensemble makes a class prediction for a new input, and the final assigned class is the one that receives the most votes from the trees. In regression tasks, the result is the average of the outputs from the different trees. Essentially, each tree votes for a class, and the class with the most votes is chosen as the final prediction (Nabian et al. 2019).

5.2.8.5 RF-Extra

In Extra-Trees, randomness is introduced in two main ways. First, as in RF, a random subset of features is used. Second, unlike RF, the split points in the trees are not chosen deterministically; instead, for each selected feature, a split value is chosen randomly, and the best of these points is used to split the node. This increases diversity among the trees in the model, helping to reduce model variance. The prediction of an Extra-Trees ensemble is given by the average (in regression) or by the mode (in classification) of the predictions from all individual trees (Geurts et al. 2006).

5.3 Results

This section presents the results obtained in the experimental evaluation of the developed predictive models. The different architectures tested are compared, evidencing their strengths and weaknesses, and the impact of methodological choices on final performance is analyzed.

5.3.1 Number of Trains Affected by Delay

The attributes closeness centrality, railway track, time of day, and train service group emerge as the most predominant, being present in 12 datasets. This observation suggests their significant relevance and central influence in the analyses addressed, emphasizing their fundamental role in the investigations conducted. Next are the attributes average daily trains, interval between trains, daily passengers at checkpoint, and total train kilometers, each appearing in eleven datasets. Subsequently, there is a group of attributes comprising the betweenness centrality measure, geographic area coverage, incident cause ID, and geographic checkpoint identification, each present in ten datasets. The day of week attribute is found in nine datasets. In turn, the month of year attribute is mentioned in eight datasets, indicating notable importance, although not as widespread as the others.

Finally, the attributes alternative transport, temperature in degrees Celsius, and wind speed in km/h, which appear with lower frequency, are present in the betweenness centrality measure, average daily train count, and four datasets, respectively. This frequency pattern may indicate a more limited application or specific relevance in certain analytical contexts. The total suppression attribute, appearing in only two datasets, stands out as the least frequent, which may suggest a very specific use or limited focus in the analyses undertaken.

5.3.1.1 Hyperparameters

The MLP model, with four dense layers and L1 regularization in the first layer to prevent overfitting, comprises 128, 64, and 32 neurons, all with 'relu' activation, and a linear output layer. We found that dropout was not useful for improvements. The data split was 80% for training and 20% for testing.

RNN: The RNN model includes two recurrent layers (64 and 32 units) with 'relu' activation, followed by a dense output layer. The data split was the same as for MLP.

The CNN model consists of three 1D convolutional layers (128, 64, and 32 filters), a flattening layer, and a dense output layer. Dropout was not effective here, and 'relu' activation was used throughout. The data split is consistent with MLP and RNN.

We used an alpha of 0.01, an initial learning rate of 0.001, and a batch size of 32. Everything else remained consistent across all models.

The Random Forest model uses 200 trees and the 'Poisson' criterion, with a fixed random seed of 42 for consistency.

The RF-Extra model: Similar to RF, the Extra-Trees model employs 200 trees and the 'Poisson' criterion with the same random seed, offering more randomness in tree construction.

A maximum depth of 20, minimum samples split of 5, and minimum samples leaf of 4 were employed for all Random Forest models.

5.3.1.2 Results

Table 7 describes the best results for datasets with approximately 8,000 records, with CNN and RNN showing particularly good performance on smaller datasets. The very small variation in indicators across various preprocessing techniques emphasizes the robustness of these models in the face of differences in data treatment.

Table 8 describes the best results for datasets with approximately 18,000 records. RF with IQR winsorization and Chi-square proved effective, achieving an MSE of 1.53, an MAE of 0.93, and an RMSE of 1.24. In contrast, the RF-Extra variant did not outperform the standard RF model.

The MLP network achieved an MSE of 1.65 and an MAE of 0.86 using IQR winsorization and RFE, demonstrating its ability to efficiently handle large data volumes.

This analysis demonstrates the importance of selecting appropriate models and preprocessing strategies in machine learning. CNN and RNN are suitable for smaller datasets, while RF and MLP perform better with larger datasets. Using Standard Deviation Removal and LASSO in an RF model on a dataset of 7,934 records resulted in an MSE of 3.85, an MAE of 1.31, and an RMSE of 1.96, highlighting its effectiveness for large data volumes.

The study deepens the understanding of how preprocessing methods impact model accuracy and efficiency. There is a clear correlation between preprocessing types and model performance, particularly regarding the MSE and MAE metrics. The effectiveness of techniques such as removal and winsorization, based on IQR and SD, depends on dataset size.

In smaller datasets (around 8,000 records), IQR-based outlier removal is efficient in CNN and RNN models. IQR removal in CNN produced an MSE of 1.49 and an MAE of 0.84, while in RNN it resulted in an MSE of 1.50 and an MAE of 0.87. This suggests that IQR is more beneficial for smaller datasets in neural network models. In contrast, for larger datasets (around 18,000 records), standard deviation-based winsorization is better suited for RF and MLP models. This method in RF led to an MSE of 3.84 and an MAE of 1.31, and in MLP, an MSE of 4.15 and an MAE of 1.17, indicating its effectiveness for larger datasets in complex data models.

Additionally, incorporating techniques such as Chi-square, LASSO, and RFE, with or without OHE, adds complexity to the analysis. Combining IQR winsorization and

LASSO in an RF model for a larger dataset achieved an MSE of 1.56 and an MAE of 0.93, showing the significant role of feature selection in enhancing model performance.

Table 7: Results for number of trains affected (≈ 8,000 records)

Dataset	Method	MSE	MAE	RMSE	Records
IQR erase with Chi-squared	CNN	1.49	0.84	1.22	8,098
IQR erase with RFE and OHE	RNN	1.49	0.85	1.22	8,098
IQR erase with LASSO and OHE	CNN	1.49	0.85	1.22	8,098
IQR erase with Chi-squared and OHE	CNN	1.50	0.83	1.22	8,098
IQR erase with LASSO	RNN	1.50	0.87	1.23	8,098
IQR erase with RFE	RNN	1.52	0.82	1.23	8,098
SD erase with LASSO and OHE	RF	3.84	1.31	1.96	7,934
SD erase with LASSO	RF	3.85	1.31	1.96	7,934
SD erase with Chi-squared and OHE	RF	3.90	1.32	1.97	7,934
SD erase with Chi-squared	RF	3.91	1.32	1.98	7,934
SD erase with RFE	RNN	4.11	1.19	2.03	7,934
SD erase with RFE and OHE	MLP	4.15	1.17	2.04	7,934

Table 8: Results for number of trains affected (≈ 18,000 records)

Dataset	Method	MSE	MAE	RMSE	Rec-ords
IQR winsorized with Chi-square	RF	1.53	0.93	1.24	17,868
IQR winsorized with Chi-square and OHE	RF	1.54	0.93	1.24	17,868
IQR winsorized with LASSO	RF	1.56	0.93	1.25	17,868
IQR winsorized with LASSO and OHE	RF	1.56	0.93	1.25	17,868
IQR winsorized with RFE and OHE	MLP	1.59	0.87	1.26	17,868
IQR winsorized with RFE	MLP	1.65	0.86	1.28	17,868
SD winsorized with LASSO	RF	9.83	1.88	3.14	17,868
SD winsorized with LASSO and OHE	RF	9.84	1.87	3.14	17,868
SD winsorized with Chi-square	RF	9.94	1.88	3.15	17,868
SD winsorized with Chi-square and OHE	RF	9.94	1.88	3.15	17,868
SD winsorized with RFE and OHE	MLP	11.12	1.64	3.33	17,868
SD winsorized with RFE	MLP	11.13	1.64	3.34	17,868
Original data	RNN	22.37	1.80	4.73	17,868
Original data with OHE	MLP	22.46	1.78	4.74	17,868
Original data with LASSO selection and OHE	MLP	22.71	1.77	4.77	17,868

Original data with Chi-square selection	MLP	23.07	1.79	4.80	17,868
Original data with Chi-square selection and OHE	MLP	23.07	1.78	4.80	17,868
Original data with LASSO selection	MLP	23.16	1.79	4.81	17,868
Original data with RFE	CNN	26.64	1.87	5.16	17,868
Original data with RFE and OHE	CNN	27.48	1.87	5.24	17,868

5.3.2 Total Number of Delay Minutes

The attributes of average daily train count, betweenness centrality measure, closeness centrality measure, and geographical control point are the most prominent, each appearing in twelve datasets.

This consistent presence suggests these variables may hold greater relevance for the analyses conducted. The attributes closeness centrality measure, train interval time, daily passenger count at control point, total kilometers per train, and rolling stock ID appear in eleven datasets each, demonstrating substantial utility across various analytical contexts. Geographical area coverage, found in ten datasets, shows moderate relevance. Wind speed in km/h, time of day, day of week, and month of year each appear in eight datasets, while incident cause ID and temperature in Celsius are present in seven datasets. Although these variables occur less frequently than the top-ranked attributes, their presence across multiple datasets indicates considerable analytical value. Interestingly, this distribution suggests that frequency of occurrence does not necessarily correlate directly with analytical impact. The attributes of alternative transport and total suppressions are the least frequent, appearing in only three and two datasets respectively. This limited occurrence indicates either restricted applicability or specialized use within the studied context.

5.3.2.1 Hyperparameters

The multilayer MLP model has three dense layers. The first layer, with 128 neurons, utilizes L1 regularization to prevent overfitting. The second layer has 64 neurons, with both employing 'relu' activation. Dropout did not significantly improve the model. It terminates with a linear output layer suitable for regression. The dataset was divided into 80% for training and 20% for testing.

The RNN model comprises two SimpleRNN layers (64 and 32 units) with 'relu' activation, capturing temporal dependencies of the data and terminating with a linear output layer. The data division is 80% for training and 20% for testing.

The CNN model presents three 1D convolutional layers (128, 64, and 32 filters) with 'relu' activation, followed by a flattening layer and a dense output layer. Dropout had no notable effect. The data division is 80% training and 20% testing.

An alpha of 0.01, an initial learning rate of 0.001, and a batch size of 32 were used. Everything else remained consistent across all models.

The random forest model includes 200 trees using the 'Poisson' criterion, selected for our data attributes, with a fixed random seed of 42 for consistent results, which is specifically designed for target variables that represent count data, such as the number of events or occurrences. This criterion ensures that the splits in the trees are optimized for the nature of the data, effectively capturing their distribution. A fixed random seed of 42 was applied to ensure consistent and reproducible results.

The RF-Extra model, similar to RF, the extra trees model uses 200 trees with the 'Poisson' criterion and a random seed of 42, offering more variability in tree construction. A maximum depth of 20, minimum sample split of 5, and minimum samples per leaf of 4 were used for all random forest models.

5.3.2.2 Results

Table 9 with the results shows that in smaller datasets of approximately 8,000 records, RF models with various preprocessing methods such as IQR removal are effective. For example, RF with IQR removal and LASSO registers an MSE of 64.19 and an MAE of 5.706, indicating good performance in predicting the number of delay minutes. In contrast, the CNN model with IQR removal and RFE (recursive feature elimination) has slightly inferior performance with an MSE of 74.77 and an MAE of 5.333, suggesting it may be less effective than RF in this scenario.

Table 10 shows the larger datasets of approximately 18,000 records; RF models with winsorized IQR and RFE variants show varied results. For example, RF with winsorized IQR and Chi-square has an MSE of 68.23 and an MAE of 5.9, showing moderate effectiveness in larger-scale predictions. RF-Extra models with original data exhibit much higher MSE values, such as 6,547.90 for RF-Extra Original, indicating reduced performance due to data complexity and volume.

Comparing models and preprocessing methods, RF models generally show more consistent performance across different preprocessing techniques than RF-Extra variants, especially for larger datasets. This highlights the importance of choosing the correct preprocessing method to optimize predictions.

Table 9: Results for the total number of delays (≈ 8,000 records)

Dataset	Method	MSE	MAE	RMSE	Records
IQR removal with LASSO and OHE	RF	64.13	5.665	8.008	7,958
IQR removal with LASSO	RF	64.19	5.706	8.012	7,958
IQR removal with Chi-square	RF	65.27	5.750	8.079	7,958
IQR removal with Chi-square and OHE	RF	65.48	5.724	8.092	7,958

IQR removal with RFE and OHE	RF	73.78	6.065	8.589	7,958
IQR removal with RFE	CNN	74.77	5.333	8.647	7,958
SD removal with LASSO	RF	223.07	9.540	14.94	7,820
SD removal with Chi-square	RF	223.26	9.540	14.940	7,820
SD removal with LASSO and OHE	RF	223.46	9.460	14.950	7,820
SD removal with Chi-square and OHE	RF	223.55	9.450	14.950	7,820
SD removal with RFE and OHE	RF	249.96	10.07	15.810	7,820
SD removal with RFE	RF	250.92	10.10	15.840	7,820

Table 10: Results for the total number of delays (≈ 18,000 records)

Dataset	Method	MSE	MAE	RMSE	Rec-ords
IQR winsorized with Chi-square	RF	68.23	5.900	8.260	17,868
IQR winsorized with LASSO	RF	68.37	5.907	8.269	17,868
IQR winsorized with Chi-square and OHE	RF	68.49	5.883	8.276	17,868
IQR winsorized with LASSO and OHE	RF	68.61	5.889	8.283	17,868
IQR winsorized with RFE and OHE	RF	72.41	6.061	8.510	17,868
IQR winsorized with RFE	RF	72.45	6.062	8.512	17,868
SD winsorized with Chi-square	RF	1,061.37	17.691	32.579	17,868
SD winsorized with Chi-square and OHE	RF	1,063.06	17.523	32.605	17,868
SD winsorized with LASSO	RF	1,117.30	18.569	33.426	17,868
SD winsorized with LASSO and OHE	RF	1,119.40	18.433	33.458	17,868
SD winsorized with RFE and OHE	RF	1,186.83	18.659	34.450	17,868
SD winsorized with RFE	RF	1,187.46	18.692	34.460	17,868
Original data with LASSO selection and OHE	RF	5,307.02	22.840	72.850	17,868
Original data with Chi-square selection and OHE	RF	5,320.49	22.860	72.940	17,868
Original data with OHE	RF	5,324.62	22.840	72.970	17,868
Original data with Chi-square selection	RF-Ex-tra	6,319.07	23.790	79.490	17,868
Original data with LASSO selection	RF	6,343.70	23.410	79.650	17,868
Original data	RF-Ex-tra	6,547.90	23.835	80.919	17,868

Original data with RFE	RF-Extra	7,091.64	27.870	84.210	17,868
Original data with RFE and OHE	RF-Extra	7,687.87	27.820	87.680	17,868

5.3.3 Number of affected passengers

The primary attributes across 11 datasets are the presence of railway track, daily passenger count, geographic area, and geographic checkpoint identification. Betweenness centrality, train headway time, day of the week, and train service group classification follow in ten datasets. Less prevalent are the average daily train count, closeness centrality, and total kilometers per train, which are present in nine datasets. The incident cause ID and rolling stock ID, found in eight datasets, to have slightly lower relevance. Time of day and temperature in degrees Celsius appear in six and five datasets, respectively. Attributes such as alternative transportation, wind speed, and month of the year, observed in four datasets each, indicate a specific focus. The least common attribute, total suppressions, is found in only two datasets.

5.3.3.1 Hyperparameters

The MLP model presents two dense layers, the first with 'relu' activation and the second with 'softmax' activation for class categorization. The 'adam' optimizer was used, with 'categorical_crossentropy' loss, and EarlyStopping was implemented (min_delta: 0.001, patience: 5). Training encompassed 100 epochs, a batch size of 32, and a validation split of 20%.

The CNN model has a 1D convolutional layer, MaxPooling, Flatten, and two dense layers. Training and compilation mirrored the MLP.

The RNN model includes a SimpleRNN layer with 80 units and a 'softmax' dense layer. It shares the MLP's compilation and training settings, including EarlyStopping. An alpha of 0.01 was used, with an initial learning rate of 0.001 and a batch size of 32. Everything else remained consistent across all models.

A RandomForestClassifier model was trained, configured with 100 decision trees, using the previously split dataset for training.

The Extra Trees model (RF-Extra) is based on the same architecture as Random Forest, using the same number of trees, but is distinguished by introducing additional randomness in the split selection process. To ensure comparability between the random forest models, the same hyperparameters were applied to both: maximum depth of 20, minimum split of 5 samples per node, and a minimum of 4 samples per leaf. The Synthetic Minority Over-Sampling Technique (SMOTE) was applied to balance class representation in cases of higher values.

5.3.3.2 Results

Table 11 describes the results for the number of passengers affected.

The data show that larger test sets result in more accurate evaluations for both CNN and RF models. Specifically, CNN accuracy improves from 0.63 to 0.80 with an expansion of the test set from 7,998 to 148,536 samples, highlighting the impact of test data volume on model generalization. CNNs exhibit notable accuracy fluctuations based on configuration and test size, especially in class handling, as observed in balanced macro averages. Conversely, RF models also show accuracy growth from 0.64 to 0.96 but differ in other metrics. The most accurate RF model has high macro and weighted averages, but a lower average area under the receiver operating characteristic curve (ROC curve), suggesting class differentiation challenges despite overall accuracy.

This analysis underscores how CNN and RF model performances vary with test set size, indicating the need for broader considerations in model selection, such as test set size, class balance, and class differentiation.

Table 11: Number of affected passengers

Dataset	Method	Accuracy	Macro average (precision, recall, F1-score)	Weighted average (precision, recall, F1-score)	Average ROC area	ROC areas per class	Records
Original SMOTE	RF	0.96	(0.97, 0.97, 0.97)	(0.96, 0.96, 0.96)	0.45	[0.99, 0, 0.43, 0.46, 0.46, 0.35, 0.52, 0.46, 0.44, 0.35, 0.47, 0.41, 0.41, 0.48, 0]	148,536
Original SMOTE Encode	RF	0.96	(0.96, 0.96, 0.96)	(0.96, 0.96, 0.96)	0.51	[0.99, 0, 0.42, 0.46, 0.46, 0.35, 0.52, 0.46, 0.44, 0.35, 0.47, 0.41, 0.42, 0.47, 0]	148,536
SD removal with Chi-square SMOTE Encode	CNN	0.80	[0.786, 0.8007, 0.788]	[0.7852, 0.7998, 0.7871]	0.98	[0.92, 1.00, 0.90, 1.00, 0.95, 0.97, 1.00, 1.00, 0.91, 1.00, 1.00, 0.90, 1.00, 0.99, 0.90]	148,536
Original	RF	0.67	(0.26, 0.19, 0.20)	(0.61, 0.67, 0.63)	0.51	[0.79, 0, 0, 0.53, 0.74, 0.84, 0.23, 0.5, 0.47, 0.79, 0.51, 0.84, 0.72, 0.52, 0]	17,868
Original with standard deviation	RF	0.66	(0.25, 0.19, 0.20)	(0.60, 0.66, 0.62)	0.51	[0.79, 0, 0, 0.52, 0.74, 0.83, 0.24, 0.50, 0.47, 0.81, 0.51, 0.94, 0.70, 0.52, 0]	17,868
SD winsorized with Chi-square	CNN	0.64	[0.28, 0.22, 0.22]	[0.56, 0.64, 0.58]	0.80	[0.80, 0.84, 0.84, 0.75, 0.85, 0.81, 0.73]	17,868

SD winsorized with RFE	RF	0.64	(0.33, 0.24, 0.25)	(0.58, 0.64, 0.59)	0.73	[0.73, 0.82, 0.73, 0.74, 0.69, 0.69, 0.7]	17,868
SD removal with LASSO	CNN	0.63	[0.34, 0.30, 0.29]	[0.57, 0.63, 0.58]	0.79	[0.85, 0.83, 0.75, 0.78, 0.77, 0.75]	7,998
SD removal with Chi-square	CNN	0.63	[0.33, 0.30, 0.29]	[0.57, 0.63, 0.59]	0.79	[0.86, 0.83, 0.74, 0.78, 0.78, 0.74]	7,998
SD winsorized with LASSO	CNN	0.63	[0.27, 0.22, 0.21]	[0.55, 0.63, 0.57]	0.78	[0.79, 0.82, 0.80, 0.73, 0.84, 0.79, 0.73]	17,868
SD removal with LASSO SMOTE En-code	CNN	0.60	[0.59, 0.60, 0.59]	[0.59, 0.60, 0.59]	0.89	[0.96, 0.97, 0.83, 0.90, 0.86, 0.81]	23,234
SD removal with LASSO SMOTE	CNN	0.58	[0.57, 0.58, 0.57]	[0.57, 0.58, 0.57]	0.88	[0.95, 0.96, 0.82, 0.88, 0.85, 0.79]	23,234
SD removal with RFE	CNN	0.57	[0.45, 0.26, 0.25]	[0.54, 0.57, 0.52]	0.76	[0.79, 0.84, 0.74, 0.77, 0.71, 0.73]	7,998
SD winsorized with RFE SMOTE	RF	0.56	(0.55, 0.56, 0.55)	(0.55, 0.56, 0.55)	0.88	(0.82, 0.97, 0.87, 0.88, 0.93, 0.82, 0.84)	70,734
SD removal with Chi-square SMOTE	CNN	0.56	[0.55, 0.56, 0.55]	[0.55, 0.56, 0.55]	0.87	[0.95, 0.95, 0.80, 0.86, 0.84, 0.79]	23,234
SD winsorized with Chi-square SMOTE En-code	CNN	0.55	[0.54, 0.55, 0.55]	[0.54, 0.56, 0.55]	0.88	[0.82, 0.96, 0.91, 0.86, 0.96, 0.84, 0.81]	70,734

SD winsorized with Chi-square SMOTE	CNN	0.53	[0.51, 0.53, 0.52]	[0.51, 0.53, 0.52]	0.87	[0.80, 0.96, 0.90, 0.84, 0.96, 0.82, 0.79]	70,734
SD removal with RFE SMOTE Encode	CNN	0.52	[0.51, 0.52, 0.51]	[0.51, 0.52, 0.51]	0.84	[0.94, 0.92, 0.78, 0.82, 0.83, 0.77]	23,234
SD winsorized with LASSO SMOTE En-code	CNN	0.47	[0.46, 0.47, 0.46]	[0.46, 0.47, 0.46]	0.84	[0.79, 0.96, 0.83, 0.84, 0.90, 0.78, 0.78]	70,734
SD removal with RFE SMOTE	CNN	0.47	[0.46, 0.47, 0.46]	[0.46, 0.47, 0.46]	0.81	[0.92, 0.90, 0.76, 0.79, 0.84, 0.73]	23,234
SD winsorized with LASSO SMOTE	CNN	0.45	[0.43, 0.45, 0.43]	[0.43, 0.45, 0.43]	0.82	[0.77, 0.95, 0.81, 0.82, 0.88, 0.75, 0.77]	70,734

5.4 Discussion

The comparative evaluation of machine learning architectures reveals distinct behavioral patterns that can be explained by the intrinsic nature of the operational data. The superior performance of Random Forest and Extra Trees models in tabular datasets is primarily attributed to their structural suitability for heterogeneous, non-linear, and partially noisy data. These ensemble tree-based methods partition the feature space into highly localized regions, efficiently capturing complex interactions between categorical and numerical variables without requiring extensive normalization or transformation. Furthermore, they naturally handle missing values and non-Gaussian distributions, which are characteristic of real-world railway datasets derived from multiple operational sources.

In contrast, neural network architectures, particularly CNN and RNN, depend on large training volumes and strong sequential correlations to achieve optimal generalization. The CP dataset, although extensive in historical scope, consists of independent disruption records rather than continuous temporal sequences. This limitation restricts the ability of neural networks to exploit temporal dependencies and makes them more susceptible to overfitting in high-dimensional spaces with limited examples per class. Additionally, the need for parameter tuning and normalization increases model sensitivity to data imbalance and scaling differences across features.

From a theoretical standpoint, the Random Forest's ensemble averaging reduces variance and mitigates overfitting by aggregating uncorrelated decision trees, yielding robust predictions even when individual trees perform sub optimally. By comparison, neural networks attempt to approximate a global non-linear mapping, which can be unstable when relationships between variables are sparse or discontinuous. Consequently, tree-based models deliver higher consistency and interpretability in tabular railway data, whereas neural networks become advantageous primarily when continuous temporal or spatial dependencies dominate the problem structure.

These predictive results and modeling insights motivate the prescriptive step taken in Chapter 6, where Transformer-based deep learning predicts full operational task sequences to support actionable replanning.

Chapter 6: Prediction of Rolling Stock Tasks with Transformer

Following the impact-prediction models developed in Chapter 5, this chapter advances from predicting what will happen to prescribing operationally coherent sequences. It introduces a domain-adapted Transformer with specialized semantic tokenization to generate full rolling-stock task sequences under disruption.

6.1 Introduction

The multidimensional predictive models developed in the previous chapter demonstrated robust capability to anticipate aggregated impacts of disruptive events, including the number of affected trains, total delay minutes, and impacted passengers. However, effective management of railway disruptions goes beyond quantifying impacts; it requires the ability to generate complete operational sequences that guide real-time resource rescheduling.

This investigation identifies a fundamental methodological gap: the scientific literature lacks approaches that explore deep learning architectures specifically adapted to predict complete sequences of operational tasks in railway contexts. Existing methods favor mathematical optimization with predefined heuristics or generic machine learning applications without consideration of the structural particularities of the railway domain. As a result, the potential of attention mechanisms to capture complex temporal and spatial dependencies inherent to rolling stock circulation remains unexplored.

This chapter develops the prescriptive dimension of the investigation through the design, implementation, and validation of a Transformer architecture specialized for the railway domain. The main innovation is the semantic tokenization strategy that preserves the hierarchical integrity of operational data, combined with a multidimensional evaluation framework based on statistical plausibility. This approach recognizes that multiple historically valid solutions can coexist for the same disruptive event, overcoming the limitation of optimization paradigms that seek unique solutions.

The empirical validation uses 371,668 operational records from CP (2016-2024), encompassing nine distinct rolling stock series and 135 documented disruption causes. The comparative analysis with LSTM architecture establishes the advantages of attention mechanisms in modeling complex operational patterns, demonstrating 75% reductions in validation error and computational efficiency four times higher.

The chapter is organized into the formal specification of the problem (section 6.2), the detailed presentation of the proposed architectures (sections 6.5 to 6.7), the description of the

experimental methodology (sections 6.8 to 6.9), and concludes with the systematic analysis of results and implications for railway operational management (sections 6.10 to 6.11).

6.2 Research Methodology

The methodology adopted in this study is founded on a comparative experimental approach that evaluates two distinct deep learning architectures for predicting operational task sequences in railway contexts: Transformer encoder-decoder models and Long Short-Term Memory (LSTM) networks. The methodological strategy integrates the analysis of real operational data, systematic hyperparameter optimization, and the development of specialized evaluation metrics for the railway domain.

The dataset used comprises 371,668 operational records of CP rolling stock, covering the temporal period from 2016 to 2024. This database is characterized by 13 operational variables distributed across eight categorical variables and five numerical variables, encompassing temporal dimensions (departure and arrival minutes, delay duration), spatial dimensions (origin and destination stations, geographical location), and operational dimensions (rolling stock series, delay type, disruption cause). The typological diversity of the rolling stock analyzed includes nine distinct series: A2240, A592, A450, L5600, A9630, A4000, L1400, L2600, and A350. These series cover different operational contexts, from regional services to intercity connections.

The proposed Transformer architecture incorporates specific adaptations for the railway domain through a specialized semantic tokenization system. This system uses delimiter tags <T>...</T> to preserve the structural integrity of operational data. The approach maintains the cohesion of operational entities and facilitates structural mapping between input and output spaces. The common vocabulary integrates fundamental special tokens: <PAD> for dimensional normalization of sequences in batch processing, <SOS> as the initial marker for the autoregressive decoding process, <EOS> for termination signaling, and <UNK> for the representation of lexical elements not observed during the training phase.

The Transformer encoder-decoder architecture implements a multi-head self-attention mechanism that processes token sequences through encoding and decoding layers. The encoder uses sinusoidal positional encoding, which is added to the embeddings to capture temporal relations. Meanwhile, the decoder employs masked attention to maintain the autoregressive property during sequence generation. In contrast, the baseline LSTM model implements a sequential encoder-decoder architecture with memory cells. These cells process information through gating mechanisms, specifically input, forget, and output gates, providing long-term information retention capabilities through recurrent structures.

The hyperparameter optimization strategy adopts a systematic three-phase approach. In the exploratory phase, 100 random samples are used to examine the multidimensional parameter space. This phase tests variations in learning rate (0.00005 to 0.0005), embedding dimensions

(128 to 512), number of attention heads (2 to 16), encoding and decoding layers (2 to 6), dropout (0.05 to 0.3), and batch size (8 to 128). The intermediate phase develops statistical analyses that incorporate stability metrics, computational efficiency, and learning speed. This phase identifies significant correlations between the number of attention heads and training stability. The final phase concentrates on exploring the most promising regions, culminating in an optimal configuration. This configuration includes a learning rate of 0.0003, embedding dimension of 512, eight attention heads, three encoder-decoder layers, dropout of 0.3, and batch size of 128.

The training process implements an 80/20 split for training and validation with a fixed seed (42) to ensure comparability between architectures. The Transformer model uses the AdamW optimizer with decoupled L2 regularization, while the LSTM model employs Adam with gradient clipping to prevent gradient explosion. Both models apply early stopping based on relative stagnation metrics and minimum improvement thresholds to prevent overfitting. The cross-entropy loss function is used during training and is computed over token sequences with appropriate masking for padding tokens.

Model evaluation implements a multidimensional statistical plausibility metric that recognizes the inherent ambiguity in predicting operational sequences. In this context, multiple historically valid solutions may exist. This metric comprises six plausibility components: station sequence, point-to-point transitions, specific positional values, temporal intervals, travel durations, and global route patterns. The methodology incorporates statistical robustness metrics, including Kullback-Leibler divergence between component distributions, evaluation of value distribution perplexity, and statistical plausibility analysis through z-scores and percentiles contextualized by incident cause.

Experimental procedures are organized into test sets structured according to the “X-Y” convention. In this convention, “X” represents the rolling stock series and “Y” indicates the target number of samples per disruption cause. This structuring allows systematic comparative analysis across different operational contexts and data representativeness levels. It facilitates the identification of consistent performance patterns and key predictability factors by disruptive event type.

Figure 14 illustrates the sequence diagram for the practical implementation of Transformer and LSTM models, dividing the process into two main phases: training and inference. The top section represents model preparation and learning from operational data, while the bottom section shows the phases where the trained models are applied to predict new task sequences and their corresponding evaluation.

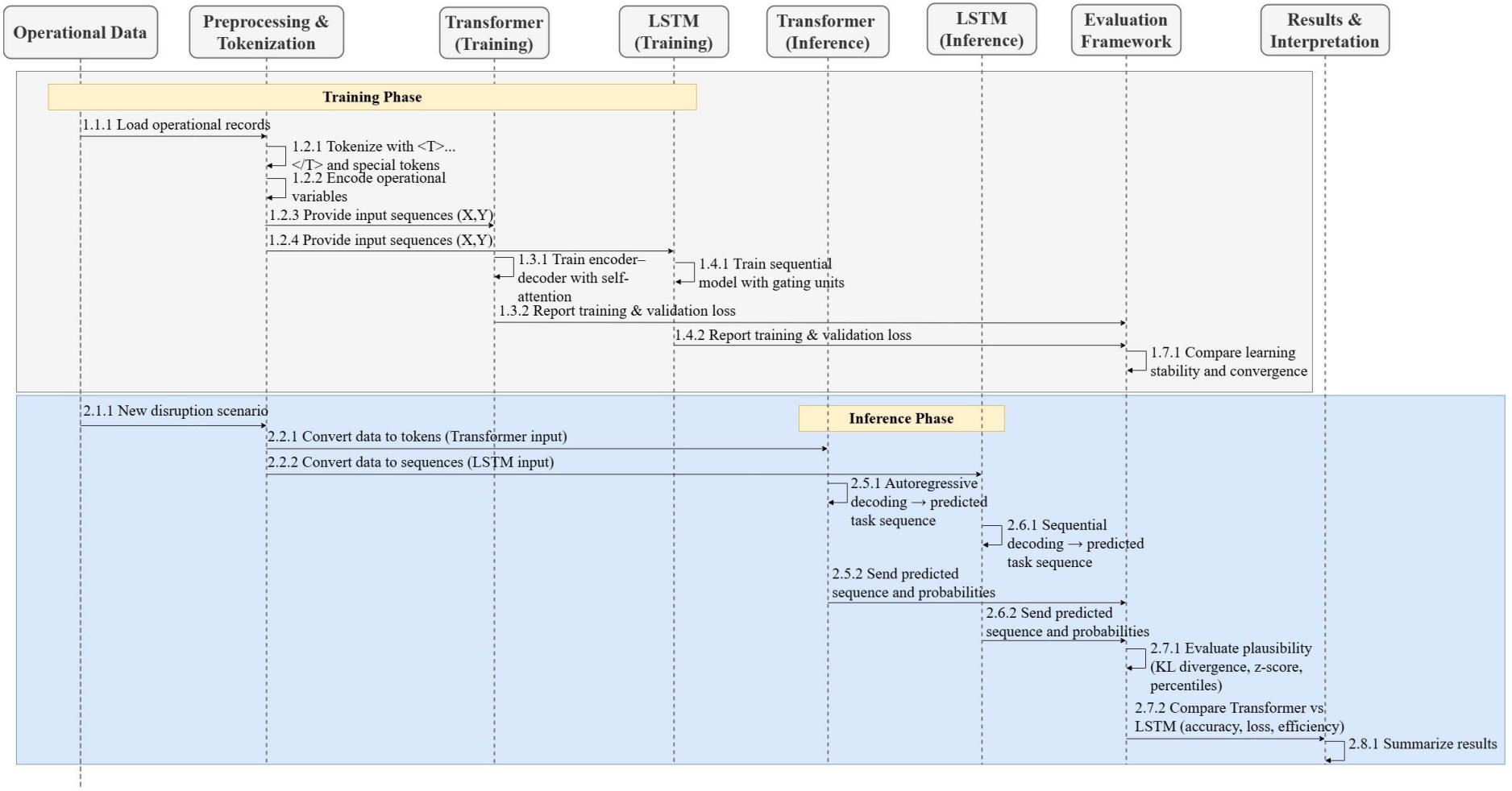

Figure 14: Sequence diagram of the practical implementation of Transformer and LSTM models

Table 12 complements Figure 14 by providing a detailed presentation of the flows represented in the diagram and describing the function and results of each step of the process.

Table 12: Description of the stages and processes of the workflow shown in Figure 14

Section	Process / Sub-step	Purpose and Description	Interconnections / Output
Operational Data	1.1.1 Load operational records	Imports real operational data from CP, containing disruptions, delays, and task attributes.	Supplies base dataset for preprocessing.
Preprocessing & Tokenization	1.2.1 Tokenize with <T>…</T> and special tokens	Creates structured symbolic representations preserving task hierarchy.	Output tokens feed Transformer encoder.
	1.2.2 Encode operational variables	Normalizes categorical and numerical variables.	Encoded features feed both models.
	1.2.3–1.2.4 Provide input sequences (X,Y)	Generates aligned input–target pairs for supervised training.	Feeds training modules.
Transformer (Training)	1.3.1 Train encoder–decoder with self-attention	Learns temporal–spatial dependencies among operational tokens.	Produces trained Transformer model.
	1.3.2 Report training & validation loss	Records convergence metrics.	Input to evaluation framework.
LSTM (Training)	1.4.1 Train sequential model with gating units	Learns task transitions using recurrent memory cells.	Produces trained LSTM model.
	1.4.2 Report training & validation loss	Provides comparative baseline to Transformer.	Input to evaluation framework.
Evaluation Framework	1.7.1 Compare learning stability and convergence	Quantifies efficiency and loss behavior between both models.	Concludes training phase.

Results & Interpretation	—	Consolidates training outcomes.	Forms reference for inference comparison.
Operational Data (new scenario)	2.1.1 New disruption scenario	Introduces unseen operational conditions for testing.	Initiates inference phase.
Preprocessing & Tokenization (Inference)	2.2.1 Convert data to tokens (Transformer input) / 2.2.2 Convert data to sequences (LSTM input)	Transforms new operational records into model-ready format.	Inputs to inference models.
Transformer (Inference)	2.5.1 Autoregressive decoding → predicted task sequence	Generates predicted operational task chain.	Sends results for evaluation.
LSTM (Inference)	2.6.1 Sequential decoding → predicted task sequence	Produces equivalent sequence via recurrent prediction.	Sends results for evaluation.
Evaluation Framework (Inference)	2.7.1 Evaluate plausibility (KL-divergence, z-score, percentiles)	Measures statistical consistency between predicted and historical sequences.	Produce plausibility metrics.
	2.7.2 Compare Transformer vs LSTM (accuracy, loss, efficiency)	Quantitatively contrasts both architectures.	Basis for result synthesis.
Results & Interpretation (Final)	2.8.1 Summarize results	Integrates findings from both phases and highlights model advantages.	Final conclusions for Chapter 6.

6.3 Problem Definition

Contemporary railway systems frequently experience operational disruptions that compromise service efficiency and passenger satisfaction (Zhang et al., 2022). These disruptions include rolling stock failures, infrastructure faults, adverse weather conditions, and external interferences. They create cascading effects that propagate through the railway network, resulting in delays, cancellations, and resource reallocations (Tiong et al., 2023). Effective disruption management requires rapid and coordinated decisions regarding timetable rescheduling and rolling stock redistribution. This represents a dynamic and multi-dimensional optimization problem with high

computational and temporal complexity (Su et al., 2024). Traditional disruption management methods, based on manual heuristics or mathematical optimization algorithms, show significant limitations when applied to large-scale networks or scenarios requiring real-time response (Veelenturf et al., 2017).

6.3.1 Proposed Approach

The approach employs Transformer and LSTM architectures to learn operational reconfiguration patterns directly from historical data, eliminating the need for predefined heuristics or artificial decompositions of the problem. The mathematical formulation of the problem can be expressed as:

$$\theta^{*} = \arg\max_{\theta} \sum_{n=1}^{N} \log P_{\theta}\left(S^{(n)} \mid H^{(n)}\right) \tag{4}$$

where θ^{*} represents the optimal model parameters, N corresponds to the number of historical observations, and $P_{\theta}(S|H)$ denotes the conditional probability of the future task sequence S given the history H. During the inference phase, the model identifies the most likely sequence through:

$$\arg\max_{S} \; P_{\theta *}(S \mid H) \tag{5}$$

This approach allows the automatic capture of complex dependencies between operational variables, facilitating the generation of rescheduling solutions that are simultaneously efficient and compatible with historically observed operational patterns.

6.3.2 Technical Problem Specification

The problem of rolling stock allocation during disruptions can be formally characterized as a sequential prediction problem under uncertainty. Given a set of historically executed operational tasks $H = (h_{1}, \ldots, h_{T})$ and a specific disruptive event, the objective is to predict a sequence of future tasks $S = (s_{1}, \ldots, s_{T})$ that is both operationally feasible and statistically plausible when compared to the observed historical patterns. Figure 15 illustrates the problem of task allocation to railway rolling stock.

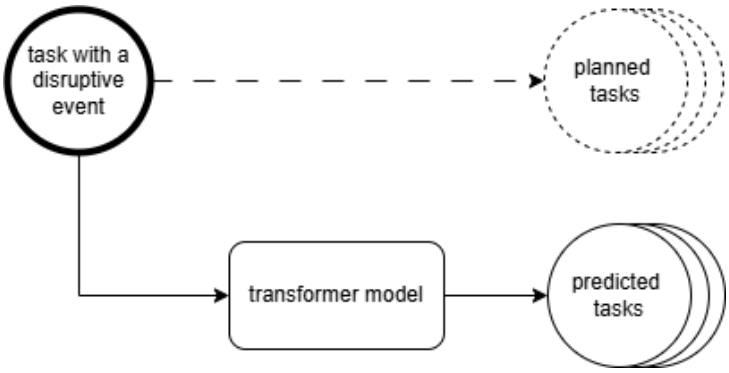

Figure 15: High-level flow of the proposed model's employability

This problem presents multiple dimensions of complexity:

- Temporal Dimension: Reallocation decisions must consider temporal dependencies between consecutive tasks, including transition times between stations, service durations, and mandatory maintenance windows.
- Spatial Dimension: The geographical distribution of disruptions and available resources influences reallocation options, requiring simultaneous consideration of multiple lines and nodes of the railway network.
- Stochastic Dimension: The random nature of disruptions and the variability in event duration introduce uncertainty in predictions, requiring approaches that incorporate robustness in the face of multiple scenarios.

6.3.3 Limitations of Existing Methods

Conventional disruption management methods have structural limitations that necessitate the exploration of alternative approaches. Exact optimization algorithms, while theoretically capable of identifying optimal solutions, face scalability problems when applied to real-scale railway networks (Zhong et al., 2019). Heuristic approaches, by contrast, rely on predefined rules that may fail to capture the complexity of real operational patterns (Liu et al., 2024).

Furthermore, existing methods often adopt a fragmented perspective of the problem, optimizing individual components (timetables, rolling stock) sequentially rather than through an integrated solution (Zhan et al., 2024). This approach can yield suboptimal solutions that do not adequately leverage the synergies between different dimensions of the operational problem.

6.4 Architecture of the Proposed Models

The approach involves a comparative analysis of two distinct architectures for task sequence prediction: a Transformer encoder-decoder model and an LSTM-based model. Both architectures are evaluated to assess their effectiveness in predicting sequential

tasks within the railway context. Specific customizations are implemented to address the unique challenges of rolling stock management during disruption scenarios.

6.4.1 Common Characteristics

Both models incorporate a common vocabulary to organize the task sequence, which is described in Table 13. The tokenization uses four fundamental special tokens to enable the sequential mechanism: <PAD> for padding sequences to uniform length in batch processing, <SOS> (Start of Sequence) as the initial marker for autoregressive decoding, <EOS> (End of Sequence) to signal sequence termination, and <UNK> (Unknown) to represent tokens not seen during training. The key feature of the proposed tokenization system is its ability to preserve the semantic structure of data through delimiter tags <T>...</T>. This format encapsulates discrete units of operational information, offering two main advantages: it maintains the integrity of operational entities and simplifies structural mapping between input and output spaces.

Table 13: Table of tags and descriptions

Tag	Description
<PAD>	For dimensional normalization of sequences in batch processing
<SOS>	Start of Sequence - initial marker for the autoregressive decoding process
<EOS>	End of Sequence - for termination signaling
<UNK>	Unknown - to represent lexical elements not observed during the training phase
<T>...</T>	Delimiter tags that encapsulate atomic units of operational information, preserving the semantic structure of data

Figure 16 illustrates the components of both proposed models, whose details are described in the following subsections.

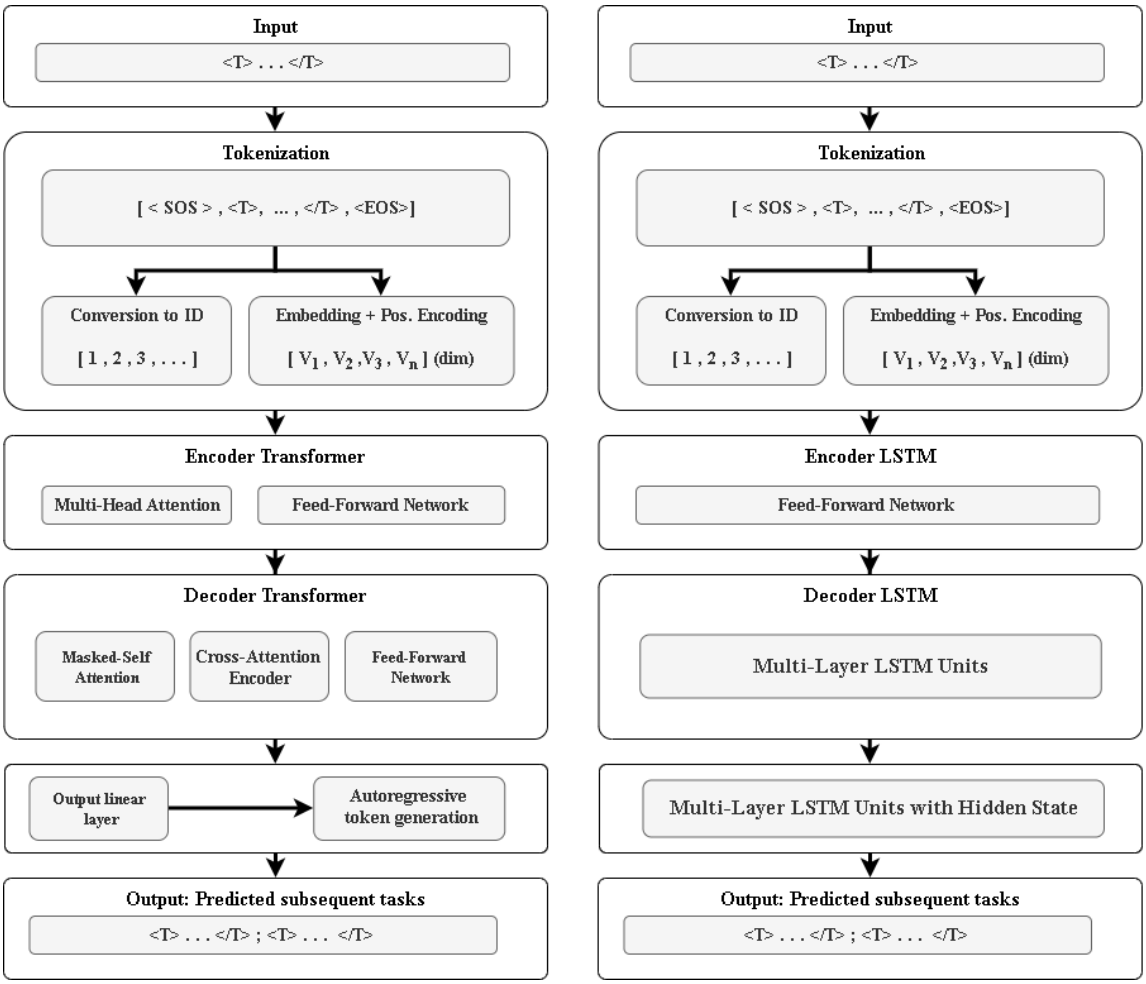

Figure 16: High-level architecture of the Transformer model (left) and the LSTM model (right)

The proposed Transformer architecture differs significantly from traditional implementations through targeted adaptations developed for the railway operational domain. Its main innovation resides in a specialized semantic tokenization strategy that employs delimiter tags <T>...</T> to preserve the structural integrity of operational data, maintaining the hierarchical relationship between railway tasks. This approach is reinforced by an embedding system carefully calibrated to capture the temporal and spatial subtleties of railway operations, thereby improving the capacity of the self-attention mechanism to detect disruption propagation patterns throughout the network.

Contrary to conventional Transformer models used in natural language processing or generic time series analysis, our architecture is explicitly optimized to identify and prioritize critical operational dependencies between sequential rolling stock tasks. Consequently, the model is better adapted to respond to the complex demands of railway management in disruption scenarios.

6.5 Transformer Architecture

The following subsections formally describe the Transformer architecture developed for the railway domain by decomposing it into fundamental components. Building upon the common

characteristics established in section 6.3.1, we present the mathematical formalization that specifies the sequential processing from tokenization and embedding to the autoregressive generation of task sequences. This formal description allows us to understand the mechanisms by which the model captures the complex operational dependencies that characterize rolling stock management in disruptive scenarios.

6.5.1 Input

Let X be a sequence of characteristics representing a past task:
The input consists of a sequence of tokens $X = (x_1, x_2, \dots, x_{13})$, where each x_i is an index in the vocabulary V.

6.5.2 Embedding and Positional Encoding

The embedding layer transforms the indices into dense vectors:

$$E(X) = \left(E_{x_1}, E_{x_2}, \dots, E_{x_n}\right) \in \mathbb{R}^{n \times d} \tag{6}$$

Where $E \in \mathbb{R}^{|V| \times d}$ is the embedding matrix and d is the embedding dimension.
The positional encoding is added to the embeddings:

$$X^0 = E(X) + PE \tag{7}$$

Where $PE \in \mathbb{R}^{n \times d}$ is the positional encoding matrix with elements:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i}/d_{model}}\right) \tag{8}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i}/d_{model}}\right) \tag{9}$$

6.5.3 Encoder

The encoder consists of N_e identical layers. For each layer $l \in [1, N_e]$:

$$Z^l = X^{l-1} + \text{MultiHead}(\text{LayerNorm}(X^{l-1}), \text{LayerNorm}(X^{l-1}), \text{LayerNorm}(X^{l-1})) \tag{10}$$

$$X^l = Z^l + \mathrm{FFN}(\mathrm{LayerNorm}(Z^l)) \tag{11}$$

Where multi-head attention is defined as:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O \tag{12}$$

$$\mathrm{head}_i = \mathrm{Attention}\,(QW_i^Q, KW_i^K, VW_i^V) \tag{13}$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

And the feed-forward network is defined as:

$$\mathrm{FFM}(x) = \max(0, xW_1 + b_1)\,W_2 + b_2 \tag{15}$$

The final output of the encoder is $M = X^{N_e}$.

6.5.4 Decoder

The decoder consists of N_d layers. For an output sequence $Y = (y_1, \ldots, y_m)$, we start with:

$$Y^0 = E(Y) + PE \tag{16}$$

For each layer $l \in [1, N_d]$:

$$A^l = Y^{l-1} + \mathrm{MaskedMultiHead}(\mathrm{LayerNorm}(Y^{l-1})) \tag{17}$$

$$B^l = A^l + \mathrm{MultiHead}(\mathrm{LayerNorm}(A^l), M, M) \tag{18}$$

$$Y^l = B^l + \mathrm{FFN}\left(\mathrm{LayerNorm}(B^l)\right) \tag{19}$$

Where MaskedMultiHead uses a look-ahead mask:

$$\mathrm{mask}(i, j) = \begin{cases} 1, & \text{if } j > i, \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

The mask is applied within the attention function:

$$\text{MaskedAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \text{mask} \cdot (-\infty)\right) V \tag{21}$$

The final output of the decoder is $K = Y^{N_d}$.

6.5.5 Linear Layer

The linear layer projects the vectors into the vocabulary space:

$$O = HW_O + b_O \in \mathbb{R}^{m \times |V|} \tag{22}$$

Where $W_O \in \mathbb{R}^{d \times |V|}, b_O \in \mathbb{R}^{|V|}$.

6.5.6 Output

The probability distribution for each position is:

$$P(y_t | y_{<t}, X) = \text{softmax}(O_t) \in \mathbb{R}^{|V|} \tag{23}$$

During training, we compute the cross-entropy loss:

$$\mathcal{L} = -\sum_{t=1}^{m} \sum_{v=1}^{|V|} 1(y_t \neq \text{pad}) \cdot y_{t,v} \log(P(y_t = v | y_{<t}, X)) \tag{24}$$

6.6 LSTM Architecture

The following subsections formally describe the LSTM architecture developed for the railway domain by decomposing its fundamental components. Building upon the common characteristics established in Section 6.4.1, the mathematical formalization specifying the sequential processing from tokenization and embedding to the autoregressive generation of task sequences is presented. This formal description enables an understanding of the mechanisms through which the model captures temporal operational dependencies via recurrent memory structures, in contrast to the

parallel attention-based approach of the Transformer applied to rolling stock management under disruption scenarios.

6.6.1 Input

Like the Transformer, the input to the LSTM consists of a sequence of tokens representing characteristics of past tasks:

The input is a sequence of tokens $X = (x_1, x_2, \dots, x_{13})$, where each x_i is an index in the vocabulary V.

6.6.2 Embedding

The LSTM embedding module transforms the indices into fixed-dimension characteristic vectors:

$$E(X) = \left(E_{x_1}, E_{x_2}, \dots, E_{x_n}\right) \in \mathbb{R}^{n \times d} \tag{25}$$

Where $E \in \mathbb{R}^{|V| \times d}$ is the embedding matrix. Unlike the Transformer, the LSTM model does not require explicit positional encoding, as the sequential nature of LSTM processing implicitly captures positional relationships.

6.6.3 LSTM Encoder

The LSTM encoder processes the input embeddings sequentially, updating its hidden state:

$$h_t, c_t = \text{LSTM}(E_{xt}, h_{t-1}, c_{t-1}) \tag{26}$$

Forget gate:

$$f_t = \sigma(W^f[h_{t-1}, E_{xt}] + b^f) \tag{27}$$

Input gate:

$$i_t = \sigma(W^I[h_{t-1}, E_{xt}] + b^I) \tag{28}$$

Cell candidate:

$$\tilde{c}_t = \tanh(W^c[h_{t-1}, E_{xt}] + b^c) \tag{29}$$

Output gate:

$$o_t = \sigma(W^o[h_t^{-1}, E_{xt}] + b^o) \tag{30}$$

Cell state update:

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (31)$$

Hidden state:

$$h_t = o_t * \tanh(c_t) \quad (32)$$

The LSTM encoder can be stacked in multiple layers to increase the model's capacity. In this case, the output of each LSTM layer serves as input for the subsequent layer.

6.6.4 LSTM Decoder

The LSTM decoder operates in an autoregressive manner, generating one token at a time. Initially, the decoder's hidden state is initialized with the final state of the encoder:

$$h^{d0} = h^e{}_n \quad (33)$$
$$c^d{}_0 = c^e{}_n$$

At each time step t, the decoder receives the embedding of the previously generated token (or the start token <SOS> at the first step) and produces a probability distribution over the next token:

$$h^d{}_t, c^d{}_t = \text{LSTM}(E_{yt}{}^{-1}, h^d{}_t{}^{-1}, c^d{}_t{}^{-1}) \quad (34)$$

6.6.5 Output Layer and Generation

The output of the LSTM decoder is processed by a linear layer followed by a softmax function to produce a probability distribution over the next token:

$$z_t = W^N h^d{}_t + b^N \quad (35)$$
$$P(y_t | y <_t, X) = \text{softmax}(z_t)$$

6.7 Comparison Between Architectures

Transformers and LSTM architectures present distinct characteristics, each with advantages and disadvantages well-documented in scientific literature. Transformers are inherently designed for parallelization owing to the absence of recurrent units, offering significant performance benefits

over sequential architecture such as LSTM (Vaswani et al., 2017). This parallel processing capability can be efficiently leveraged on specialized hardware like Graphics Processing Unit (GPU), whereas LSTMs, due to their sequential structure, face limitations in parallelization (Shazeer et al., 2017). Nevertheless, computational performance also depends on factors such as model complexity, batch size, implementation efficiency, and hardware specifications.

For dependency capture, Transformers employ a self-attention mechanism that captures both fine-grained and high-level contexts, while LSTM are particularly effective in modeling long-term dependencies through their memory units (Cao et al., 2024). This distinction proves crucial in sequential prediction tasks, such as those examined in this study, where Transformers outperformed LSTM by 1-25% for long time series, demonstrating superior capacity to link past and future events (Pölz et al., 2024).

In terms of long-range memory, Transformer architectures, through self-attention and positional encoding, offer equal or superior performance to RNN and LSTM variants while being more computationally efficient and avoiding the limitations of these models. LSTMs may experience information degradation with very long sequences, whereas Transformers maintain effectiveness even when handling distant dependencies (Su et al., 2023).

Transformers also address a central limitation of RNNs and LSTM concerning sequence parallelization. Recurrent architectures process elements sequentially, introducing temporal dependencies, while Transformers use self-attention to directly access any position in a sequence, requiring a constant number of operations regardless of sequence length (Vaswani et al., 2017).

6.8 Experimental Configuration

This section describes the experimental configuration used to evaluate the performance of the proposed models in the prediction of railway operational sequences.

6.8.1 Dataset

This study analyzed 371,668 operational records of railway rolling stock tasks. Table 14 presents 13 variables that describe characteristics of tasks and disruptive events, categorized into eight categorical types and five numerical types. These variables encompass temporal, spatial, and operational dimensions in the railway domain.

Table 14: Dataset variables

#	Variable	Description	Data Type	Data. Range / Values	Unique Values

1	Rolling Stock Series	Type of railway rolling stock	Categorical	A2240, A592, A450, L5600, A9630, A4000, L1400, L2600, A350	9
2	Day of the Week	Day of the week in which the task occurs	Categorical	Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday	7
3	Departure Minute of Day	Minute of the day at which the task starts	Numerical	0–1440	1285
4	Arrival. Minute of Day	Minute of the day at which the task ends	Numerical	0–1440	1375
5	Departure Location	Departure station	Categorical	Portuguese railway network stations	284
6	Arrival Location	Arrival station	Categorical	Portuguese railway network stations	284
7	Disruption Start Minute	Minute of the day when the disruptive event began	Numerical	0–1440	1424
8	Geographic Region	Region corresponding to the railway line	Categorical	List of regions	40
9	Disruption Location	Initial location of the disruption on the railway network	Categorical	Portuguese railway network stations	629
10	Cause	Reason assigned by the railway operator for the disruption	Categorical	Operational causes defined by the operator CP	135

11	Delay Type	Whether the delay is primary (originates in the task) or secondary (caused by another task)	Categorical	Primary (Pri) and Secondary (Sec)	2
12	Delay Duration (minutes)	Duration of the delay	Numerical	1–894	326
13	Number of Trains Involved	Number of other tasks affected by the same disruptive event	Numerical	1–10,824	291

6.8.2 Hyperparameter Selection

A three-phase hyperparameter optimization strategy was implemented that balances exploration and refinement. In the exploration phase, 100 random samples were used to examine the multidimensional space, testing parameters such as learning rate (0.00005-0.0005), embedding dimensions (128-512), attention heads (2-16), encoding/decoding layers (2-6), dropout (0.05-0.3), and batch size (8-128).

In the intermediate phase, statistical analyses were developed beyond validation loss, incorporating metrics for stability, computational efficiency, and learning speed. Significant correlations were identified between the number of attention heads and training stability.

In the final phase, exploration was concentrated on the most promising regions previously identified. The optimal configuration achieved a validation loss of 0.5519 with a learning rate of 0.0003, an embedding dimension of 512, eight attention heads, three encoder–decoder layers, a dropout rate of 0.3, and a batch size of 128.

The experiments partially illustrated in Figure 17 revealed an optimal configuration with a learning rate of 0.0003 and an embedding dimension of 512. They also demonstrate eight attention heads, three layers in the encoder, three layers in the decoder, a dropout rate of 0.3, and a batch size of 128, achieving a validation loss value of 0.183, significantly lower than the average of the tested configurations.

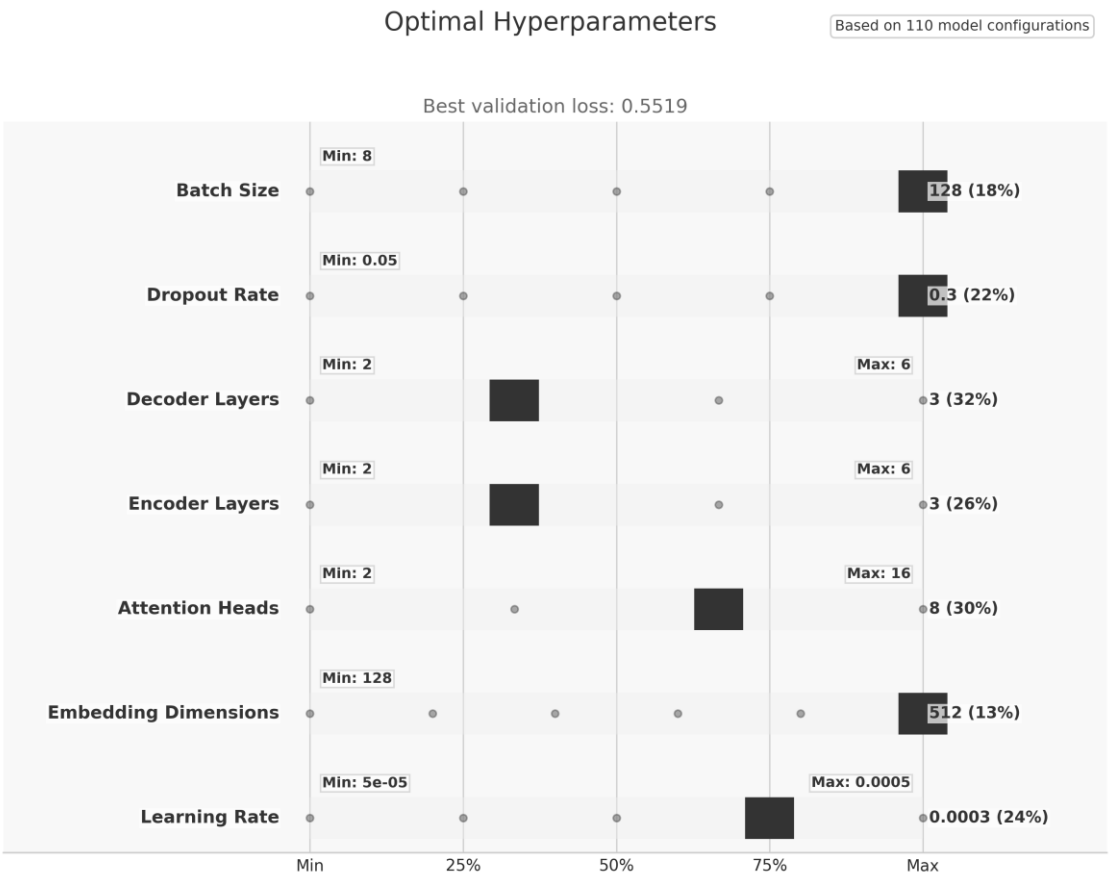

Figure 17: Hyperparameter optimization results

6.9 Model Training

In the input phase, both Transformer and LSTM models process specialized data containing specific task information delimited by structural tags <T>...</T>. This process involves advanced lexical analysis through regular expressions, allowing precise content extraction while preserving semantic integrity. A hierarchical vocabulary management system, which recognizes special tokens and padding elements for dimensional normalization, ensures architectural consistency in both models.

In the embedding layer, token indices are projected into continuous vector spaces through linear transformations. The Transformer incorporates sinusoidal positional encoding to address the inherent positional invariance of attention networks (Vaswani et al., 2017). In contrast, the recurrent LSTM architecture naturally retains sequential order, eliminating the need for explicit positional encoding, although with limited capacity to model long-range dependencies (Graves & Schmidhuber, 2005).

The Transformer encoder employs multi-layer processing with parallelized self-attention using normalized scalar products. Conversely, the LSTM encoder processes data sequentially through memory cells and gating mechanisms (input, forget, and output), allowing long-term

retention and selective forgetting of information (Gers et al., 1999). Although sequential nature improves stability, it restricts parallelism and increases computation time.

Layer normalization in the Transformer uses local statistics with trainable parameters for adaptive invariance. The LSTM employs recurrent normalization in hidden states to mitigate the vanishing gradient problem, a known limitation of RNNs (Graves & Schmidhuber, 2005). Feed-forward networks in the Transformer and linear projections in the LSTM model nonlinear relationships between sequence elements through distinct mechanisms (Vaswani et al., 2017).

For optimization, the Transformer uses AdamW with decoupled L2 regularization (Liu et al., 2021), while the LSTM typically applies Adam with gradient clipping to prevent explosion, a common issue in deep recurrent networks (Srivastava et al., 2015). Both models employ an 80/20 training-validation split with fixed seed (42) to ensure comparability.

The Transformer decoder uses causal attention through upper triangular masking and cross-attention for contextual conditioning. The LSTM decoder follows an autoregressive approach where each output depends on the sequentially updated internal state. Both models conclude with a linear transformation followed by softmax normalization to produce token probabilities. During inference, the Transformer leverages incremental caching of representations, while the LSTM updates its internal state step by step (Graves & Schmidhuber, 2005).

Early stopping based on a relative stagnation metric and minimum improvement threshold is implemented in both models to prevent premature termination and underfitting. Regularization is implicit in the Transformer and explicit in the LSTM (e.g., dropout in recurrent connections) (Hewamalage et al., 2021).

LSTM hyperparameters are based on established neural network literature. A learning rate of 0.001 balances convergence speed and stability. The embedding dimension of 256 aligns with findings demonstrating that values between 100-300 optimize semantic richness and efficiency. A hidden dimension of 512 is based on original LSTM recommendations (Beck et al., 2024), aiming to capture complex sequential patterns without attention mechanisms. Two layers were chosen to avoid overfitting or underfitting, and a dropout rate of 0.3 reflects empirical evidence supporting rates between 0.2 and 0.5 for effective regularization.

Experimental results in Table 15 highlight substantial performance differences. Transformer Test 3 achieved the lowest validation loss (0.240), a 75% improvement over the best LSTM result (0.9471 in Test 1). Despite similar parameter counts (10.05M vs. 9.71M), the Transformer was more than four times faster per epoch (461.74s vs. 1829.32s). On average, Transformer models exhibited 35.3% lower training loss and 72.4% lower validation loss than LSTM models, indicating consistent superiority. Notably, Transformer Test 1 converged in the first epoch, while others required more prolonged training, underlining the architecture's sensitivity to hyperparameter configurations.

Table 15: Parameterization of the training performed

Parameter/Metric	Transformer			LSTM		
	Test 1	Test 2	Test 3	Test 1	Test 2	Test 3
Batch Size	128	128	128	128	128	128
Learning Rate	0.0003	0.0001	0.000075	0.001	0.0005	0.00075
Embedding Dim	512	256	384	256	256	256
Num. Heads	8	4	4	0	0	0
Num. Encoder Layers	3	2	2	2	2	2
Num. Decoder Layers	3	2	2	2	2	2
Dropout	0.3	0.5	0.4	0.3	0.5	0.4
Max Seq. Length	512	512	512	512	512	512
Epochs Completed	3	20	20	15	9	7
Final Training Loss	0.231	0.264	0.221	0.3140	0.3988	0.3939
Best Validation Loss	0.363	0.285	0.240	0.9471	1.2034	1.0706
Best Epoch	1	20	19	13	7	7
Average Time per Epoch (sec)	938.81	284.17	461.74	1829.32	1943.75	1834.55
Vocabulary	2305	2305	2305	2299	2299	2299

Total Parameters	2.4E+07	4868865	1E+07	9712891	9712891	9712891

In Figure 18, the Transformer model (left graph) demonstrates faster and more stable convergence in both training and validation. Across all learning rate configurations (0.0001, 0.0005, 0.00075) and dropout rates (0.3, 0.5, 0.4), the loss function decreases sharply in the first five epochs, stabilizing below 0.25 to 0.30 after the tenth epoch. Similarly, validation loss decreases rapidly and stabilizes between 0.30 and 0.35, indicating strong generalization, minimal overfitting, and robustness to hyperparameter variations. In contrast, the LSTM model (right graph) converges more slowly and exhibits a pronounced gap between training loss and validation loss. Training loss starts above 1.5 and decreases to approximately 0.30 after 8 to 10 epochs, while validation loss remains above 1.0 even after 15 epochs, suggesting a significant generalization gap. Although the LSTM training loss continues to decrease, its performance indicates limited generalization capacity, likely stemming from architectural constraints of its sequential nature.

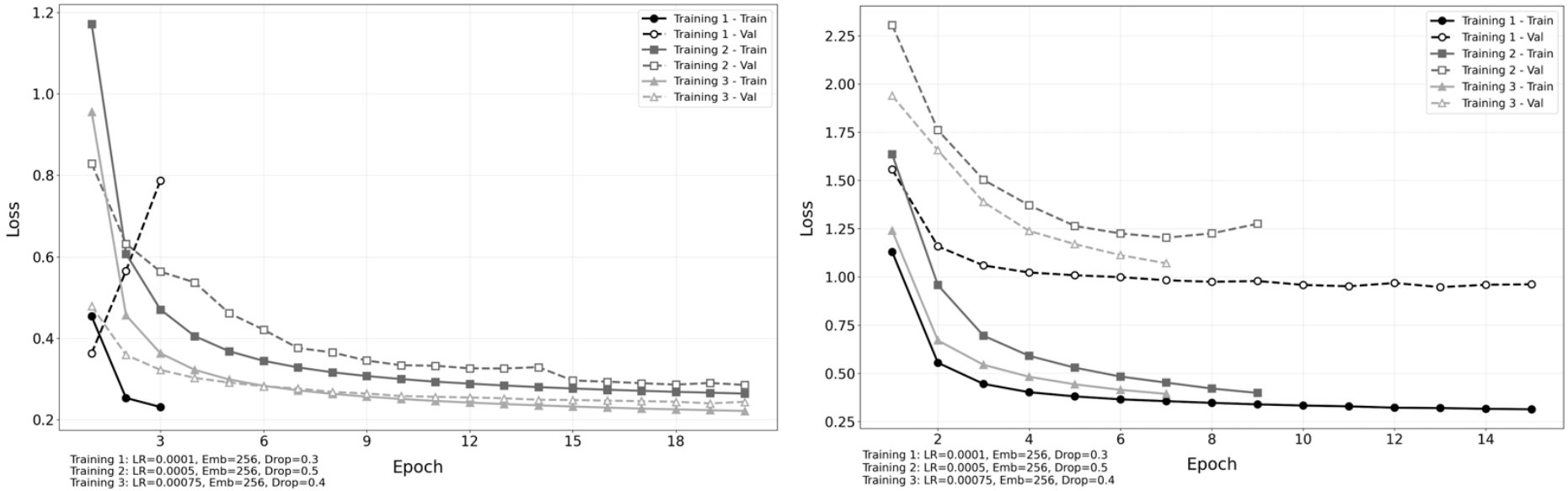

Figure 18: Training and validation loss: Transformer (left) vs. LSTM (right)

Figure 19 compares the performance of Transformer and LSTM models across multiple training sessions. In the first experiment with the Transformer (epoch 1), the training loss was relatively low (0.36); however, the validation loss increased to 0.79, indicating early overfitting. At epoch 20, training and validation losses converged to 0.285, approaching the optimal threshold of 0.2397 that the dashed line indicates. The best performance occurred at epoch 19, with training and validation losses of 0.24 and 0.2437, respectively. This result reflects optimal convergence with minimal deviation from the dataset characteristics.

In contrast, the right panel presents the LSTM results. In the first session (epoch 13), training and validation losses were 0.9471 and 0.9617, respectively, with the training loss matching the optimal threshold (0.9471) and the validation loss slightly exceeding it. Performance deteriorated in subsequent runs: at epoch 7, training and validation losses increased to 1.20 and 1.28, and in a later session, both stabilized at 1.0706. These results demonstrate that, unlike the Transformer, the LSTM model does not show improvement with additional training and remains distant from optimal performance.

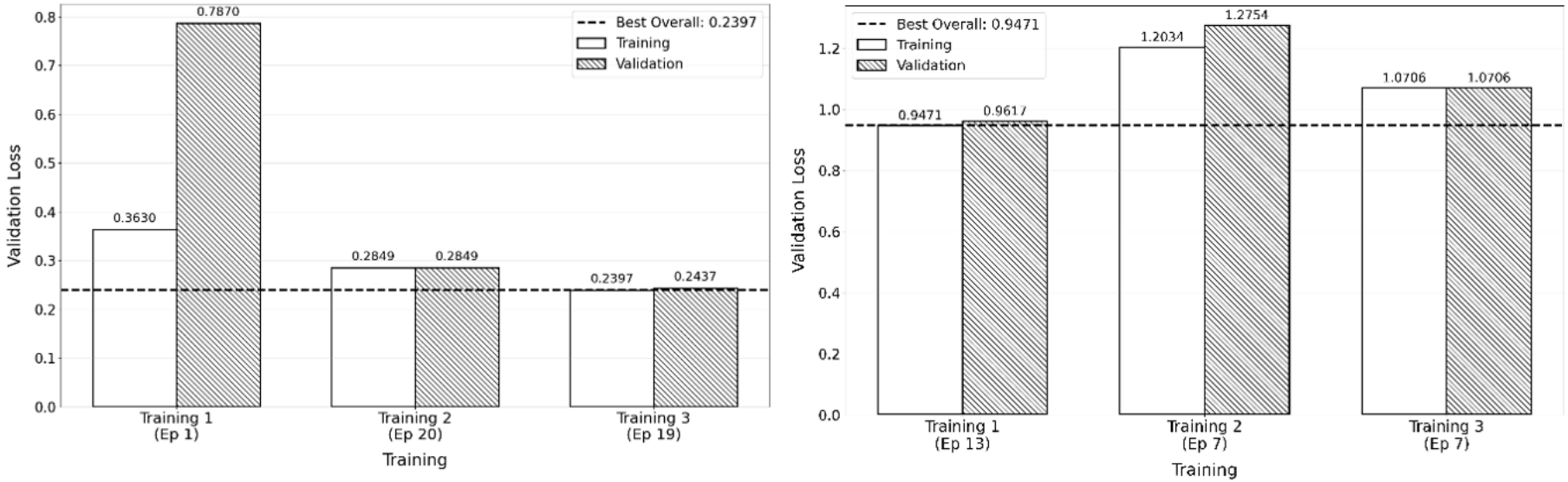

Figure 19: Progression of validation loss of Transformer and LSTM models

Both models were evaluated using a multicomponent statistical plausibility metric that recognizes the inherent ambiguity in predicting operational sequences, where multiple historically valid solutions may coexist. The metric encompassed six plausibility dimensions: station sequence, point-to-point transitions, specific positional values, temporal intervals, travel durations, and global route patterns. This multidimensional framework enabled an objective comparison of prediction quality while considering the unique generative characteristics of each architecture. Figure 20 provides a detailed illustration of the construction process for the statistical plausibility metric.

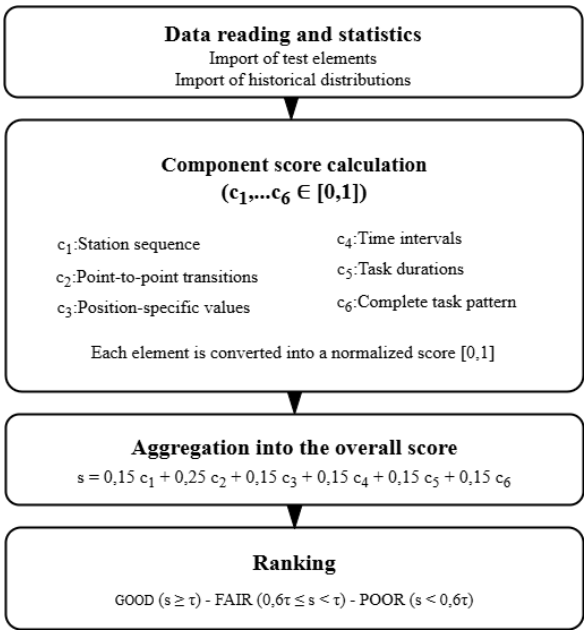

Figure 20: Statistical plausibility evaluation flow

The methodology incorporates statistical robustness metrics to enhance the analysis, including Kullback-Leibler divergence between component distributions (Grewal, 2011). This procedure evaluates the perplexity of value distribution and examines statistical plausibility through z-scores and percentile scores in the context of incident cause. Additionally, the analysis considers predicted frequency, expected frequency, and relative scores.

6.10 Results

Table 16 presents a detailed comparison of Transformer and LSTM models across multiple test sets, each corresponding to a specific rolling stock series and case record count. The “X-Y” naming convention identifies each test set, where “X” represents the rolling stock series (e.g., A9630, A350) and “Y” indicates the target number of samples per disruption cause (1, 2, 4, 8, 16, or 32). For example, A9630-1 contains one record for each distinct cause, while A9630-32 is intended to

include 32 examples for each reason. Some sets show slightly fewer records than the theoretical maximum (causes × samples), indicating causes with limited occurrences in the original dataset.

The results highlight the superior performance of the Transformer, as demonstrated by rigorous quantitative metrics. Notably, the Transformer achieves 100% classification accuracy in sets A9630-1, A9630-32, and A9630-8, compared with 90%, 89%, and 90% for the LSTM in the same sets. This clear disparity underscores the Transformer's enhanced capacity to discriminate complex data patterns.

Additional support for the Transformer's robustness comes from analysis of relative means, a normalized metric that quantifies predictive efficiency through the ratio between predicted mean and expected mean. This measure provides crucial insight into model calibration, where values close to 1.0 indicate optimal correspondence between predictions and real observations. In the highlighted sets, the Transformer registers values of 1.17, 1.36, and 1.31, while the LSTM produces 1.43, 1.42, and 1.27, indicating greater precision and reduced prediction variance. This stability persists despite significant variation in test set sizes, from 67 to 1,277 records in the examples, and up to 3,459 in the complete set.

The Transformer also demonstrates a more consistent predicted mean of approximately 0.50 across the three sets, compared with the LSTM's 0.38. The predicted mean represents the average value of predictions generated by the machine learning model for each test set, constituting a fundamental indicator of the central tendency of predictions produced by the algorithm. This consistency in the Transformer contrasts with the expected mean, which corresponds to the average value observed in the real or historical data of each test set, serving as an empirical reference to evaluate prediction accuracy. Expected mean values vary between 0.34 and 0.48 across the different sets analyzed.

A joint analysis of these three metrics reveals that the Transformer's self-attention mechanism produces richer and contextually more informative latent representations, improving generalization and adaptability to diverse data distributions. The convergence between predicted and expected means, evidenced by relative mean values close to unity, suggests a superior capacity for modeling complex patterns in railway data, with direct implications for operational efficiency and resource management in transportation systems.

Table 16: Results obtained by test set

Test set	Test records	Predicted mean		Expected mean		Relative mean		Good Ranking %	
		Transf.	LSTM	Transf.	LSTM	Transf.	LSTM	Transf.	LSTM
A9630-1	67	0.52	0.38	0.48	0.34	1.17	1.43	100	90
A9630-32	1277	0.50	0.38	0.45	0.34	1.36	1.42	100	89
A9630-8	438	0.50	0.38	0.44	0.35	1.31	1.27	100	90

A9630-16	757	0.50	0.39	0.45	0.35	1.35	1.53	99	90
A9630-2	131	0.53	0.39	0.49	0.34	1.48	1.56	99	91
A9630-4	243	0.49	0.39	0.46	0.35	1.19	1.6	98	92
L2600-16	285	0.34	0.58	0.29	0.56	1.72	1.36	98	95
L2600-2	46	0.36	0.57	0.34	0.55	1.27	1.42	98	97
A350-16	563	0.57	0.58	0.55	0.55	1.42	1.47	97	96
A350-2	143	0.58	0.56	0.55	0.53	1.47	1.5	97	96
A350-32	797	0.56	0.55	0.53	0.52	1.50	1.42	97	96
A350-8	379	0.58	0.58	0.55	0.55	1.54	1.54	97	97.
A450-2	202	0.44	0.33	0.40	0.29	1.41	1.4	97	86
A450-4	386	0.44	0.32	0.38	0.29	1.41	1.52	97	83
A450-8	722	0.44	0.33	0.37	0.3	1.55	1.54	97	86
L2600-8	157	0.36	0.32	0.30	0.29	1.80	1.43	97	83
A350-4	241	0.55	0.33	0.52	0.3	1.42	1.24	96	82
A450-16	1306	0.43	0.31	0.37	0.29	1.53	1.5	96	81
A450-32	2247	0.43	0.41	0.37	0.35	1.44	1.53	96	93
A592-1	117	0.41	0.43	0.37	0.37	1.47	1.53	96	96
L1400-2	161	0.38	0.44	0.34	0.4	1.90	1.41	96	96
L2600-32	502	0.36	0.43	0.31	0.37	1.62	1.44	96	96
L2600-4	85	0.35	0.44	0.31	0.38	1.27	1.41	96	96
A350-1	81	0.58	0.44	0.56	0.37	1.36	1.55	95	97
A592-16	1429	0.39	0.41	0.34	0.37	1.47	1.47	95	95
A592-2	229	0.40	0.39	0.35	0.34	1.37	1.47	95	95
A592-32	2528	0.39	0.4	0.34	0.35	1.54	1.37	95	95
A592-4	431	0.39	0.39	0.34	0.34	1.39	1.54	95	95
A592-8	792	0.39	0.39	0.34	0.34	1.55	1.39	95	95
L1400-1	87	0.37	0.39	0.35	0.34	1.31	1.55	94	95
L1400-4	282	0.37	0.52	0.33	0.48	1.70	1.17	94	100
A450-1	105	0.41	0.5	0.35	0.45	1.53	1.35	93	99
L1400-16	727	0.38	0.53	0.34	0.49	1.60	1.48	93	99
L1400-8	461	0.37	0.5	0.33	0.45	1.56	1.36	93	99
A2240-8	971	0.39	0.49	0.35	0.46	1.60	1.19	92	98
L1400-32	1050	0.38	0.5	0.34	0.44	1.68	1.31	92	99

L2600-1	24	0.34	0.37	0.27	0.35	1.88	1.31	92	94
A2240-4	501	0.39	0.38	0.34	0.34	1.56	1.6	91	92
A2240-1	130	0.38	0.38	0.34	0.34	1.43	1.9	90	95
A2240-32	3459	0.39	0.38	0.35	0.34	1.53	1.68	90	91
L5600-2	245	0.36	0.37	0.32	0.33	1.67	1.7	90	93
A2240-16	1866	0.38	0.37	0.34	0.33	1.42	1.56	89	93
A2240-2	257	0.38	0.34	0.35	0.27	1.27	1.88	89	91
L5600-1	126	0.34	0.34	0.30	0.29	1.30	1.72	87	97
A4000-1	119	0.33	0.36	0.29	0.34	1.40	1.27	86	97
A4000-2	228	0.33	0.36	0.30	0.31	1.54	1.62	86	96
L5600-8	892	0.35	0.35	0.32	0.31	1.47	1.27	86	96
L5600-32	2878	0.35	0.36	0.32	0.3	1.50	1.8	85	96
L5600-4	475	0.36	0.34	0.32	0.3	1.50	1.3	85	87
L5600-16	1632	0.34	0.34	0.32	0.32	1.45	1.45	84	84
A4000-16	1427	0.32	0.36	0.29	0.32	1.52	1.67	83	89
A4000-32	2435	0.32	0.35	0.29	0.32	1.43	1.5	82	84
A4000-4	437	0.33	0.36	0.30	0.32	1.24	1.5	82	85
A4000-8	813	0.31	0.35	0.29	0.32	1.50	1.47	81	85

Figure 21 presents the distribution of predicted frequency scores for various disruptive event causes across nine datasets (A350-32, A450-32, A592-32, A2240-32, A4000-32, A9630-32, L1400-32, L2600-32, and L5600-32). Box plots were used to statistically characterize these distributions and enable comparative analysis between causes and datasets. In the A350-32 dataset, causes such as “Collision with infrastructure” and “Documentation problem” exhibit elevated medians (>0.6), indicating strong predictive confidence, while “To be determined” and “Passenger disruption” show lower values. The wide range of scores in categories such as “Train preparation” reflects considerable variability. The A450-32 dataset generally shows lower median values (0.2–0.5) but reveals relatively elevated scores for “Schedule conflict” and “Engine problem” and lower scores for “Passenger incident” and “Animal collision.” Outliers are evident in causes such as “RS failure” and “Loading problem.” The A592-32 dataset resembles A350-32 but includes more outliers, particularly in “Traffic congestion” and “Equipment failure.” The A4000-32 and A9630-32 datasets show decreasing scores, with frequent medians below 0.2 and asymmetric distributions. This trend continues in the L1400-32 and L2600-32 datasets, while L5600-32 exhibits partial recovery (medians between 0.2–0.6) and more symmetric distributions.

Overall, the analysis indicates superior predictive performance in the A350 and A450 datasets compared to the L series, except for L5600. Causes related to infrastructure, documentation, and operational conflicts are more predictable, while those involving passengers, animals, and signaling failures present greater predictive challenges.

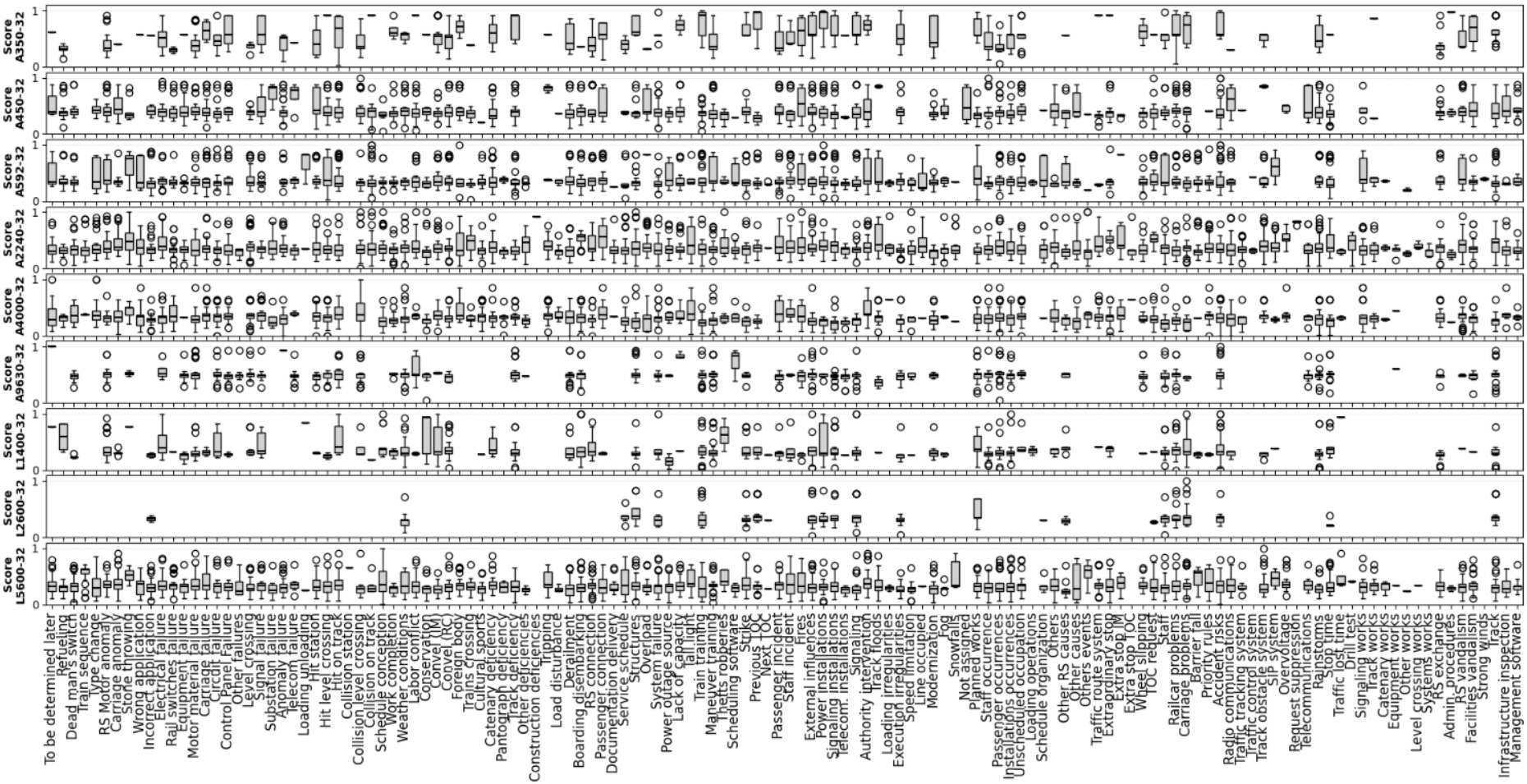

Figure 21: Relationship between test set and disruptive event cause (Transformer model)

Figure 22 presents an analysis of various performance metrics of the Transformer model using box plots arranged in horizontal rows, allowing direct comparisons between datasets. For the predicted frequency score, the A350 series (A350-8, A350-2, A350-1) stands out with mean values around 0.58, significantly above the global average of 0.39, while the A4000 series shows the lowest scores at approximately 0.32. The A350 series achieves superior performance due to its operational consistency, as these units primarily serve consistent regional routes, creating more predictable patterns that the Transformer model can capture effectively. In contrast, the lower scores of the A4000 likely reflect the challenges in modeling its less frequent but more diverse long-distance routes.

The expected frequency score reveals a similar pattern: A350-1, A350-2, and A350-16 exceed 0.54, while L2600-1, L2600-16, and A4000-32 remain below 0.30. The relative score has a global average of 1.49, indicating that predictions generally exceed expectations based on historical frequencies. The L series stands out in this metric, with L1400-2, L2600-1, and L2600-8 exceeding 1.80. While the A350 series demonstrates higher absolute frequency values due to its route consistency, the L series achieves superior relative performance despite operating on variable medium-distance routes and occasional special services. The model appears to have developed a robust capacity to identify patterns in more complex operational scenarios, particularly for the L series, which must adapt to changing route assignments.

KL divergence shows substantial variability between datasets, with significant outliers. Perplexity, as a metric that quantifies the model's uncertainty in its predictions, indicates that the Transformer model is, on average, undecided between approximately 4.60 options. Given the inherent complexity of predicting operational disruptions, we consider this an acceptable level of predictive certainty. Plausibility metrics with z-scores close to zero and mean percentiles close to 50 exhibit symmetric and consistent distributions, indicating adequate calibration across data types. Validation with complementary metrics confirms that the model achieves over 95% “GOOD” plausibility in more than 10 datasets, including 100% in multiple A9630 subsets (likely due to its operation on a closed-loop railway line, creating highly predictable patterns) and 96-97% in the A350 series. The global average relative score was 1.49, reaching 1.90 in cases such as L1400-2. We find the L1400-2 result particularly impressive considering that this series manages regional routes and special services outside planned routes. KL divergence remained low in datasets with elevated plausibility, and normalized metrics showed stable distributions, confirming the structural fidelity of the model's predictions across diverse operational contexts.

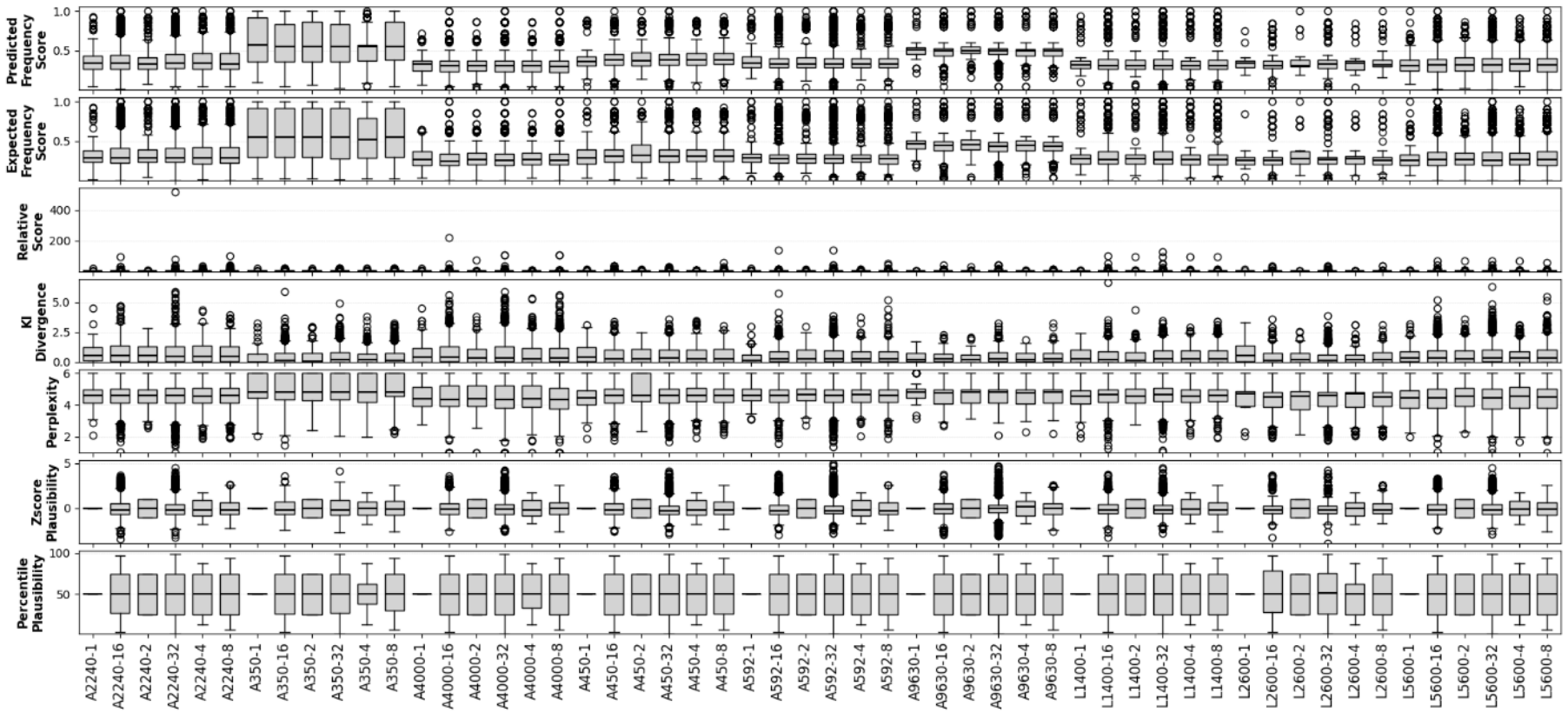

Figure 22: Performance metrics of the model by test set (Transformer model)

Figure 23 illustrates the LSTM model's performance, revealing substantial variation in predictive accuracy across the nine evaluated datasets. In A350-32, the model performs moderately well, with medians around 0.5 to 0.6 for "Collision with infrastructure" and "Documentation problem," while other causes vary between 0.3 and 0.4. The consistent operation of the A350 on fixed regional routes generates predictable patterns, which explains its relatively strong performance in the LSTM model. Meanwhile, categories such as "To be determined" and "Passenger disruption" perform poorly with values below 0.2, while "Train preparation" and "Schedule conflict" exhibit considerable dispersion across the dataset despite the route consistency.

The A450-32 series shows a general performance decline, with medians typically between 0.2 and 0.4. Although "Schedule conflict" and "Engine problem" stand out (approximately 0.45), "Animal collision" and "Passenger incident" exhibit central values close to zero, and outliers appear in "RS failure" and "Loading problem." This diminished performance compared with A350-32 aligns with the A450's operation across more varied regional routes, which introduces greater variability in the prediction task.

Despite the broader route coverage of the A592 series, the A592-32 dataset shows notable similarity with the A350-32 performance pattern, although with increased noise. It maintains strong medians (0.45 to 0.55) for causes related to infrastructure and documentation, while exhibiting more outliers in the "Traffic congestion" and "Equipment failure" categories. This performance consistency suggests that the LSTM model successfully identifies and leverages recurrent operational patterns, even when confronted with moderate route variation.

In A2240-32, model performance diminishes, with most medians below 0.3 and asymmetric distributions, especially for "Construction delay," "Technical problem," and "Schedule conflict," which rarely exceed 0.4. This notable decline likely reflects the challenges posed by the A2240's extensive coverage across diverse routes and regional services, creating a more complex prediction landscape for the model.

The A4000-32 and A9630-32 datasets exhibit severe performance degradation, characterized by medians below 0.2, flattened distribution boxes, and long whiskers indicating high dispersion with rare instances of elevated scores. This marked decline in A4000-32 can be attributed to its operational profile of irregular service across diverse long-distance routes, which creates sparse and irregular patterns that challenge the model's predictive capabilities.

Conversely, the weak results of A9630-32 are unexpected, as its closed-loop operation should theoretically create more predictable patterns. This suggests that the LSTM struggles with specific operational characteristics despite the apparent predictability.

The L1400-32 and L2600-32 exhibit notably weak predictive capacity, with medians close to zero (0.0 to 0.1) and minimal variance. This performance deficit corresponds directly to their highly variable operational profiles. L1400 primarily serves unpredictable special services outside standard routes, while L2600 navigates changing medium-distance routes and occasionally

supplements long-distance services. These inconsistent operational patterns create significant learning challenges for the LSTM model, which struggles to identify reliable predictive signals within such variable service contexts.

The L5600-32 set exhibits a modest performance recovery, with symmetric distributions and medians of 0.2 to 0.4 for collisions with infrastructure, schedule conflicts, and documentation problems. Although still inferior to A350-32 and A592-32, this relative improvement over other L series units likely reflects the L5600's structured long-distance service patterns, which provide operational consistency despite serving multiple routes and regions.

Overall, the LSTM model performs best in A350-32 and A592-32, particularly for causes related to infrastructure, documentation, and operational conflicts, which correlates with their more consistent operational patterns. In contrast, events involving passengers, animals, and signaling remain the most difficult to predict, and datasets representing services with highly variable or special routes (A4000-32, A9630-32, L1400-32, and L2600-32) consistently show inferior robustness compared with the Transformer in similar contexts.

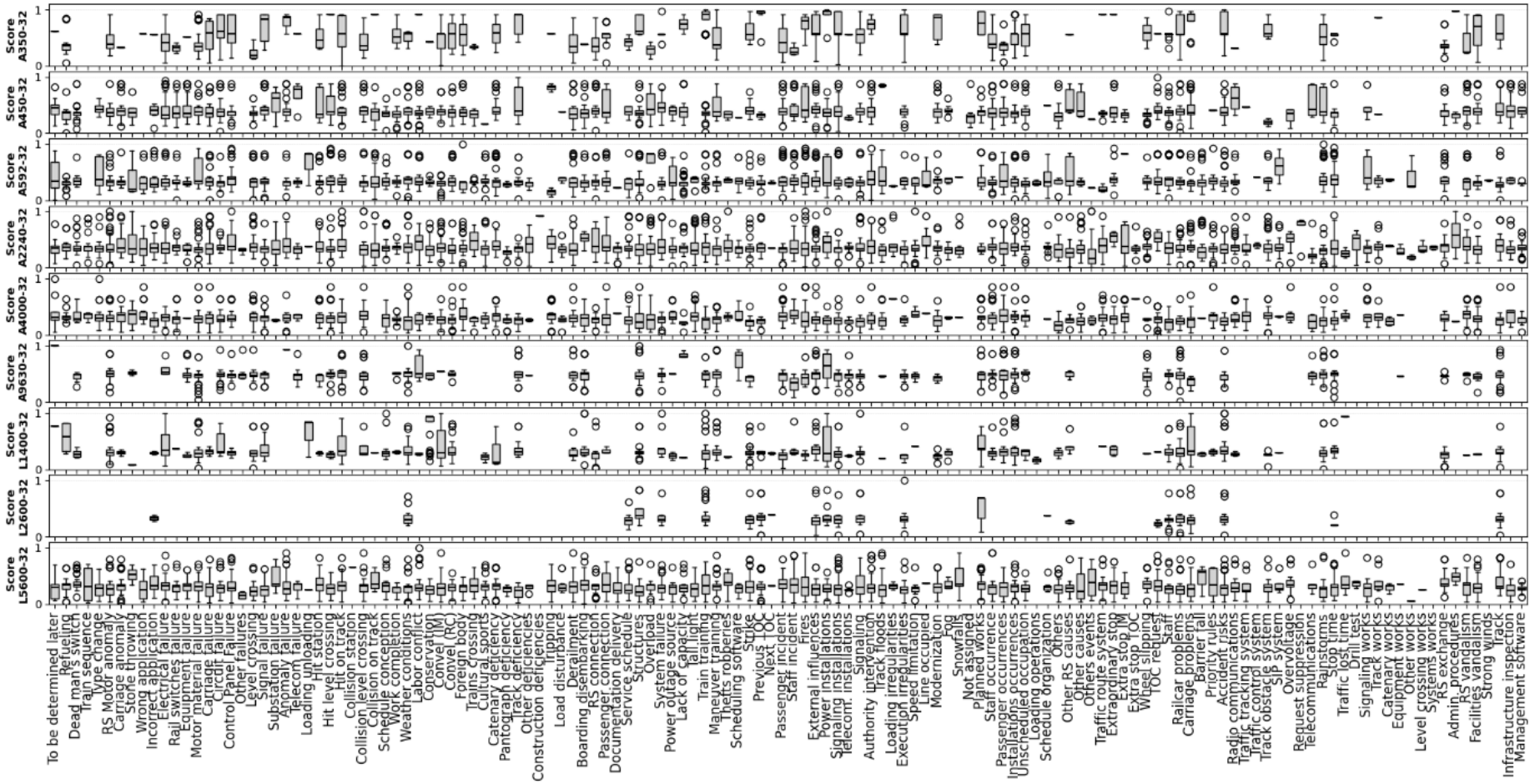

Figure 23: Relationship between test set and disruptive event cause (LSTM model)

In Figure 24, models of the A350 series (A350-8, A350-2, and A350-1) exhibit predicted frequency scores with an average around 0.58, significantly above the global average of approximately 0.39. In contrast, the A4000 series registers the lowest values, around 0.32. Expected frequency scores follow a comparable pattern. A350 subsets consistently achieve values above 0.54, whereas models L2600-1, L2600-16, and A4000-32 achieve significantly lower scores below 0.30. The relative score averages 1.49, indicating that predictions generally exceed historical expectations. Notably, L series subsets, particularly L1400-2, L2600-1, and L2600-8, exceed 1.80 on average, demonstrating stronger relative performance despite lower absolute frequencies. KL divergence varies widely between subsets and includes several outliers, reflecting dataset heterogeneity. In contrast, perplexity remains consistently close to 4.6 with limited dispersion, indicating stable overall model fit. Plausibility metrics show symmetric and consistent distributions, with z-scores centered close to zero and percentiles averaging around 50. These findings suggest well-calibrated performance across datasets. Comparison of plausibility statistics with classical metrics such as KL divergence further validates these results. “GOOD” plausibility rates exceed 95% in over ten subsets, namely A9630 (100%) and A350 (96% to 97%).

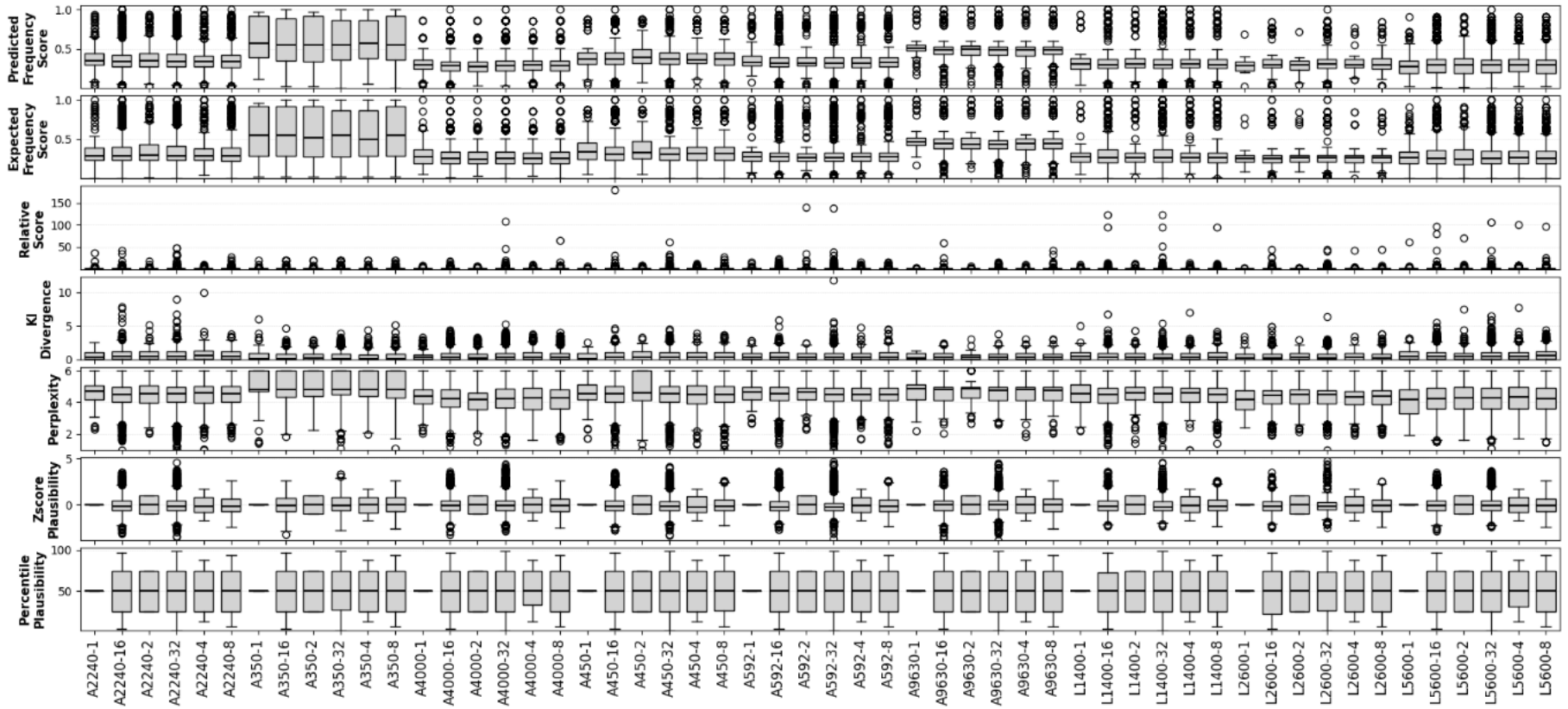

Figure 24: Performance metrics of the model by test set (LSTM model)

The results reveal distinct performance patterns between Transformer and LSTM architectures for predicting railway operational sequences under disruptive events. Quantitative analysis demonstrates the statistical superiority of the Transformer model while exposing specific behavioral differences in how these architectures respond to various disruption types and rolling stock configurations.

6.11 Discussion

The observed superiority of the Transformer model over the LSTM architecture can be attributed to the distinct mechanisms by which each captures dependencies within operational task sequences. LSTMs process information sequentially, which inherently limits their ability to model long-range dependencies and increases training time due to the absence of parallelization. In contrast, the Transformer leverages a self-attention mechanism that allows each token to attend to all others in the sequence simultaneously. This global context awareness enables the model to identify relationships between distant operational events, such as disruptions occurring at spatially separated stations but within the same rolling-stock chain, thereby improving predictive coherence and plausibility.

Furthermore, the specialized semantic tokenization developed in this research preserves the hierarchical structure of railway data by encoding spatial and temporal contexts through delimiters and positional embeddings. This representation allows the Transformer to learn not only the order of tasks but also their relational significance within the operational topology. The multi-head attention layers effectively decompose the learning process into multiple contextual subspaces, enhancing sensitivity to disruption causes, task transitions, and region-specific dependencies. As a result, the Transformer achieves lower validation loss and higher plausibility because it captures a richer multidimensional representation of operational behavior, something that recurrent networks, constrained by their sequential memory flow, cannot efficiently approximate.

The empirical and methodological contributions presented here conclude the artifact development process and provide the foundation for the integrative discussion, final conclusions, and directions for future work presented in Chapter 7.

Chapter 7: Conclusion

Building on the modeling and sequence-generation results reported in Chapter 6, this chapter synthesizes the overall contributions of the thesis, answers the research questions, and reflects on the implications for disruption management in passenger railways. It consolidates theoretical insights, methodological advances, and empirical evidence, and outlines limitations and actionable directions for future work.

7.1 Main Results

The first investigation conducted a systematic literature review on railway disruption management, analyzing 28 scientific publications from the last decade and identifying three critical structural limitations documented by Fang et al. (2015) and Schipper et al. (2015): extreme geographical fragmentation (68% of studies concentrated in the Netherlands context), predominance of traditional linear programming paradigms (60.7% of methodological approaches), and systematic under exploration of artificial intelligence techniques (only 3.6% of studies incorporated machine learning methodologies). This analysis contrasted markedly with the persistent methodological fragmentation identified by Dollevoet et al. (2017), where few approaches could effectively integrate multiple simultaneous problems, establishing a conceptual framework that transcends the conventional dichotomy between infrastructure management and the operational perspective of railway operators.

The second investigation materialized the first systematic application of advanced machine learning techniques to the Portuguese railway context, using 89,338 records of real disruptive events from CP in the 2015-2022 period. This approach overcame the limitations of simulation-based scenario studies identified by Nielsen et al. (2012). The main innovation consisted of a multidimensional approach that simultaneously predicts three critical operational metrics, transcending the single-objective fragmentation that characterizes the literature. Random Forest models with IQR winsorization achieved an MSE of 1.53 and MAE of 0.93 for larger datasets, establishing new performance benchmarks that surpass the predictive accuracy results documented by Yaghini et al. (2013) in conventional artificial neural networks.

This investigation introduced operational variables previously unexplored in the literature, including network topological indicators such as the betweenness centrality indicator. These variables contrasted with the predominance of approaches based exclusively on aggregated historical data identified in the systematic review. The rigorous application of the CRISP-DM methodology enabled systematic comparisons between multiple architectures, empirically demonstrating the advantages and limitations of each approach in different operational contexts.

The third investigation represented the technological culmination through the pioneering development of Transformer architecture specifically adapted to the railway domain, using 371,668

operational records (2016-2024). This definitively overcame the computational limitations documented by Louwerse and Huisman (2014), where problems involving more than one hundred trains became intractable. The Transformer model achieved a 75% reduction in validation error compared to conventional LSTM models (0.240 vs 0.947) and demonstrated computational efficiency four times superior (461.74 vs 1,829.32 seconds per epoch), representing a qualitative leap that renders obsolete the linear programming methods documented by Cacchiani et al. (2014) and the branch-and-price limitations of Lusby et al. (2017).

The architectural innovations included specialized semantic tokenization strategies that preserve the hierarchical integrity of railway operational data and a multidimensional evaluation framework based on statistical plausibility, transcending the conventional point metrics documented in the literature. Validation through 27 specific test sets revealed classification rates of 100% in multiple operational scenarios, contrasting markedly with the 89-92% documented by previous studies by Pölz et al. (2024) in time series applications.

The results demonstrate a systematic overcoming of the limitations identified in international literature, establishing new methodological foundations that transcend the persistent methodological fragmentation identified by Corman and Meng (2015). The holistic multidimensional approach definitively resolves the dominant tendency toward decomposition into isolated subproblems identified in the critical analysis, simultaneously integrating three critical metrics through advanced deep learning architectures.

This convergence demonstrates how scientifically rigorous research and practical applicability converge, showing how problem-oriented research can generate transformative scientific contributions that definitively overcome the limitations of the paradigms established by Corman et al. (2012) and Dollevoet et al. (2017).

7.2 Practical Implications

The practical implications of this research extend beyond the academic domain, offering concrete opportunities for operational integration within railway control centers such as those of Comboios de Portugal (CP). The predictive and prescriptive models developed in this thesis could be embedded into real-time decision support tools that assist dispatchers and operations managers in managing disruptions more efficiently. In practice, the multidimensional predictive models (Chapter 5) could continuously monitor network conditions and provide early warnings of high-impact disruptions, quantifying expected delay minutes, affected trains, and passenger volumes. The Transformer-based architecture (Chapter 6) could then generate operationally coherent recovery sequences, recommending alternative rolling stock allocations and timetable adjustments. When integrated into CP's operational management systems, these models would enable a transition from reactive to proactive control, improving service regularity, reducing cascading delays, and enhancing passenger satisfaction. Moreover, their implementation would contribute to more

sustainable and cost-efficient railway operations by optimizing the use of rolling stock and energy resources. These findings demonstrate the tangible potential of artificial intelligence to strengthen the resilience and decision-making capacity of modern railway systems.

7.3 Future Work

The scientific contributions developed in this research open multiple promising directions for future investigation, structured into six main strategic dimensions.

Extending specialized Transformer architectures is an immediate priority. This involves exploring transfer learning techniques to adapt the models to different operational contexts and railway networks in other geographic regions. Integrating federated learning techniques offers a strategic opportunity to amplify predictive capabilities by aggregating operational knowledge from multiple railway operators while simultaneously preserving the confidentiality of sensitive data. Developing hybrid architectures that combine the attention capabilities of Transformer models with classical optimization techniques would enable the creation of solutions that integrate predictive accuracy with optimality guarantees.

Expanding the operational scope through the incorporation of additional impact metrics is a fundamental strategic direction. Developing models capable of predicting impacts on energy consumption, carbon emissions, and passenger satisfaction would provide a more holistic approach to disruption management, aligning with contemporary sustainability objectives of the railway sector. Integrating Internet of things (IoT) sensor data and real-time monitoring systems would represent a transformative opportunity to amplify predictive capabilities, enabling proactive detection of potential disruptions before they manifest operationally.

Transferring the methodological innovations developed to other transportation domains presents significant potential. Air, road, and maritime transport share similar structural characteristics that could benefit from semantic tokenization and multidimensional evaluation approaches. Application to urban logistics and traffic management represents a particularly relevant opportunity, given the increasing complexity of urban mobility systems and the need for integrated multimodal solutions.

Developing practical interfaces and operational systems is a fundamental priority to maximize the impact of these scientific contributions. Creating real-time decision support systems that incorporate multidimensional predictions into an operational dashboard architecture would provide dispatchers with practical tools for effective disruption management, integrating advanced visualization capabilities with proactive alerts and actionable recommendations.

The future work directions identified provide a comprehensive roadmap to amplify and consolidate the innovations developed. They demonstrate how the convergence between scientific rigor and practical applicability can generate transformative contributions that transcend the

limitations of established paradigms, contributing to the continuous evolution of more resilient, efficient, and sustainable transportation systems.

References

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended Long Short-Term Memory. http://arxiv.org/abs/2405.04517

Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications, 44(9), 875–886. https://doi.org/10.1080/1206212X.2021.1974663

Bešinović, N. (2020). Resilience in railway transport systems: a literature review and research agenda. Transport Reviews, 0(0), 1–22. https://doi.org/10.1080/01441647.2020.1728419

Boateng, V. A., & Yang, B. (2023). A Global Modeling Pruning Ensemble Stacking With Deep Learning and Neural Network Meta-Learner for Passenger Train Delay Prediction. IEEE Access, 11(May), 62605–62615. https://doi.org/10.1109/ACCESS.2023.3287975

Borndörfer, R., Grimm, B., Reuther, M., & Schlechte, T. (2017). Template-based re-optimization of rolling stock rotations. Public Transport, 9(1–2), 365–383. https://doi.org/10.1007/s12469-017-0152-4

Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., & Wagenaar, J. (2014). An overview of recovery models and algorithms for real-time railway rescheduling. Transportation Research Part B: Methodological, 63, 15–37. https://doi.org/10.1016/j.trb.2014.01.009

Cao, K., Zhang, T., & Huang, J. (2024). Advanced hybrid LSTM-transformer architecture for real-time multi-task prediction in engineering systems. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-55483-x

Cats, O., West, J., & Eliasson, J. (2016). A dynamic stochastic model for evaluating congestion and crowding effects in transit systems. Transportation Research Part B: Methodological, 89, 43–57. https://doi.org/https://doi.org/10.1016/j.trb.2016.04.001

Chen, X., Ma, Z., & Li, Z. (2022). Unplanned Disruption Analysis and Impact Modeling in Urban Railway Systems. Transportation Research Record, 2676(10), 16–27. https://doi.org/10.1177/03611981221088221

Corman, F., & Meng, L. (2015). A review of online dynamic models and algorithms for railway traffic management. IEEE Transactions on Intelligent Transportation Systems, 16(3), 1274–1284. https://doi.org/10.1109/TITS.2014.2358392

Corman, F., D'Ariano, A., Pacciarelli, D., & Pranzo, M. (2011). Optimal inter-area coordination of train rescheduling decisions. Transportation Research Part E: Logistics and Transportation Review, 48(1), 71–88. https://doi.org/10.1016/j.tre.2011.05.002

Corman, F., D'Ariano, A., Pacciarelli, D., & Pranzo, M. (2012). Bi-objective conflict detection and resolution in railway traffic management. Transportation Research Part C: Emerging Technologies, 20(1), 79–94. https://doi.org/10.1016/j.trc.2010.09.009

Dollevoet, T., Huisman, D., Kroon, L. G., Veelenturf, L. P., & Wagenaar, J. C. (2017). Application of an iterative framework for real-time railway rescheduling. Computers and Operations Research, 78, 203–217. https://doi.org/10.1016/j.cor.2016.08.011

Fabella, V. M., & Szymczak, S. (2021). Resilience of railway transport to four types of natural hazards: An analysis of daily train volumes. Infrastructures, 6(12). https://doi.org/10.3390/infrastructures6120174

Fang, W., Yang, S., & Yao, X. (2015). A Survey on Problem Models and Solution Approaches to Rescheduling in Railway Networks. IEEE Transactions on Intelligent Transportation Systems, 16(6), 2997–3016. https://doi.org/10.1109/TITS.2015.2446985

Gao, T., Chen, J., & Xu, H. (2023). Data-driven train delay prediction incorporating dispatching commands: An XGBoost-metaheuristic framework. IET Intelligent Transport Systems. https://doi.org/10.1049/itr2.12461

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: continual prediction with LSTM. 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), 2, 850–855 vol.2. https://doi.org/10.1049/cp:19991218

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghaemi, N., Zilko, A. A., Yan, F., Cats, O., Kurowicka, D., & Goverde, R. M. P. (2018). Impact of railway disruption predictions and rescheduling on passenger delays. Journal of Rail Transport Planning and Management, 8(2), 103–122. https://doi.org/10.1016/j.jrtpm.2018.02.002

Gkiotsalitis, K., & Cats, O. (2018). Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. Transportation Research Part C: Emerging Technologies, 88, 187–207. https://doi.org/https://doi.org/10.1016/j.trc.2018.01.026

Golightly, D., & Dadashi, N. (2017). The characteristics of railway service disruption: implications for disruption management. Ergonomics, 60(3), 307–320. https://doi.org/10.1080/00140139.2016.1173231

Grandhi, B. S., Chaniotakis, E., Thomann, S., Laube, F., & Antoniou, C. (2021). An estimation framework to quantify railway disruption parameters. IET Intelligent Transport Systems, 15(10), 1256–1268. https://doi.org/10.1049/itr2.12095

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5–6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Grewal, M. S. (2011). Kalman Filtering. In M. Lovric (Ed.), International Encyclopedia of Statistical Science (pp. 705–709). Springer. https://doi.org/10.1007/978-3-642-04898-2_395

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent Neural Networks for Time Series Forecasting: Current status and future directions. International Journal of Forecasting, 37(1), 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Huang, P., Lessan, J., Wen, C., Peng, Q., Fu, L., Li, L., & Xu, X. (2020). A Bayesian network model to predict the effects of interruptions on train operations. Transportation Research Part C: Emerging Technologies, 114(August 2019), 338–358. https://doi.org/10.1016/j.trc.2020.02.021

Klumpenhouwer, W., & Shalaby, A. (2022). Using Delay Logs and Machine Learning to Support Passenger Railway Operations. Transportation Research Record, 2676(9), 134–147. https://doi.org/10.1177/03611981221085561

Kohl, N., Larsen, A., Larsen, J., Ross, A., & Tiourine, S. (2007). Airline disruption management—Perspectives, experiences and outlook. Journal of Air Transport Management, 13(3), 149–162. https://doi.org/https://doi.org/10.1016/j.jairtraman.2007.01.001

König, E., & Schön, C. (2020). Railway delay management with passenger rerouting considering train capacity constraints. European Journal of Operational Research, 288, 450–465. https://doi.org/10.1016/j.ejor.2020.05.055

Kroon, L., Maróti, G., & Nielsen, L. (2015). Rescheduling of Railway Rolling Stock with Dynamic Passenger Flows. Transportation Science, 49(2), 165–184. https://doi.org/10.1287/trsc.2013.0502

Li, Z. C., Wen, C., Hu, R., Xu, C., Huang, P., & Jiang, X. (2021). Near-term train delay prediction in the Dutch railways network. International Journal of Rail Transportation, 9(6), 520–539. https://doi.org/10.1080/23248378.2020.1843194

Li, Z., Huang, P., Wen, C., Tang, Y., & Jiang, X. (2020). Predictive models for influence of primary delays using high-speed train operation records. Journal of Forecasting, 39(8), 1198–1212. https://doi.org/10.1002/for.2685

Liu, J., Canca, D., Lv, H., & Ni, S. (2024). Demand-adapted train timetabling with coupling-decoupling operations on a bidirectional intercity railway line. Computers and Industrial Engineering, 189. https://doi.org/10.1016/j.cie.2024.109999

Louwerse, I., & Huisman, D. (2014). Adjusting a railway timetable in case of partial or complete blockades. European Journal of Operational Research, 235(3), 583–593. https://doi.org/10.1016/j.ejor.2013.12.020

Lusby, R. M., Haahr, J. T., Larsen, J., & Pisinger, D. (2017). A Branch-and-Price algorithm for railway rolling stock rescheduling. Transportation Research Part B: Methodological, 99, 228–250. https://doi.org/10.1016/j.trb.2017.03.003

Marković, N., Milinković, S., Tikhonov, K. S., & Schonfeld, P. (2015). Analyzing passenger train arrival delays with support vector regression. Transportation Research Part C: Emerging Technologies, 56(September 2018), 251–262. https://doi.org/10.1016/j.trc.2015.04.004

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.

Nabian, M. A., Alemazkoor, N., & Meidani, H. (2019). Predicting Near-Term Train Schedule Performance and Delay Using Bi-Level Random Forests. Transportation Research Record, 2673(5), 564–573. https://doi.org/10.1177/0361198119840339

Nielsen, L. K., Kroon, L., & Maróti, G. (2012). A rolling horizon approach for disruption management of railway rolling stock. European Journal of Operational Research, 220(2), 496–509. https://doi.org/10.1016/j.ejor.2012.01.037

Nishi, T., Ohno, A., Inuiguchi, M., Takahashi, S., & Ueda, K. (2017). A Combined column generation and heuristics for railway short-term rolling stock planning with regular inspection constraints. Computers and Operations Research, 81, 14–25. https://doi.org/10.1016/j.cor.2016.11.025

Pölz, A., Blaschke, A. P., Komma, J., Farnleitner, A. H., & Derx, J. (2024). Transformer Versus LSTM: A Comparison of Deep Learning Models for Karst Spring Discharge Forecasting. Water Resources Research, 60(4). https://doi.org/10.1029/2022WR032602

Schipper, D., & Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. Journal of Rail Transport Planning and Management, 8(1), 42–55. https://doi.org/10.1016/j.jrtpm.2017.12.003

Schipper, D., Gerrits, L., & Koppenjan, J. F. M. (2015). A dynamic network analysis of the information flows during the management of a railway disruption. European Journal of Transport and Infrastructure Research, 15(4), 442–464. https://doi.org/10.18757/ejtir.2015.15.4.3091

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. http://arxiv.org/abs/1701.06538

Shires, J. D., Ojeda-Cabral, M., & Wardman, M. (2019). The impact of planned disruptions on rail passenger demand. Transportation, 46(5), 1807–1837. https://doi.org/10.1007/s11116-018-9889-0

Smiti, A. (2020). A critical overview of outlier detection methods. Computer Science Review, 38, 100306. https://doi.org/10.1016/j.cosrev.2020.100306

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2, 2377–2385.

Su, B., D'Ariano, A., Su, S., Wang, Z., & Tang, T. (2024). A data-driven mixed-integer linear programming approach for real-time rescheduling of urban rail transit under rolling stock faults. Transportation Research Part C: Emerging Technologies, 169. https://doi.org/10.1016/j.trc.2024.104893

Su, H., Peng, S., Mo, S., & Wu, K. (2022). Neural Network-Based Hybrid Forecasting Models for Time-Varying Passenger Flow of Intercity High-Speed Railways. Mathematics, 10(23). https://doi.org/10.3390/math10234554

Su, L., Zuo, X., Li, R., Wang, X., Zhao, H., & Huang, B. (2023). A Systematic Review for Transformer-based Long-term Series Forecasting. https://doi.org/10.1007/s10462-024-11044-2

Tiong, K. Y., Ma, Z., & Palmqvist, C. W. (2023). A review of data-driven approaches to predict train delays. Transportation Research Part C: Emerging Technologies, 148. https://doi.org/10.1016/j.trc.2023.104027

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. http://arxiv.org/abs/1706.03762

Veelenturf, L. P., Kroon, L. G., & Maróti, G. (2017). Passenger oriented railway disruption management by adapting timetables and rolling stock schedules. Transportation Research Part C: Emerging Technologies, 80, 133–147. https://doi.org/10.1016/j.trc.2017.04.012

Wang, P., & Zhang, Q. P. (2019). Train delay analysis and prediction based on big data fusion. Transportation Safety and Environment, 1(1), 79–88. https://doi.org/10.1093/tse/tdy001

Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2013). Railway passenger train delay prediction via neural network model. Journal of Advanced Transportation, 47(3), 355–368. https://doi.org/https://doi.org/10.1002/atr.193

Yin, J., Pu, F., Yang, L., D'Ariano, A., & Wang, Z. (2023). Integrated optimization of rolling stock allocation and train timetables for urban rail transit networks: A benders decomposition approach. Transportation Research Part B: Methodological, 176, 102815. https://doi.org/https://doi.org/10.1016/j.trb.2023.102815

Yue, Y., Han, J., Wang, S., & Liu, X. (2017). Integrated Train Timetabling and Rolling Stock Scheduling Model Based on Time-Dependent Demand for Urban Rail Transit. Computer-Aided Civil and Infrastructure Engineering, 32(10), 856–873. https://doi.org/10.1111/mice.12300

Zhan, S., Xie, J., Wong, S. C., Zhu, Y., & Corman, F. (2024). Handling uncertainty in train timetable rescheduling: A review of the literature and future research directions. Transportation Research Part E: Logistics and Transportation Review, 183(January), 103429. https://doi.org/10.1016/j.tre.2024.103429

Zhang, D., Peng, Y., Zhang, Y., Wu, D., Wang, H., & Zhang, H. (2022). Train Time Delay Prediction for High-Speed Train Dispatching Based on Spatio-Temporal Graph Convolutional Network. IEEE Transactions on Intelligent Transportation Systems, 23(3), 2434–2444. https://doi.org/10.1109/TITS.2021.3097064

Zhong, Q., Lusby, R. M., Larsen, J., Zhang, Y., & Peng, Q. (2019). Rolling stock scheduling with maintenance requirements at the Chinese High-Speed Railway. Transportation Research Part B: Methodological, 126, 24–44. https://doi.org/10.1016/j.trb.2019.05.013

Zhou, H., Qi, J., Yang, L., Shi, J., Pan, H., & Gao, Y. (2022). Joint optimization of train timetabling and rolling stock circulation planning: A novel flexible train composition mode. Transportation Research Part B: Methodological, 162, 352–385. https://doi.org/10.1016/j.trb.2022.06.007

Zhou, Y., Zhou, L., Wang, Y., Yang, Z., & Wu, J. (2017). Application of multiple-population genetic algorithm in optimizing the train-set circulation plan problem. Complexity, 2017. https://doi.org/10.1155/2017/3717654