Public Transport
https://doi.org/10.1007/s12469-024-00380-9

ORIGINAL RESEARCH

Data-driven insights to reduce uncertainty from disruptive events in passenger railways

Luís Marques[1] · **Sérgio Moro**[1] · **Pedro Ramos**[1]

Accepted: 19 November 2024

Abstract

This study investigates the predictive modeling of the impact of disruptive events on passenger railway systems, using real data from the Portuguese main operator, Comboios de Portugal. We develop models using neural networks and decision trees, using key features such as the betweenness centrality indicator, railway track, time of day, and the train service group. Conclusively, these attributes significantly predict the impact on the proposed models. The research reveals the superior performance of neural network models, such as convolutional neural networks and recurrent neural networks in smaller datasets, while decision tree models, particularly random forest, stand out in larger datasets. The findings of this study unveil new attributes that can be employed as predictors. Additionally, they confirm, within this study's context, the effectiveness of certain traits previously recognized in the literature for mitigating the uncertainty associated with the uncertainty of the impact of disruptive events in passenger railway systems.

Keywords Disruptive events · Railway systems · Neural networks · Decision tree

1 Introduction

In passenger railway transportation systems, the occurrence of disruptive events poses a significant challenge, affecting punctuality, train regularity, and the overall passenger experience (Bešinović 2020). During disruptions in railway operations, the service offering can be affected by a reduction in available seats or a modification

✉ Luís Marques
luis_carlos_marques@iscte-iul.pt

Sérgio Moro
sergio.moro@iscte-iul.pt

Pedro Ramos
pedro.ramos@iscte-iul.pt

[1] ISTAR, ISCTE-IUL, Lisbon, Portugal

Published online: 12 February 2025

in train quality. Generally, available seats may be reduced due to insufficient commercial supply or the substitution of rolling stock with an inferior category. These measures aim to maintain service viability and customer satisfaction during disruptions. Train suppression is a drastic measure involving the withdrawal of services in response to technical issues, adverse weather conditions, or extraordinary events. Occasional suppressions also regulate schedules. During disruptions, it may be necessary to introduce additional trains outside the regular schedule to meet unexpected demand or to reroute rolling stock. Train delays are common during disruptions and can result from technical failures, weather conditions, or congestion. Punctuality is essential for passenger satisfaction, and managing delays involves quickly resolving problems and efficiently communicating with passengers (König 2020; Tiong et al. 2023).

Railway networks are exposed to various types of disturbances on a daily basis. Minor railway system disturbances can be managed by modifying train schedules without altering the tasks of rolling stock and crews. On the other hand, disruptions are relatively large incidents that require changes in both the train schedules and the tasks of the rolling stock and crews. These disruptions can be caused by a variety of human or equipment-related failures (rolling stock or infrastructure), leading to inconvenience for passengers and inefficiency in the railway system. In addition to the consequence of constraining passengers' journeys, disturbances cause significant financial losses (Huang et al. 2020). In rail operations, disruptive events can be classified into two main types: minor disturbances and major disruptions. Minor disturbances are short-term events that cause small interruptions to the normal operation of the rail system, such as delays in passenger boarding, minor technical issues, or brief weather interruptions, and can be managed with very little adjustments to schedules and resource allocations. Major disruptions, on the other hand, are long-term events that require extensive rescheduling, caused by internal factors such as critical failures in tracks or rolling stock, and external factors such as extreme weather conditions or accidents. These require substantial changes in operational plans and coordinated efforts to minimize impacts (Nielsen et al. 2012). Ge et al. (2022) propose a comprehensive classification of disruptions in transportation systems, categorizing them across various dimensions. Disruptions can be planned, such as maintenance, or unplanned, like delays and terrorist attacks. The probability of occurrence includes frequent events, such as demand fluctuations, and rare events, like volcanic eruptions. Impacts can vary widely in magnitude, ranging from minor occurrences such as limited access due to escalator malfunctions, to large-scale impacts like terrorist attacks that have serious implications for the continuity of railway service provision. Occurrences can be anticipated before the trip or during, with durations that may be short or long. Causes can be natural, like floods, or human-made, such as illegal track occupation. The scope varies from local to regional or global, and the location can be internal or external.

Reducing the uncertainty regarding the impact on the operational circulation of disruptive events in passenger rail transport is one of the daily goals of passenger rail transport companies (Ghaemi et al. 2018). Due to the characteristics of this type of transportation and its constant exposure to various points of failure, including the impact of weather conditions, operationality of infrastructure and rolling stock, as

well as the - sometimes - irregular behavior of passengers (Artan and Sahin 2022), the railway system is highly interconnected in its operation. This means that a disruptive event in one train can impact others, as the use of infrastructure is competitive, and the effects of disruptive events can have a sequential impact dynamic (Sajan and Kumar 2021). Inherent to this type of transport are disruptive events that can result in train cancellations, significant operational delays, and even accidents with material and human damage (Huang et al. 2020). Railway research literature demonstrates significant results in predicting and analyzing delays and interruptions. The study by Fabella and Szymczak (2021) revealed a considerable impact of natural disasters, such as floods and landslides, on the German railway network. Grandhi et al. (2021) showed that neural networks are more effective in predicting delays in the Danish Railway, indicating the crucial role of weather and specific attributes of disruptive events. Chen et al. (2022) found that factors such as time of day and weather conditions significantly affect delays in the Hong Kong Mass Transit Railway. Huang et al. (2020) achieved 96.6% accuracy in predicting delays using a hybrid model based on Bayesian networks in China. Golightly and Dadashi (2017) identified 26 relevant attributes for railway transportation interruptions through interviews with operational personnel in Great Britain. Boateng and Yang (2023) and Gao et al. (2023) implemented advanced machine learning techniques, such as XGBoost and deep neural networks, significantly improving accuracy in predicting delays.

In the railway context, the clear distinction between passenger operators and traffic controllers is an area requiring further investigation, as indicated by Spanninger et al. (2022). The interests of passenger operators do not always align with those of traffic controllers, who focus on the availability of infrastructure for train circulation. Operators, on the other hand, must ensure that passengers reach their destinations and that rolling stock and crews are appropriately positioned to meet schedules (Wang and Zhang 2019). This study exclusively focuses on attributes pertinent to passenger operators, omitting aspects of infrastructure management, such as platform use and single-track crossings. We did not identify models in the literature that differentiate operators and infrastructure managers with the same focus and objectives as this study, which prioritizes the specific needs of passenger railway operators (Leng and Weidmann 2017).

This study stands out for its innovative and comprehensive approach, integrating a wide range of attributes that provide a robust analysis of the impact of disruptive events on the railway. The originality of the work is evidenced by the introduction of new predictive variables not previously used in studies. It provides a detailed comparison of different evaluation methods, offering valuable insights to reduce the uncertainty of disruptive events on the railway. The article aims to contribute to the investigation of the impact of these events, focusing on the perspective of passenger rail operators, particularly regarding their operations. The main objectives and contributions of this research are:

- Identify the attributes that best explain the impact of disruptive events.
- Propose a model for predicting the total number of trains affected in a disruptive event.
- Propose a model for predicting the total delay of trains.

- Propose a model for predicting the total number of passengers affected.

In this investigation, data from the railway operator CP—Comboios de Portugal were used. The dataset spans from 2015 to 2022. We utilized operational data from the operator, including occurrences, delays, infrastructure characteristics, and climatic data collected from the Visual Crossing weather platform. CP is a public company controlled by the Portuguese State and is the largest passenger railway operator in Portugal, covering a significant portion of the mainland territory. It operates urban services in the main cities of Portugal as well as regional and long-distance services.

This article is structured as follows. In Sect. 2, we conduct a literature review. Section 3 describes the methodology, including a comprehensive analysis of the methodology, the used models, and the description of the data. Section 4 addresses data processing, including the management of extreme values and dimensionality reduction, and presents the used datasets. In Sect. 5, we discuss the obtained results, interpreting them and relating them to the existing literature, as well as acknowledging the limitations of the study. Finally, Sect. 6 concludes the article and suggests directions for future research.

2 Literature review

We conducted a literature review using the Web of Science and Scopus databases, focusing on academic articles and journals published in English with no restrictions regarding the publication date. The aim of this review was to identify methodologies that can be integrated to enhance the resilience and efficiency of railway operations in the face of adverse impacts, providing valuable insights for improving disruption management in the railway transportation sector. Initially, we performed the search using specific terms in the title, abstract, and keyword fields. In the Web of Science, the terms "disruption," "duration prediction," and "disruption management" combined with "railway" resulted in 545, 1, and 57 articles, respectively. In Scopus, the same terms returned 609, 1, and 58 articles. These results indicate a greater number of studies on Scopus concerning the topic, particularly regarding "railway disruptions" and "disruption management." To ensure the relevance of the selected articles, we applied elimination criteria in two stages. In the first stage, we excluded articles published in conference proceedings, those focused on railway infrastructure management, or those exclusively dedicated to infrastructure exploration without considering the impact on operations. After thoroughly reading the remaining articles, we applied additional elimination criteria, discarding studies that addressed planned disruptions (such as infrastructure improvement works), those that focused on socioeconomic analyses, and articles that did not have as their main objective the prediction of impacts on passenger railway operations. Following these rigorous filtering and selection stages, we identified 15 relevant articles to include in the literature review. These articles were analyzed to explore how different methodologies can be combined to strengthen operational resilience and improve disruption

management, contributing to a more efficient and proactive management of railway operations.

Fabella and Szymczak (2021) investigated the vulnerability of the German railway network to disasters such as floods, landslides, forest fires, and fallen trees. They used count data regression models, including negative binomial regression, combining daily rail traffic data with geospatial information on disruptive events. They concluded that floods significantly reduce the number of trains in operation, highlighting the need for more comprehensive data to analyze multiple simultaneous events.

Several studies have utilized machine learning models to predict rail delays. Grandhi et al. (2021) used neural networks to predict the duration and total delays of incidents on the Danish railway. They identified weather as a critical factor, though they faced challenges related to the accuracy of manually input data. Similarly, Yaghini et al. (2013) applied an artificial neural network model to predict delays in passenger trains on Iranian railways, using three different approaches for input definition: normalized real number, binary encoding, and binary encoding set. While the model demonstrated high accuracy, it requires large volumes of data.

Huang et al. (2020) developed a Bayesian network-based model to predict the effects of rail disruptions in China, focusing on delay propagation, the number of affected trains, and delay time. Although the model achieved high accuracy, it faced difficulties due to complexity and the need for specialized knowledge. Klumpenhouwer and Shalaby (2022) applied machine learning techniques, such as random forest regression and elastic net, to improve rail operations in the GO Rail network in Ontario. The models were effective in predicting delays, but had limitations in modeling delay propagation. The authors considered that expanding the dataset could mitigate some of these limitations.

In another group of studies, regression techniques were used to analyze rail delays. Chen et al. (2022) analyzed delays on Hong Kong's Mass Transit Railway caused by unplanned disruptions using quantile regression models. The results indicated that factors such as time of day and weather conditions were significant, but the study lacked information on procedures during disruptions and infrastructure data. Wang and Zhang (2019) used a boosting regression tree model to predict passenger train delay times in China, considering factors like weather conditions, the number of trains passing through each station, and delay history. The results showed that traffic volume and train conditions significantly impact delays and that these delays propagate to subsequent trains.

Some researchers have developed hybrid approaches and utilized advanced techniques to improve delay prediction. Boateng and Yang (2023) proposed a pruned ensemble learning approach (PST-NN) to predict passenger train delays in the USA, combining multiple deep learning sub-models (ANN, DNN, and CNN) and using meta-learning (MLP) to improve prediction accuracy. The magnitude-based pruning technique was applied to reduce model complexity and increase computational efficiency. The results showed that PST-NN significantly outperformed benchmark models in terms of accuracy and prediction error, demonstrating an improvement of up to 85.22% over existing models.

Gao et al. (2023) proposed an advanced model for predicting high-speed train delays in China, using the XGBoost algorithm combined with meta-heuristic algorithms for hyperparameter optimization. Based on a dataset of 1.9 million records over 38 months, including arrival and departure times, dispatch commands, and delay propagation information, the methodology demonstrates that dispatch commands and spatiotemporal relationships are crucial for prediction accuracy. The optimized model achieved an root mean square error (RMSE) of 2.56 min, an mean absolute error (MAE) of 1.78 min, and an R^2 of 0.87. However, it faces limitations due to computational complexity, reliance on high-quality historical data, and challenges in generalizing to other railway networks. Su et al. (2022) presented a hybrid methodology combining variational mode decomposition (VMD) with neural networks (MLP, GRU, and Bi-LSTM) to predict passenger flow in high-speed railways in China. Using historical data from the Guangzhou–Zhuhai track, the methodology decomposed the time series into stable subsequences processed by neural networks, showing that hybrid models are more accurate than individual models.

Nabian et al. (2019) proposed a two-level random forest approach to predict passenger train delays in the Netherlands, comparing it with several other techniques, such as gradient boosting, Adaboost, SVM, extra tree, logistic regression, decision tree, KNN, and naive Bayes. The study used data including scheduled times, historical train performance, crew schedules, rolling stock circulation, infrastructure data, and weather conditions, analyzing approximately 10 million data points over 13 weeks. Golightly and Dadashi (2017) identified characteristics of rail disruptions in Great Britain, revealing differences between events that stop and delay traffic, but with a focus solely on the British context, limiting its international applicability.

Marković et al. (2015) and Li et al. (2021) explored support vector regression (SVR) and artificial neural network (ANN) models to predict passenger train delays in Serbia and the Netherlands, respectively. Marković et al. (2015) analyzed data from 727 trains, considering variables such as train category, scheduled arrival time, infrastructure, distance, travel time, and intervals between trains. They concluded that SVR outperforms ANN in predictive accuracy and generalization, highlighting infrastructure as a critical factor in delays. Li et al. (2021) used the random forest (RF) model to predict delays in the Dutch railway network, analyzing historical rail operation data collected between September and December 2017, 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 ANN, XGBoost, and gradient boosted decision trees (GBDT), showing high accuracy but highlighting limitations due to dependence on specific historical data.

Li et al. (2020) used XGBoost and SVR to predict high-speed train delays in China, using data from the Wuhan–Guangzhou track. XGBoost achieved an MAE of 0.57 for the number of affected trains, while SVR achieved an MAE of 4.45 min for total delay time. However, the dependence on specific historical data may limit the generalization of the models to other tracks, and the absence of exogenous factors, such as weather conditions, may affect accuracy in complex situations. A literature review reveals that studies on predictive modeling of delays in railway systems use various approaches, such as machine learning, Bayesian networks, and quantile regression. Machine learning models, like neural networks and random forests, are

widely used due to their ability to handle large volumes of data and the complexity of delays. However, many studies face limitations related to data quality and comprehensiveness, the need for specialized knowledge for complex modeling, and difficulties in generalizing models to different geographical and operational contexts. Additionally, integrating exogenous variables, such as weather conditions and dispatch commands, is frequently highlighted as crucial for improving predictive accuracy but presents significant computational challenges.

The present study aims to overcome the limitations identified in the existing literature by addressing the need for more comprehensive data and automated data collection methods to enhance model reliability. By incorporating an extensive and diverse dataset spanning from 2015 to 2022, the research enables a robust analysis of simultaneous events in different contexts. The inclusion of additional variables, such as infrastructure and weather conditions, improves model generalization and the understanding of the factors causing delays. The research tests models in various railway contexts, increasing the robustness and applicability of predictions. Furthermore, the integration of detailed incident logs and the consideration of delay propagation within the network enhance predictive accuracy and the management of rare events. Advanced data preprocessing techniques and feature selection are employed to efficiently handle large volumes of data, improving model performance and interpretability.

3 Methodology

3.1 Comprehensive methodology analysis

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

Figure 1 illustrates the six phases of CRISP-DM, which are detailed below:

1. Business understanding: Recognizing the challenges faced by the railway and public transportation sectors, particularly due to disruptive events. The goal is to develop predictive models to mitigate these impacts, enhancing the resilience and robustness of the railway system operated by Comboios de Portugal (CP). It is essential to create models that predict and manage the impacts of disruptions more accurately.

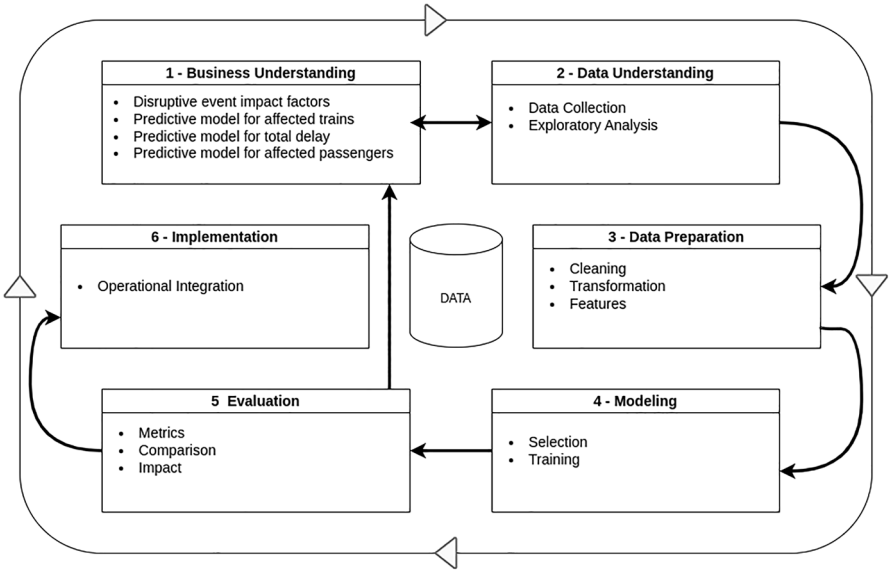

Fig. 1 The six phases of CRISP-DM, adapted from Martinez-Plumed et al. (2021)

2. Data understanding: Real data provided by CP includes information on railway infrastructure, train schedules, records of disruptive events, and operational data. Climatic data were collected from the Visual Crossing platform. Exploratory analysis involved identifying and analyzing key characteristics, 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 duplicate data to ensure the quality and accuracy of the model inputs. Techniques such as the removal of outliers using the interquartile range (IQR) and standard deviation (SD) were employed to handle extreme values. Data transformation included converting data into a suitable format for modeling, including normalization and encoding of categorical variables. Feature engineering was applied to create new characteristics, such as centrality metrics, that enhance the performance of predictive models. Techniques like the Chi-square test, LASSO, and recursive feature elimination (RFE) were used for dimensionality reduction and selection of the most relevant characteristics (Acito 2023).
4. Modeling: Multilayer perceptron (MLP) networks are particularly effective in modeling complex non-linear relationships, such as those between variables like the number of daily passengers, train intervals, the average daily number of trains, the number of trains affected by delays, total minutes of delay, and the number of affected passengers. These networks can capture complex patterns and subtle interactions between input and output variables, making them suitable for non-linear predictions. Convolutional neural networks (CNNs), although traditionally used for image data, can also be applied to time series and spatially structured data. In the context of CP data, CNNs can detect spatial and temporal patterns

in passenger data, including the number of trains affected by delays, by learning hierarchies of attributes. This capability is useful for identifying trends and patterns. Recurrent neural networks (RNNs) 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 historical data-based predictions. Random forest combines multiple decision trees to reduce the risk of overfitting, capturing complex interactions between variables. Extra-trees use 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, mean square error (MSE) and root mean square error (RMSE) are used to evaluate the models' predictive effectiveness. One-hot encoding (OHE) is applied to the 'Geographic Area' variable to convert its categories into binary values, enabling 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 implementing the developed models in CP's daily operations to predict the impacts of disruptive events.

3.2 Used models

3.2.1 Multilayer perceptron (MLP)

An MLP (multilayer perceptron) is composed of multiple layers of neurons, where each neuron in a layer is connected to all neurons in the preceding and succeeding layers. The output of a neuron a_i given by $a_i = f\left(\sum_j w_{ij} x_i + b_i\right)$, 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).

3.2.2 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: $h_t = f(W_{hh} h_{t-1} + W_{xh} x_t + b_h)$, where f is a non-linear activation function, W_{hh} and W_{xh} are weight matrices, and b_h is a bias vector (Su et al. 2022).

3.2.3 Convolutional neural network (CNN)

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

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

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

3.2.4 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 results 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).

3.2.5 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 the diversity among the trees in the model, helping to reduce the model's variance. The prediction of an extra-trees ensemble is given by the average (in regression) or the mode (in classification) of the predictions from all the individual trees (Geurts et al. 2006).

3.2.6 Hyperparameters

Our hybrid approach for hyperparameter optimization merges grid and random search (Belete and Huchaiah 2022). First, grid search provides a thorough analysis using a predefined matrix of values, including learning rate and neuron count. Then, having identified a promising range, we switched to random search for a faster, less resource-intensive exploration. This combination yields an optimal hyperparameter combination under given assumptions, balancing computational cost and efficiency, resulting in a model with good fit to training data and effective generalization to new data. For our grid strategy, we used the values described in Table 1 and Table 2.

3.3 Data description

The dataset for this study was provided by the passenger railway operator CP. This study analyzed the period from 2015 to 2022. Due to the COVID-19 pandemic, reliable data on the number of passengers per day at the control point was not available for the years 2021 and 2022. Consequently, these values were replaced with the average from the other years under analysis. The periods of lockdown hindered CP's ability to provide this variable with reliability for those years. We found 166,471 records of disruptive events for the selected period. For analysis purposes, only records with affected trains and assigned delay minutes were selected, resulting in 89,338 records of disruptive events. Figure 2 illustrates the betweenness centrality

Table 1 Hyperparameters combined in neural networks

Hidden layer sizes	Alpha	Learning rate init	Batch size	Activation
[(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)]	[0.0001, 0.001, 0.01]	[0.001, 0.01]	[32, 64]	[‘sigmoid’, ‘tanh’,’relu’, ‘adam’, ‘softmax’]

Table 2 Hyperparameters combined in random forests

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

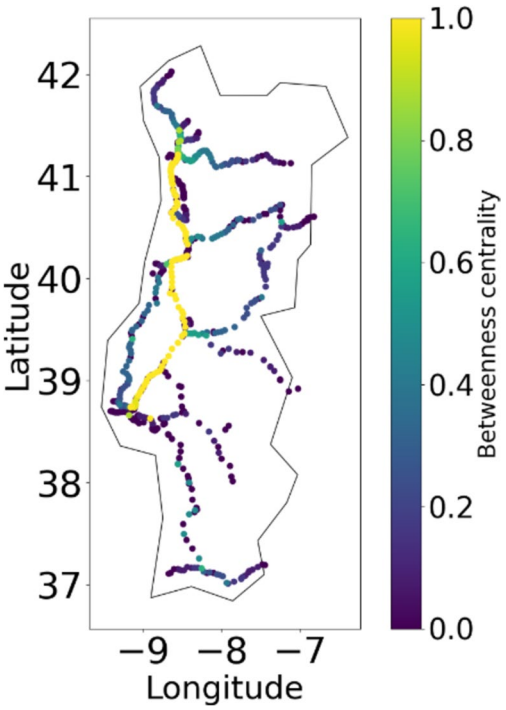

Fig. 2 Betweenness centrality of the railway network under analysis

of the railway network under analysis. This illustration is particularly useful for understanding the critical points and areas with the greatest influence; the closer to 1 (lighter color), the greater is the influence, and the closer to 0 (darker color), the smaller is the influence on the network.

Table 3 describes the variables of the dataset initially collected for the experiments undertaken in this study. Variables 1 through 26 are independent, and variables 27 through 29 are dependent.

3.4 Data processing

The analysis focused on three aspects: the number of delayed trains, total delay minutes, and affected passengers. Passenger numbers were classified into value classes to mitigate the complexity of forecasting a discrete value, while the other variables were analyzed using regression. The lower-class range was set from 1 to

Table 3 Description of variables

#	Short form	Description	Data type	Data range	Number of unique values
1	AltModes	Number of alternative modes of transportation	Num	0 to 87	32
2	Suppressions	Total number of suppressions occurred	Num	0 to 1498	36
3	UnplannedTrains	Total number of unplanned trains	Num	0 to 39	19
4	DailyTrains	Average daily number of trains at the 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	Damage to the infrastructure	Cat	0 to 1	2
10	RollStockDam	Damage to rolling stock	Cat	0 to 1	2
11	AssistReq	Request for assistance	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 the control point	Num	28 to 164,359	339
16	Km	Number of kilometers (the train that had the origin of the disruptive event)	int64	4265 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	Geographical control point	Cat	1008 to 95,125	416
20	SerialID	Rolling stock serial identification	Cat	350 to 9630	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	Hour of the day	Cat	0 to 23	24
24	DayWeek	Day of the week	Cat	0 to 6	7
25	Month	Month of the year	Cat	1to 12	12
26	ServGroup	Train service group	Cat	1to 39	29
27	TrainsDelay	Number of trains affected by delay	Num	1to 3839	129
28	TotDelay	Total number of minutes of delay	Num	1to 25,001	792
29	PaxAff	Number of affected passengers	Num	1to 615,226	4967

300, with 300 representing the minimum occupancy of the smallest train operated. This approach establishes a direct relationship between the predicted class and the rolling stock. Figure 3 illustrates the distribution of the 29 variables showing varying data distributions, skewness, and kurtosis. Variables 1–3 had less variation and

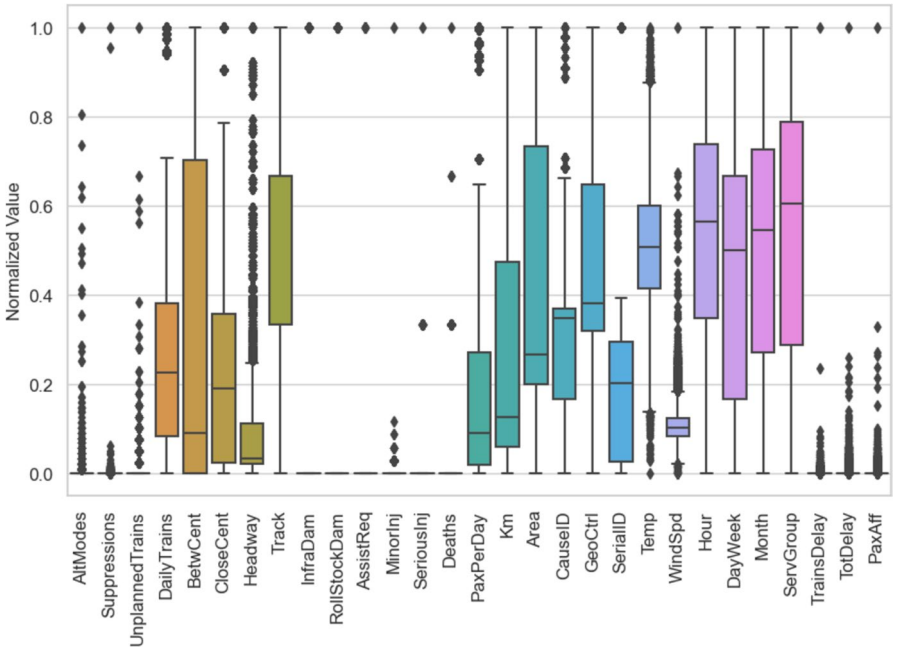

Fig. 3 Original dataset (normalized data)

lower median values, while 19–21 showed more variation and outliers, indicating a broader range. Variables 22–24 were more uniform, with fewer outliers. The whiskers in the data mostly extended to 1.5 times the interquartile range, highlighting common data ranges. This visual analysis underscored differences in centrality and dispersion among the variables.

3.4.1 Addressing extreme values

Two commonly used techniques were employed for the removal of outliers: 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 often used to identify outliers, which are defined as values that lie below $Q1 - 1.5$ IQR or above $Q3 + 1.5$ IQR.

3.4.2 Dimensionality reduction

Reducing dimensionality helps identify redundant variables and stabilize models. One common technique is the Chi-square test, which checks for association or independence between two categorical variables. It compares observed and expected frequencies to determine if variables are independent. The result is compared to a critical value from the Chi-square distribution (Yaghini et al. 2013). If the result 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, which can lead some coefficients to become exactly zero. This indicates that the features associated with these coefficients have no impact on the model, thereby reducing the number of used features (Klumpenhouwer and Shalaby 2022).

Similarly, recursive feature elimination (RFE) is a feature selection method that systematically removes the least important features from a model. Initially, the model is fitted with all features, and they are ranked based on the absolute value of their coefficients. The feature with the smallest coefficient is removed, and the model is refitted. This process is repeated until the desired number of features is achieved (Tiong et al. 2023).

Both LASSO and RFE are effective in simplifying a model while retaining the most significant features, improving both the model's performance and interpretability.

3.4.3 Datasets

In this section, we demonstrate the data processing for each independent attribute under study. The analysis focused on several key-independent attributes, including the number of alternative modes of transportation, total suppressions, and unplanned trains. It examined the average daily number of trains at the facility, betweenness and closeness centrality indicators, and the headway between trains. The report highlighted railway track conditions, infrastructure and rolling stock damage, assistance requests, and the number of minor and serious injuries, as well as recorded deaths. Additionally, it detailed passenger numbers at the control point, kilometers traveled by the disrupted train, the geographic area of incidents, cause group identities, control points, rolling stock IDs, and environmental factors such as temperature, wind speed, time of day, week, and month. The analysis also included train service groups. The number of trains affected by delay, the total number of minutes of delay, and the number of passengers affected were considered dependent variables subject to dimensionality reduction.

3.5 Number of trains affected by delay

Figure 4 illustrates the removal of outliers. For easier graphical visualization, the values were logarithmically transformed. The left graph uses standard deviation, and the right graph employs the IQR to predict delay minutes. Only attributes with significant differences (p-value > 0.05) are shown. Both graphs include average daily train count, centralities, train interval, presence of tracks, total kilometers, geographical area, incident cause, control point, circulating material, day of the week, and service group. Exclusively in the standard deviation method are alternative transportation, total suppressions, temperature, and month. Exclusively in the IQR method are passenger count at the control point and time of day.

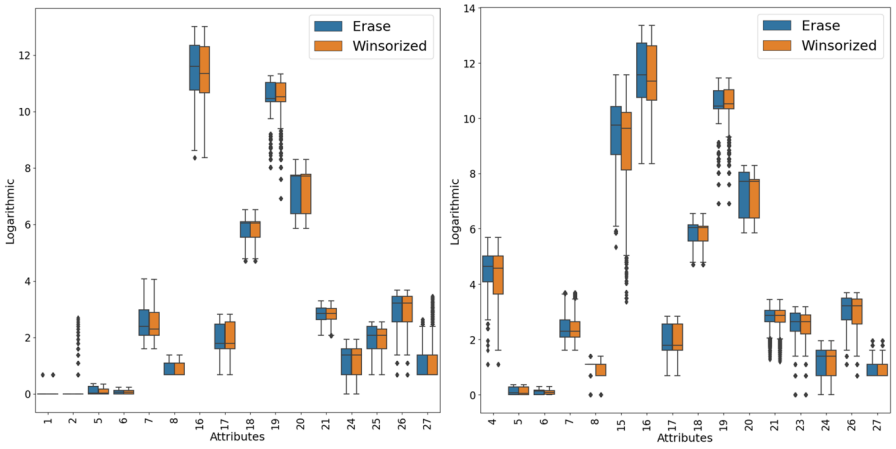

Fig. 4 Attributes with statistically significant differences (number of trains affected by delay). Variables are numbered as per Table 3

3.6 Total number of minutes of delay

Figure 5 illustrates the removal of outliers using standard deviation (left) and the IQR (right) to predict delay minutes. The common attributes include average daily train count, centralities, train interval, presence of tracks, total kilometers, geographical area, incident cause ID, control point, rolling stock ID, day of the week, and service group classification. Exclusively in the standard deviation method are alternative transportation, total suppressions, temperature, and month. Exclusively in the IQR method are daily passenger count at the control point and time of day.

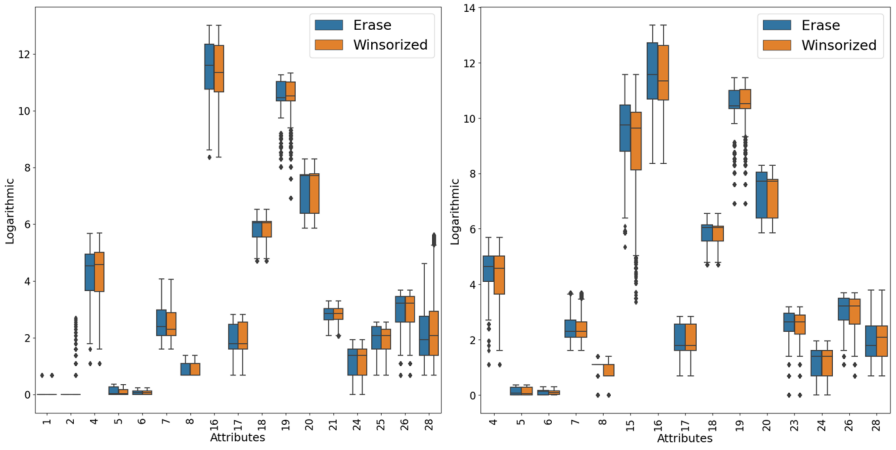

Fig. 5 Attributes with statistically significant differences (total number of minutes of delay). Variables are numbered as per Table 3

3.7 Number of affected passengers

Figure 6 illustrates the removal of outliers for the number of passengers, using standard deviation (left) and the IQR (right), with statistically significant attributes (p value < 0.05). Common to both methods are betweenness and closeness centralities, train interval, presence of tracks, train kilometers, covered area, incident cause ID, control point ID, rolling stock ID, day of the week, and service group classification. Exclusively in the standard deviation method are alternative transportation, total suppressions, temperature, and month. Exclusively in the IQR method are average daily train count, daily passenger count at control points, and time of day.

4 Results

4.1 Number of trains affected by delay

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

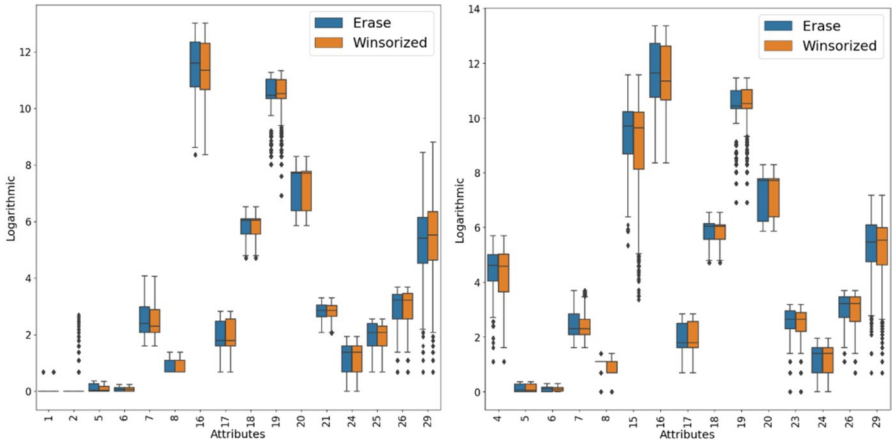

Fig. 6 Attributes with statistically significant differences (number of affected passengers). Variables are numbered as per Table 3

Finally, the attributes alternative transportation, temperature in Celsius, and wind speed in km/h, which appear less frequently, 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 attribute total suppression, appearing only in two datasets, stands out as the least frequent, which might suggest a very specific use or limited focus in the undertaken analyses.

4.1.1 Hyperparameters

MLP: Our MLP model, featuring 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 dropout unhelpful for improvement. The data split was 80% training, 20% 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 mirrors the MLP.

CNN: Our CNN 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 process. The data split is consistent with the 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.

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

RF-Extra: 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.

We used a max depth of 20, minimum samples split of 5, and a minimum samples leaf of 4 for all random forest models.

4.1.2 Results

Table 4 describes the best results for datasets with approximately 8,000 records, with CNN and RNN performing particularly well on smaller sets. The very small variation in indicators across various preprocessing techniques underscores the robustness of these models in handling differences in data treatment.

Table 5 describes the best results for datasets with approximately 8,000 records. RF with IQR winsorization and Chi-squared proved effective with 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 RF.

MLP: With an MSE of 1.65 and an MAE of 0.86 using IQR winsorization and RFE, the MLP demonstrated its capacity to efficiently handle large volumes of data.

This analysis demonstrates the importance of selecting appropriate models and preprocessing strategies in machine learning. CNN and RNN are suitable for smaller datasets, whereas RF and MLP perform better with larger datasets. For example, using Standard Deviation Erase and LASSO in an RF model on a 7,934-record

Table 4 Results for the number of affected trains (≈ 8000 records)

Dataset	Method	MSE	MAE	RMSE	Records
IQR erase with Chi-squared	CNN	1.49	0.84	1.22	8098
IQR erase with RFE and OHE	RNN	1.49	0.85	1.22	8098
IQR erase with LASSO and OHE	CNN	1.49	0.85	1.22	8098
IQR erase with Chi-squared and OHE	CNN	1.50	0.83	1.22	8098
IQR erase with LASSO	RNN	1.50	0.87	1.23	8098
IQR erase with RFE	RNN	1.52	0.82	1.23	8098
SD erase with LASSO and OHE	RF	3.84	1.31	1.96	7934
SD erase with LASSO	RF	3.85	1.31	1.96	7934
SD erase with Chi-squared and OHE	RF	3.90	1.32	1.97	7934
SD erase with Chi-squared	RF	3.91	1.32	1.98	7934
SD erase with RFE	RNN	4.11	1.19	2.03	7934
SD erase with RFE and OHE	MLP	4.15	1.17	2.04	7934

Table 5 Results for the number of affected trains (≈ 18,000 records)

Dataset	Method	MSE	MAE	RMSE	Records
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

dataset 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 furthers understanding of how preprocessing methods impact model accuracy and efficiency. A clear correlation exists between preprocessing types and model performance, particularly with

MSE and MAE metrics. The effectiveness of techniques like erasing and winsorization, based on IQR and SD, depends on the dataset size.

In smaller datasets (around 8,000 records), IQR-based outlier erasing is efficient in CNN and RNN models. For example, IQR erase in CNN yielded 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 IQR is more beneficial for smaller datasets in neural network models. In contrast, for larger datasets (about 18,000 records), winsorized standard deviation suits RF and MLP models better. For instance, 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 efficacy for larger datasets in complex data models.

Additionally, incorporating techniques like Chi-square, LASSO, and RFE, with or without OHE, adds complexity to the analysis. For instance, 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.

4.2 Total number of minutes of delay

The attributes average daily train count, betweenness centrality measure, closeness centrality measure, and geographical control point are the most prominent, each featured in twelve datasets, suggesting their constant presence and potentially higher relevance. Subsequently, attributes closeness centrality measure, train headway time, daily passenger count at control point, total kilometers per train, and Rolling stock ID, present in eleven datasets each, follow in importance. These attributes demonstrate significant utility across various analyses. The attribute geographic area coverage, found in ten datasets, indicates moderate relevance. The attributes wind speed in km/h, hour of day, day of week, and month of year, mentioned in eight datasets each, and the attributes Incident cause ID and temperature in Celsius, appearing in seven datasets, imply considerable importance but are not as central as the more frequently occurring variables. This distribution might suggest that there is no direct correlation between frequency of occurrence and impact. Finally, the attributes alternative transportation and total suppressions are the least frequent, appearing in only three and two datasets, respectively, indicating their limited use or specific application in the studied context.

4.2.1 Hyperparameters

MPL: Our multilayer perceptron neural network has three dense layers. The first layer, with 128 neurons, uses L1 regularization to avoid overfitting. The second layer has 64 neurons, both employing ‘relu’ activation. Dropout did not significantly improve the model. It ends with a linear output layer suitable for regression. The dataset was split into 80% training and 20% testing.

RNN: The RNN model comprises two SimpleRNN layers (64 and 32 units) with ‘relu’ activation, capturing temporal data dependencies and ending with a linear output layer. The data split is 80% for training and 20% for testing.

CNN: Our CNN features 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 split is 80% training and 20% testing.

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.

RF: The random forest model includes 200 trees using the 'Poisson' criterion, selected for our data's 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, capturing its distribution effectively. A fixed random seed of 42 was applied to ensure consistent and reproducible results.

RF-Extra: 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. We use a max depth of 20, minimum sample split of 5, and a minimum samples leaf of 4 for all random forest models.

4.2.2 Results

Table 6 with the results show that in smaller datasets of approximately 8000 records, RF models with various preprocessing methods like IQR erasure are effective. For example, RF with IQR erasure and LASSO records an MSE of 64.19 and an MAE of 5.706, indicating good performance in predicting the number of minutes of delay. In contrast, the CNN model with IQR erasure and RFE (recursive feature elimination) has a slightly lower 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 7 shows the larger datasets of approximately 18,000 records; RF models with IQR winsorized and RFE variants show varied results. For instance, RF with IQR winsorized and Chi-square has an MSE of 68.23 and an MAE of 5.9, showing

Table 6 Results for the total number of delays (≈ 8000 records)

Dataset	Method	MSE	MAE	RMSE	Records
IQR erase with LASSO and OHE	RF	64.13	5.665	8.008	7958
IQR erase with LASSO	RF	64.19	5.706	8.012	7958
IQR erase with Chi-square	RF	65.27	5.750	8.079	7958
IQR erase with Chi-square and OHE	RF	65.48	5.724	8.092	7958
IQR erase with RFE and OHE	RF	73.78	6.065	8.589	7958
IQR erase with RFE	CNN	74.77	5.333	8.647	7958
SD erase with LASSO	RF	223.07	9.540	14.94	7820
SD erase with Chi-square	RF	223.26	9.540	14.940	7820
SD erase with LASSO and OHE	RF	223.46	9.460	14.950	7820
SD erase with Chi-square and OHE	RF	223.55	9.450	14.950	7820
SD erase with RFE and OHE	RF	249.96	10.07	15.810	7820
SD erase with RFE	RF	250.92	10.10	15.840	7820

Table 7 Results for the total number of delays (≈ 18,000 records)

Dataset	Method	MSE	MAE	RMSE	Records
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	1061.37	17.691	32.579	17,868
SD winsorized with Chi-square and OHE	RF	1063.06	17.523	32.605	17,868
SD winsorized with LASSO	RF	1117.30	18.569	33.426	17,868
SD winsorized with LASSO and OHE	RF	1119.40	18.433	33.458	17,868
SD winsorized with RFE and OHE	RF	1186.83	18.659	34.450	17,868
SD winsorized with RFE	RF	1187.46	18.692	34.460	17,868
Original data with LASSO selection and OHE	RF	5307.02	22.840	72.850	17,868
Original data with Chi-square selection and OHE	RF	5320.49	22.860	72.940	17,868
Original data with OHE	RF	5324.62	22.840	72.970	17,868
Original data with Chi-square selection	RF-Extra	6319.07	23.790	79.490	17,868
Original data with LASSO selection	RF	6343.70	23.410	79.650	17,868
Original data	RF-Extra	6547.90	23.835	80.919	17,868
Original data with RFE	RF-Extra	7091.64	27.870	84.210	17,868
Original data with RFE and OHE	RF-Extra	7687.87	27.820	87.680	17,868

moderate efficacy in larger-scale predictions. RF-Extra models with original data exhibit much higher MSEs, like 6547.90 for RF-Extra Original, indicating reduced performance due to data complexity and volume.

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

4.3 Number of affected passengers

The primary attributes in 11 datasets are railway track presence, daily passenger count, geographic area, and geographical control point identification. Following these are betweenness centrality, train headway time, day of week, and train service group classification in ten datasets. Less prevalent are average daily train count, closeness centrality, and total kilometers per train, present in nine datasets. Incident cause ID and rolling stock ID, found in eight datasets, have a slightly lower relevance. The hour of day and temperature in Celsius appear in six and five datasets, respectively. Attributes like alternative transportation, wind speed, and month of year, noted in four datasets each, indicate a specific focus. The least common attributes, total suppressions, is only in two datasets.

4.3.1 Hyperparameters

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

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

RNN: Our RNN model includes an 80-unit SimpleRNN layer and a 'softmax' dense layer. It shares the MLP's compilation and training settings, including EarlyStopping. 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.

RF: The RandomForestClassifier model with 100 trees was trained on the split dataset.

RF-Extra: Similar to RF, the Extra-Trees model, with the same tree count, offers a more randomized approach. We used a max depth of 20, minimum samples split of 5, and a minimum samples leaf of 4 for all random forest models.

Synthetic minority over-sampling technique (SMOTE) was applied to balance class representation in cases of higher values.

4.3.2 Results

Table 8 describes the results for the number of affected passengers.

The data shows 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 a test set expansion from 7,998 to 148,536 samples, highlighting the impact of the test data volume on the model generalization. CNNs exhibit notable accuracy fluctuations based on configuration and test size, especially in class handling, as seen in balanced macro averages. Conversely, RF models also show accuracy growth from 0.64 to 0.96, but differ in other metrics. The most precise RF model has high macro and weighted averages, yet a lower mean area under the receiver operating characteristic curve (receiver operating characteristic (ROC) curve), suggesting class differentiation challenges despite the 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.

5 Discussion

Attributes such as closeness centrality, type of track, time of day, and train service group classification are significantly correlated with efficiency in predicting disruptive events, according to Fabella and Szymczak (2021), who also highlight the importance of weather conditions. CNN and RNN models are effective with smaller datasets, while RF and MLP perform better with larger datasets, as stated

Table 8 Number of affected passengers

Dataset	Method	Accuracy	Macro average (precision, recall, F1-score)	Weighted average (precision, recall, F1-score)	Average ROC area	Class ROC areas	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
Standard Deviation Erase 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 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
Standard deviation 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
Standard deviation 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
Standard Deviation Erase 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]	7998
Standard Deviation Erase 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]	7998
Standard deviation 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
Standard Deviation Erase with LASSO SMOTE Encode	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

Table 8 (continued)

Dataset	Method	Accuracy	Macro average (precision, recall, F1-score)	Weighted average (precision, recall, F1-score)	Average ROC area	Class ROC areas	Records
Standard Deviation Erase 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
Standard Deviation Erase 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]	7998
Standard deviation 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
Standard Deviation Erase 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
Standard deviation winsorized with Chi-square SMOTE Encode	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
Standard deviation 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
Standard Deviation Erase 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
Standard deviation winsorized with LASSO SMOTE Encode	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
Standard Deviation Erase 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
Standard deviation 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

by Grandhi et al. (2021) and Chen et al. (2022). Preprocessing methods such as winsorization and standard deviation impact model results, according to Huang et al. (2020) and Ilalokhoin et al. (2022). Our study reinforces these findings, emphasizing that closeness centrality and average daily train count are crucial for predicting delays, aligning with Chen et al. (2022). For predicting total minutes of delay, we found that the average daily train count, betweenness centrality measure, and headway were prominent attributes. These findings are consistent with Grandhi et al. (2021), which demonstrated that weather variables, such as temperature and winter months, are essential in predicting total delays, with neural networks outperforming other predictive models. Our analysis of affected passengers by delays also showed parallels with previous studies. Attributes such as track presence, daily passenger count, and identification of the geographic control point were identified as crucial, appearing in eleven datasets. These results align with Klumpenhouwer and Shalaby (2022), highlighting the effectiveness of machine learning models like random forest regression and elastic net in predicting delays, identifying signal failures and fatal incidents as influential attributes.

Our approach introduces new predictor variables, not thoroughly explored in previous studies, as described in Table 9. The inclusion of centrality measures, infrastructure characteristics, and environmental factors offers a more robust and contextualized analysis of railway delays. Unlike previous studies that primarily focused on operational and infrastructure variables, our research incorporates a broader range of indicators, enhancing the prediction and management of disruptive events.

The machine learning models used in our study are especially CNN and RNN for smaller datasets, and RF for larger datasets. For example, CNN with IQR erase and Chi-squared achieved an MSE of 1.49 and MAE of 0.84 in smaller datasets, while RF with IQR winsorized and Chi-square obtained an MSE of 1.53 and MAE of 0.93 in larger datasets. These results are consistent with the literature, where neural network and random forest models frequently outperform traditional methods in delay prediction.

The importance of machine learning models is corroborated by Yaghini et al. (2013), who demonstrated the high accuracy and low training time of predictive models, facilitating the minimization of delays and future operational problems. Additionally, Li et al. (2020) highlighted the effectiveness of models based on eXtreme Gradient Boosting (XGBoost) in predicting the number of trains affected by primary delays, reinforcing the applicability of advanced machine learning techniques. A unique aspect of our study is the detailed analysis of different preprocessing techniques and their impact on predictive models. We demonstrated that the effectiveness of techniques like IQR and SD erase, as well as feature selection methods like LASSO and Chi-squared, varies with the dataset size. This level of detail provides valuable insights that can improve the implementation of predictive models in different contexts, an area that, despite its importance, is less emphasized in the existing literature. For example, Gao et al. (2023) demonstrated that the inclusion of dispatch commands and the spatiotemporal propagation relationships of delays significantly improved predictive accuracy, with a coefficient of determination (R^2) of 0.87. These results highlight the importance of sophisticated preprocessing and feature selection techniques to improve the accuracy of predictive models.

Table 9 Contributions of novel variables

Variable	Contribution to the model	Difference in relation to literature
Average daily train count	Increases accuracy in predicting total minutes of delay	Not used by Grandhi et al. (2021) as a main variable
Measure of intermediation centrality	Improves identification of critical points in the rail network	Absent in previous studies as a critical attribute
Interval between trains	Crucial for predicting delays and impact on passengers	Innovative use to combine with daily passenger count
Presence of tracks	Identification of faults and incidents	Not detailed by Klumpenhouwer and Shalaby (2022)
Daily passenger count	Directly correlates with the impact of delays	New when combined with geographic control variables
Identification of geographic control point	Enables region-specific analyses	New approach for geographic data segmentation

The three variables analyzed in this study offer an essential tool for passenger rail transport companies, such as CP, to minimize the uncertainty arising from disruptive events. By studying these variables in an integrated manner, companies can obtain a more holistic and accurate view of the possible consequences and adequately prepare to face them. Predicting attributes related to disruptive events enables CP to assess its risk tolerance and implement contingency plans more effectively. For example, if a particular event causes a disruption in rail service, CP can strategically cancel certain trains, adjust train frequency sequences, or use replacement rolling stock. Additionally, understanding the impact of a disruptive event allows CP to activate alternative transportation options, such as buses, to ensure passengers reach their destinations with little impact. This approach can enhance passenger satisfaction by reducing complaints and increasing revenue by avoiding passenger refunds. Alternatively, when feasible, passengers can be redirected through existing services unaffected by the disruptive event. As CP is responsible for guaranteeing public service, it may face penalties from the state regulator in cases of non-compliance with commercial offerings.

Our conclusions, essential for the operator CP, highlight the limitations of our study in terms of scope and data. Although the data source follows the International Union of Railways standard 450, a framework developed by the International Union of Railways (UIC) to promote cooperation among railway operators, infrastructure managers, and other stakeholders, and focused on performance evaluation of the network related to railway traffic operation for quality analysis, including delay coding and cause assignment processes. Future investigations should validate the model with data from various operators, testing its universality and flexibility in different contexts. Additionally, merging data from multiple sources can enhance the accuracy and applicability of the model.

6 Conclusion

This study stands out for its innovative and comprehensive approach, integrating a diverse set of attributes that offer a robust analysis. Unlike previous studies, which often limit themselves to operational and infrastructure variables, this study incorporates centrality indicators, infrastructure and rolling stock damage, environmental and temporal factors, and human and social variables such as the number of injuries and fatalities recorded. Additionally, the study uses real data specific to the Portuguese rail operator Comboios de Portugal (CP), allowing for a contextualized analysis of the Portuguese reality. The originality of the study is also highlighted by the introduction of innovative predictor variables and a detailed comparison between different evaluation methods, such as convolutional neural networks (CNN), recurrent neural networks (RNN), and decision trees, offering valuable insights for crisis management and contingency planning in scenario simulations. In summary, the article significantly contributes to the existing literature by providing a holistic and contextualized analysis of rail delays, potentially influencing policies and practices in rail operations management and incident response. We explored neural network models and decision trees, identifying

key variables and evaluating the models' effectiveness across various data scenarios. For the attributes of delayed trains, CNN and RNN models performed equally well. The RF method yielded the best results for total delay minutes, while CNNs were most efficient for predicting passenger numbers, achieving superior outcomes. The findings highlight the importance of attributes such as closeness centrality measure, railway track presence, hour of day, and train service group classification in predicting delays and impacts on railway services. Our work finds notable parallels in the existing literature. For example, Fabella and Szymczak (2021) examined the vulnerability of the German railway network to natural hazards, highlighting the relevance of variables such as railway track presence and hour of day. This observation strongly resonates with our findings, underscoring the importance of these factors in predicting delays due to natural disasters. Additionally, the approach of Grandhi et al. (2021) in structuring impact estimates of disruptive events is reflected in the efficacy we observed in neural network models, such as CNN and RNN, for smaller datasets. This consistency extends the applicability of these models in various railway contexts. The study of Chen et al. (2022) on delays in Hong Kong's Mass Transit Railway also underscored the importance of variables like weather conditions and types of tracks, aligning with our identification of similar variables and reinforcing the universality of these factors in delay prediction. Lastly, the approach of Huang et al. (2020), which used a hybrid model based on Bayesian networks to predict consequences of disruptions, highlights the inherent complexity of modeling these events. Our study complements this perspective by demonstrating the effectiveness of specific preprocessing techniques and selecting appropriate models for datasets of different sizes.

The implementation of the model developed in this study will bring significant practical benefits. More accurate prediction of the impacts of disruptive events will enable efficient resource management, with a better prioritization of responses and the allocation of alternative transport options. Effective communication of delay information will reduce frustrations and complaints, improving passenger satisfaction (Sogbe et al. 2024). The ability to simulate and predict future impacts will facilitate the planning of preventive strategies, increasing operational resilience. With the reduction of delays and cancellations, CP will avoid financial losses, ensuring a more consistent service. In summary, the model will enhance punctuality, operational efficiency, and passenger experience, promoting more reliable and efficient railway operations. In summary, while our study offers a significant contribution to the field, it is crucial to recognize its limitations to the specific context of CP. Incorporating data from more railway operators and additional variables could enhance understanding of sector dynamics, validate this study's results, and improve predictive models' global applicability in railway operations.

Author contributions The authors confirm contribution to the paper as follows: LM contributed to the study conception and design, literature search and review, research design, methodological approach, and main manuscript writing; SM: supervision of research, research design, reviewing and editing; PR: supervision of research, research design, reviewing and editing.

Funding Open access funding provided by FCT|FCCN (b-on).

Data availability The data from this study can be obtained upon request from the corresponding author, subject to applicable ethical or legal restrictions.

Declarations

Conflict of interest The authors declare that they have no competing interests, financial or non-financial, directly or indirectly related to the work submitted for publication.

References

Acito F (2023) Data preparation. In: Acito F (ed) Predictive analytics with KNIME: analytics for citizen data scientists. Springer, Switzerland, pp 53–83

Artan MS, Sahin I (2022) Exploring patterns of train delay evolution and timetable robustness. IEEE Trans Intell Transp Syst 23(8):11205–11214. https://doi.org/10.1109/TITS.2021.3101530

Belete DM, Huchaiah MD (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. Int J Comput Appl 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. Transp Rev 40:457–478. https://doi.org/10.1080/01441647.2020.1728419

Boateng VA, Yang B (2023) A global modeling pruning ensemble stacking with deep learning and neural network meta-learner for passenger train delay prediction. IEEE Access 11:62605–62615. https://doi.org/10.1109/ACCESS.2023.3287975

Chen X, Ma Z, Li Z (2022) Unplanned disruption analysis and impact modeling in urban railway systems. Transp Res Rec 2676(10):16–27. https://doi.org/10.1177/03611981221088221

Fabella VM, Szymczak S (2021) Resilience of railway transport to four types of natural hazards: an analysis of daily train volumes. Infrastructures 6(12):174. https://doi.org/10.3390/infrastructures6120174

Gao T, Chen J, Xu H (2023) Data-driven train delay prediction incorporating dispatching commands: an XGBoost-metaheuristic framework. IET Intel Transp Syst 18:1777–1796. https://doi.org/10.1049/itr2.12461

Ge L, Voß S, Xie L (2022) Robustness and disturbances in public transport. Public Transport 14:191–261. https://doi.org/10.1007/s12469-022-00301-8

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

Ghaemi N, Zilko AA, Yan F, Cats O, Kurowicka D, Goverde RMP (2018) Impact of railway disruption predictions and rescheduling on passenger delays. J Rail Transp Plan Manag 8(2):103–122. https://doi.org/10.1016/j.jrtpm.2018.02.002

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 BS, Chaniotakis E, Thomann S, Laube F, Antoniou C (2021) An estimation framework to quantify railway disruption parameters. IET Intel Transp Syst 15(10):1256–1268. https://doi.org/10.1049/itr2.12095

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. Transp Res Part C: Emerg Technol 114:338–358. https://doi.org/10.1016/j.trc.2020.02.021

Ilalokhoin O, Pant R, Hall JW (2022) A multi-track rail model for estimating journey impacts from extreme weather events: a case study of Great Britain's rail network. Int J Rail Transp 10(2):133–158. https://doi.org/10.1080/23248378.2021.1891582

Klumpenhouwer W, Shalaby A (2022) Using delay logs and machine learning to support passenger railway operations. Transp Res Rec 2676(9):134–147. https://doi.org/10.1177/03611981221085561

König E (2020) A review on railway delay management. Public Transp 12(2):335–361. https://doi.org/10.1007/s12469-020-00233-1

Leng N, Weidmann U (2017) Discussions of the reschedule process of passengers, train operators and infrastructure managers in railway disruptions. Transp Res Procedia 27:538–544. https://doi.org/10.1016/j.trpro.2017.12.034

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

Li ZC, Wen C, Hu R, Xu C, Huang P, Jiang X (2021) Near-term train delay prediction in the Dutch railways network. Int J Rail Transp 9(6):520–539. https://doi.org/10.1080/23248378.2020.1843194

Marković N, Milinković S, Tikhonov KS, Schonfeld P (2015) Analyzing passenger train arrival delays with support vector regression. Transp Res Part C: Emerg Technol 56:251–262. 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 MJ, Flach P (2021) CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Trans Knowl Data Eng 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 MA, Alemazkoor N, Meidani H (2019) Predicting near-term train schedule performance and delay using bi-level random forests. Transp Res Rec 2673(5):564–573. https://doi.org/10.1177/0361198119840339

Nielsen LK, Kroon L, Maróti G (2012) A rolling horizon approach for disruption management of railway rolling stock. Eur J Oper Res 220(2):496–509. https://doi.org/10.1016/j.ejor.2012.01.037

Sajan, G. V., & Kumar, P. (2021). Forecasting and Analysis of Train Delays and Impact of Weather Data using Machine Learning. 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021, 1–8. https://doi.org/10.1109/ICCCNT51525.2021.9580176

Smiti A (2020) A critical overview of outlier detection methods. Comput Sci Rev 38:100306. https://doi.org/10.1016/j.cosrev.2020.100306

Sogbe E, Susilawati S, Pin TC (2024) Scaling up public transport usage: a systematic literature review of service quality, satisfaction and attitude towards bus transport systems in developing countries. Public Transp. https://doi.org/10.1007/s12469-024-00367-6

Spanninger T, Trivella A, Büchel B, Corman F (2022) A review of train delay prediction approaches. J Rail Transp Plan Manag 22:100312. https://doi.org/10.1016/j.jrtpm.2022.100312

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):4554. https://doi.org/10.3390/math10234554

Tiong KY, Ma Z, Palmqvist CW (2023) Analyzing factors contributing to real-time train arrival delays using seemingly unrelated regression models. Transp Res Part A: Pol Pract 174:103751. https://doi.org/10.1016/j.tra.2023.103751

Wang P, Zhang QP (2019) Train delay analysis and prediction based on big data fusion. Transp Safe Environ 1(1):79–88. https://doi.org/10.1093/tse/tdy001

Yaghini M, Khoshraftar MM, Seyedabadi M (2013) Railway passenger train delay prediction via neural network model. J Adv Transp 47(3):355–368. https://doi.org/10.1002/atr.193