European Transport
Research Review

REVIEW
Open Access

AP-GRIP evaluation framework for data-driven train delay prediction models: systematic literature review

Tiong Kah Yong[1,2*], Zhenliang Ma[3] and Carl-William Palmqvist[1,2]

Abstract

The surging demand for Intelligent Transportation Systems (ITS) to deliver advanced train-related Information for dispatchers and passengers has spurred the development of advanced train delay prediction models. Despite considerable efforts devoted to developing methodologies that can be used to model train operation conditions and produce anticipated train delays, the evaluation strategies for train delay prediction models remain under-researched, particularly evident when accuracy is always found to be the only determinant in model selection. The absence of a standardised evaluation procedure for assessing the effectiveness of these prediction models has hindered the practical implementation of these models. To bridge this gap, the study conducted a systematic literature review on data-driven train delay prediction models and introduced the novel AP-GRIP (Accuracy, Precision, Generalisability, Robustness, Interpretability, Practicality) evaluation framework. The framework covers six key aspects across overall, spatial, temporal, and train-specific dimensions, providing a systematic approach for the comprehensive assessment of train delay prediction models. Each aspect and dimension is thoroughly discussed and synthesised with its definitions, measuring metrics, and important considerations. A critical discussion clarifies several interactions, such as predetermined objectives, desired outputs, model type, benchmark models, and data availability, resulting in a logical framework for assessing train delay prediction models. The proposed framework uncovers inadequate prediction patterns, offering insights on when, where, and why the prediction models excel and fall short, assisting end-users in determining model suitability for specific prediction tasks.

Keywords Train delay prediction, Data-driven, Machine learning, Performance evaluation

1 Introduction

The combination of high capacity utilisation and heterogeneous train traffic poses challenges to the railway system, rendering it susceptible to delays, particularly when multiple trains share limited track infrastructure. As rail operations become more interconnected, a single train exceeding its allotted track occupation time can cause delays to spread across the network, disrupting overall train operations. To optimise capacity utilisation while maintaining service standards, predicting train delays has become a key focus for both academia and industry to improve traffic control quality and mitigate delay propagation throughout the network.

Several review papers focusing on train delay prediction models have been published. For instance, Spanninger et al. [1] reviewed the applicability, including the merits and drawbacks of various techniques (data-driven and event-driven) used to predict train delays. This is followed by Tiong et al. [2], who synthesise a structural

*Correspondence:
Tiong Kah Yong
kah_yong.tiong@tft.lth.se
[1] Department of Technology and Society, Lund University, P. O. Box 118, 221 00 Lund, Sweden
[2] K2 Swedish Knowledge Centre for Public Transport, Bruksgatan 8, 222 36 Lund, Sweden
[3] Department of Civil and Architectural Engineering, KTH, 114 28 Stockholm, Sweden

Springer Open

framework for the development of train delay prediction models using data-driven approaches. The focus on data-driven model development is driven by advancements in technology and the growing availability of data. This can be observed when the data-driven train delay prediction is progressively shifting from statistical approaches [3] to machine learning methods [4, 7], and now with a trendy preference over hybrid models [8, 9], which combine the strengths of multiple approaches with data-driven methods for more robust predictions. Regardless of the significant efforts devoted to developing data-driven models for predicting the mobility of trains in complex railway networks, previous studies paid little attention to the evaluation techniques important for ensuring the effective implementation of the data-driven train delay prediction models in real-world scenarios.

Since model evaluation is a key part of the model development process, Tiong et al. [2] offer a brief introduction to a few basic evaluation components for assessing the performance of data-driven train delay prediction models. However, it lacks sufficient detail, such as selection of metrics, and does not provide comprehensive evaluation components to fully understand the strengths and weaknesses of different models across various scenarios. Considering the pivotal role of data-driven approaches in railway research, we extend the work of Tiong et al. [2] by conducting a systematic review of existing data-driven train delay prediction literature to identify evaluation components that can be employed together to evaluate model performance, subsequently outlining potential model evaluation procedures for train delay prediction. Drawing insights from the existing literature, we introduce the novel AP-GRIP (Accuracy, Precision, Generalisability, Robustness, Interpretability, Practicality) evaluation framework. This framework covers six key evaluation aspects across overall, spatial, temporal, and train-specific dimensions for the comprehensive assessment of train delay prediction models. Each aspect and dimension is thoroughly discussed and synthesised with its definitions, measuring metrics, and important considerations.

The paper makes two main contributions. Firstly, it enhances the evaluation component introduced by Tiong et al. [2] by offering an extended and detailed description of evaluation aspects and analysis levels. Secondly, it introduces the AP-GRIP framework, the first standardised performance evaluation framework available for train delay prediction models. To the best of the authors' knowledge, no other study provides a standardised evaluation framework to guide the selection of the best-performing model, thus addressing a notable gap in the existing literature. The study is expected to make a significant contribution to assist the end-users in systematically assessing model suitability for specific train delay prediction tasks while revealing the inadequacy in the prediction patterns.

The paper is structured as follows: Section 2 outlines the systematic literature review process. Section 3 elaborates on six evaluation aspects within the proposed evaluation framework. Section 4 explores the various levels of analysis. Section 5 offers a detailed overview of the evaluation framework. Section 6 concludes the paper and discusses future research.

2 Systematic literature review

The study is an extension of the work by Tiong et al. [2] and therefore follows a similar systematic literature review methodology adopted in the study, with some refinements to ensure the research remains current. More specifically, the methodology includes employing the structured literature review described by Denyer and Pilbeam [11]. The review process is illustrated in Fig. 1. The search was conducted using Web of Science and Scopus databases, focusing on academic journals and conference papers in English, without publication year restrictions. The literature search was updated in October 2024 to include the latest articles. The keywords related to railway transport, such as "train" and "rail*", were specifically searched in the title to avoid confusion with terms like "training" and "train validation," which could lead to an overwhelming volume of literature from unrelated fields like medicine if searched in the title, abstract, and keywords of the literature. Subsequently, additional searches were conducted in the title, abstract, and keywords of the literature using terms related to train delay prediction, such as "delay", "forecast*" and "predict*", combine with phrases like "data-driven", "machine learning", "regression", "artificial intelligence", "deep learning", "neural network", and "statistical regression" to further narrow down the search towards data-driven approaches. The results from both databases were combined, and duplicated results were eliminated.

All 1039 papers were subjected to a comprehensive full-text review using the predefined set of exclusion and inclusion criteria to ensure their quality and relevance. Articles were excluded when access to full-text versions was unavailable; when the articles were purely qualitative; focused on factors such as rail velocity rather than train events from the timetable perspective; or emphasised purely mathematical perspectives, optimisation, simulation, and queuing theory, given the study's focus on data-driven train delay prediction. At this stage, a total of 48 studies were included. Subsequently, both forward and backward snowballing search strategies, as described by Wohlin [12], were employed. To ensure the inclusion of only relevant articles, the newly identified 31

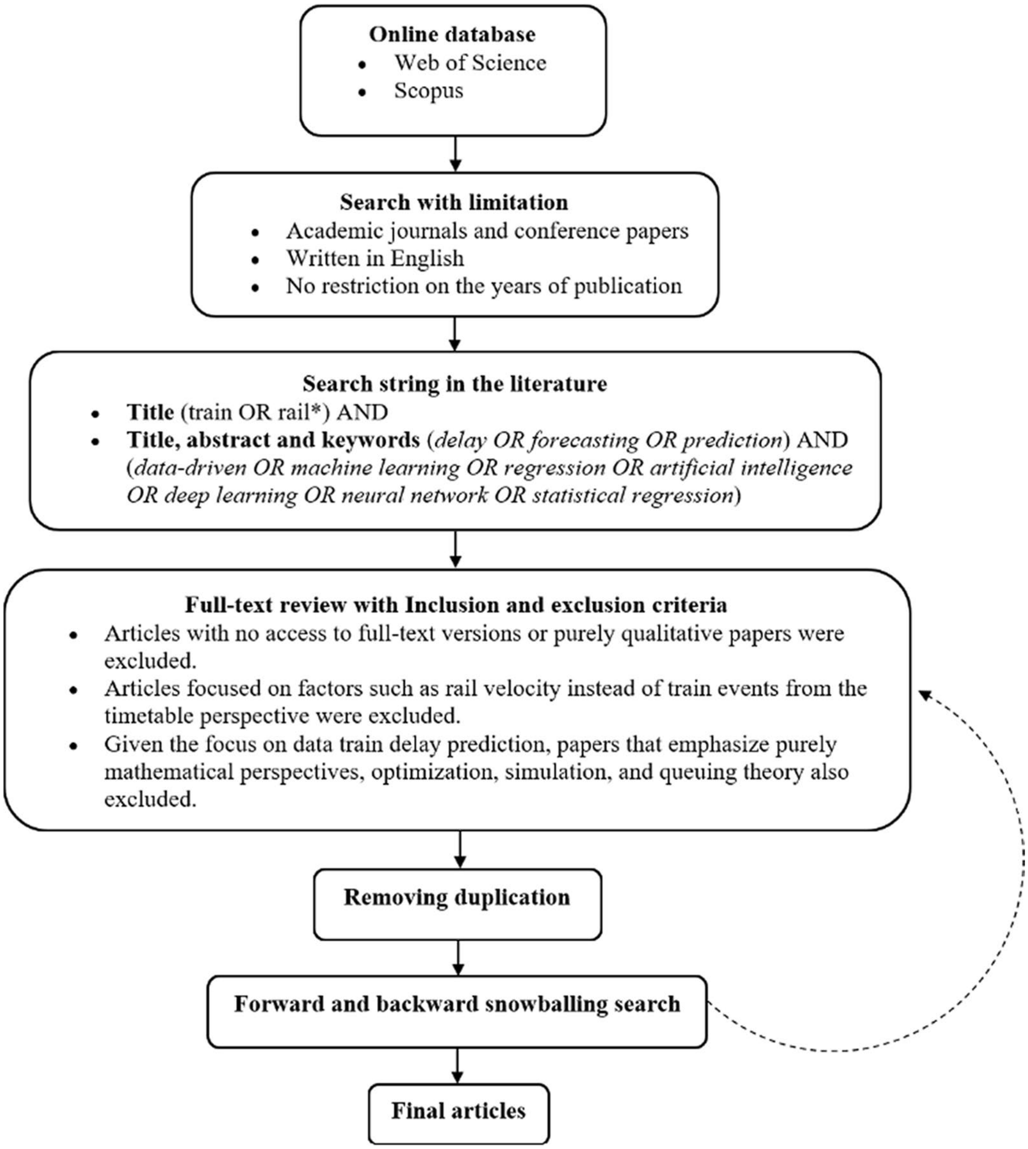
Online database
Web of Science
Scopus
Search with limitation
Academic journals and conference papers
Written in English
No restriction on the years of publication
Search string in the literature
Title (train OR rail*) AND
Title, abstract and keywords (delay OR forecasting OR prediction) AND (data-driven OR machine learning OR regression OR artificial intelligence OR deep learning OR neural network OR statistical regression)
Full-text review with Inclusion and exclusion criteria
Articles with no access to full-text versions or purely qualitative papers were excluded.
Articles focused on factors such as rail velocity instead of train events from the timetable perspective were excluded.
Given the focus on data train delay prediction, papers that emphasize purely mathematical perspectives, optimization, simulation, and queuing theory also excluded.
Removing duplication
Forward and backward snowballing search
Final articles

Fig. 1 Systematic literature review process

papers underwent full-text reading, where compliance with the inclusion and exclusion criteria was assessed, leading to the removal of 12 papers. The process ended at the first iteration when no new paper was added. In total, 67 papers were ultimately included in this study.

3 Evaluation aspects—AP-GRIP

A preliminary review reveals the complexity of evaluating train delay prediction models. To address various aspects of this complexity, the paper broadly categorised evaluation into six key aspects, as shown in Fig. 2. The metrics and approaches employed by existing literature for each evaluation aspect are detailed in Table 1–2, where an

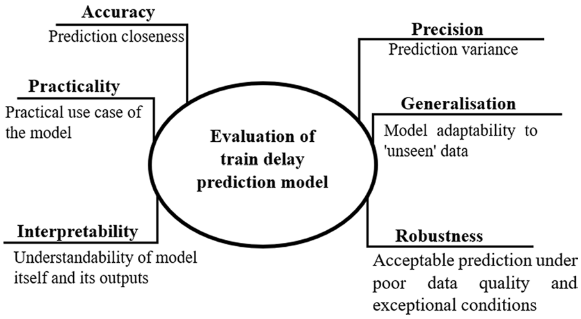
Accuracy
Prediction closeness
Practicality
Practical use case of the model
Interpretability
Understandability of model itself and its outputs
Evaluation of train delay prediction model
Precision
Prediction variance
Generalisation
Model adaptability to 'unseen' data
Robustness
Acceptable prediction under poor data quality and exceptional conditions

Fig. 2 Layered analysis of evaluation aspects

Table 1 Assessment approaches by literature for the relevant evaluation aspects (Accuracy, Precision, Generalisation, and Robustness)

Author	Accuracy				Precision	Generalisation		Robustness
	Scale-dependent measures	Percentage error measures	Relative error-based measures	Others		Data fitting	Model validity	
Vafaei and Yaghini [72]	RMSE					IV: TTS		
Wang et al. [73]	RMSE, MAE					IV: TTS; EV:GV		
Boateng and Yang [69]	MAE, RMSE					IV: TVTS		
Wu et al. [19]	RMSE, MAE, MedAE	SMAPE			Boxplot of RMSE, MAE, MedAE	IV: TVTS; EV:TV		Robustness to data quality
Pineda-Jaramillo et al. [74]	RMSE, MAE	MAPE				IV: CV, TTS	GF	
Tiong et al. [75]	RMSE							
Gao et al. [76]	MSE, ME					IV: TTS		Robustness to different prediction tasks, number of front trains, train selection schemes
Luo et al. [78]	MAE, RMSE				Mean, STD	IV: TVTS	RD	
Wiese [79]	RMSE					IV: TTS	Plot of predicted vs actual	
Gao et al. [77]	RMSE, MAE				Boxplot of RMSE	IV: TTS		Robustness to long delays
Li et al. [10]	MSE, RMSE, MAE					IV: CV, TTS	GF, RD, CDF	
Liu et al. [44]	MAE	MAPE				IV: CV		
Luo, Peng, et al. [105]	MAE, RMSE					IV: CV, TTS	GF	
Wu et al. [20]	RMSE, ME, MAE	SMAPE				IV: CV, TTS		Robustness to different train traffic demand corresponding to night time, daytime, and AM peak
Meng et al. [42]	RMSE	MAPE, APE			Mean, median	IV: CV, TTS		Robustness to different delay sizes, incident
Liu et al. [43]	MAE					IV: CV, TTS	RD	Robustness to different sample sizes
Huang et al. [8]	MAE, RMSE	MAPE					CDF	Robustness to different delay clusters
Tiong et al. [81]	RMSE, MAE					IV: TTS		
Lapamonpinyo et al. [82]	RMSE, MAE					IV: TTS		
Luo, Huang, et al. [80]	MAE, RMSE				Boxplot of prediction error	IV: TVTS	RD, GF	
Wang [83]	MAE, RMSE					IV: CV, TTS		

Table 1 (continued)

Author	Accuracy				Precision	Generalisation		Robustness
	Scale-dependent measures	**Percentage error measures**	**Relative error-based measures**	**Others**		**Data fitting**	**Model validity**	
Klumpenhouwer and Shalaby [84]	MAE, MSE, NRMSE					IV: CV, TTS		
Kusonkhum et al. [85]		MAPE, Accuracy				IV: TTS		Robustness to passenger demand, rain
Tiong et al. [86]	RMSE					IV: TTS		
Chen et al. [45]	MSE, RMSE, MAE					IV: CV, TVTS	GF	Robustness to different weather, peak hour
Bao et al. [46]	RMSE, MAE					IV: TTS		Robustness to different data size, delay sizes
Laifa et al. [87]	RMSE, MAE					IV: TTS		
Rößler et al. [53]	MAE					IV: CV		
Shi et al. [88]	MAE, RMSE	C1			Min, Mean, Max, LB, UB, Median	IV: CV, TTS	GF, RD, CDF	Robustness to different delay sizes, delay scenario
Grandhi et al. [49]	RMSE					IV: TTS	RD	
Wu et al. [63]	MAE, RMSE	MAPE				IV: TTS		
Zhang et al. [65]	MAE, RMSE					IV: CV, TTS	RD	
Pradhan et al. [89]	MSE							
Huang, Wen, Fu, Peng and Li [91]	MAE	MAPE			Boxplots of actual and predicted delays, prediction errors	IV: CV, TTS	RD	
Li et al. [6]	MAE, RMSE	Lessthan				EV:TV		Robustness to different delay sizes
Li et al. [4]	RMSE, MAE					IV: TTS		
Huang, Wen, Fu, Lessan, et al. [90]	MAE	MAPE		ROC, AUC		IV: CV, TVTS	RD, CDF	Robustness to different delay sizes
Taleongpong et al. [7]	RMSE, MAE	MAAPE, AADAE					GF	
Li et al. [10]	MAE	Accuracy, MAPE, Lessthan		ROC, AUC		IV: TTS; EV:TV		
Huang, Wen, Fu, Peng and Tang [92]	RMSE, MAE					IV: CV, TVTS		Robustness to different data size, data dimension, delay trains
Gao et al. [58]	MAE					IV: CV, TTS	CDF	
Oh et al. [35]		Accuracy				IV: TTS	GF	Robustness to peak hour

Table 1 (continued)

Author	Accuracy				Precision	Generalisation		Robustness
	Scale-dependent measures	Percentage error measures	Relative error-based measures	Others		Data fitting	Model validity	
Shi and Xu [47]	MAE, RMSE					IV: CV, TTS	GF	Robustness to delay scenario
Ji et al. [61]	MSE, RMSE					IV: TTS		
Wen et al. [34]	MAE, RMSE					IV: CV, TTS; EV: GV	RD, GF	Robustness to peak hour
Nair et al. [70]	RMSE	C						Robustness to different delay sizes, time horizon, train services, train classes, and operational status
Nabian et al. [50]	RMSE					IV: CV, TVTS		
Wang and Zhang [93]						EV:TV	Boxplots of actual vs predicted delays	
Mou et al. [29]	RMSE, MAE					IV: TTS	CDF, RD, GF	
Jiang et al. [94]	RMSE					EV:TV	RD	
Barbour, Martinez Mori, et al. [22]			RMAE		Min, Mean, Max percent improvement over the baseline	IV: CV		
Lulli et al. [68]	Average accuracy							
Oneto et al. [95]	Average Accuracy							
Ghaemi et al. [96]				ROC		IV: TTS	GF	
Barbour, Samal, et al. [104]			RMAE		Mean, Max percent improvement over the baseline	IV: CV, TTS		
Jiang et al. [97]	MSE					IV: TTS	GF, RD	
Wen et al. [31]	RMSE						F-test, and t-tests, RD	
Oneto et al. [98]	Average accuracy							
Lee et al. [51]		Accuracy				IV: CV, TVTS		Robustness to different delay scenario
Li et al. [99]	RMSE	MAPE				IV: TVTS	GF	
Oneto et al. [100]	Average accuracy					IV: CV, TTS		
Marković et al. [27]	MSE				CI of R2	IV: CV, TVTS	ANOVA, RD	
Kecman and Goverde [30]	MSE				Boxplots of prediction errors	IV: CV, TTS	ANOVA, RD	
Pongnumkul et al. [101]	MAE							

Table 1 (continued)

Author	Accuracy				Precision	Generalisation		Robustness
	Scale-dependent measures	Percentage error measures	Relative error-based measures	Others		Data fitting	Model validity	
Yaghini et al. [102]		Accuracy		C		IV: CV, TVTS; EV:TV		Robustness to different input encoded
Gorman [3]	MSE, MAE, ME	MAPE						
Peters et al. [103]	MSE							

Accuracy: MAE = Mean Absolute Error; RMSE = Root Square Mean Error; MAPE = Mean Absolute Percentage Error; MSE = Mean Square Error; ROC = Receiver Operating Characteristics; AUC = Area Under the Curve; C = Forecast correctness; SMAPE = Symmetric Mean Absolute Percentage Error; APE = Absolute Percentage of Error; NRMSE = Normalised Root Square Mean Error; ME = Mean Error; RMAE = Relative Mean Absolute Percentage Error; MedAE = Median Absolute Error

Precision: Min = Minimum; Max = Maximum; UB = Upper Boundary; LB = Lower Boundary

Generalisation: IV: Internal Validation; CV: Cross Validation; TVTS: Train-validate-test Split; TTS: Train-test Split; EV: External Validation; GV = Geographical Validation; TV = Temporal Validation; ANOVA = Two-way analysis of variance; GF = Goodness-of-fit plots; RD = Residual Distribution Plot; CDF = Cumulative Distribution Functions

Table 2 Assessment approaches by literature for the relevant evaluation aspects (Interpretability and Practicality) and level of analysis

Author	**Interpretability**	**Practicality**	**Level of analysis**					
	Explanation (1)	**Explanation (2)**	**Applicability**	**Consistency (level)**	**Overall**	**Spatial (level)**	**Temporal (interval)**	**Tran-specific (level)**
Vafaei and Yaghini [72]					✓	✓ (stn)		
Wang et al. [73]					✓			
Boateng and Yang [69]					✓			✓ (train id)
Wu et al. [19]					✓			
Pineda-Jaramillo et al. [74]		SHAP			✓			
Tiong et al. [75]		SUR				✓ (stn)		
Gao et al. [76]					✓			
Luo et al. [78]						✓(train line, stn)		
Wiese [79]	FI				✓			
Gao et al. [77]		SHAP			✓			
Li et al. [10]						✓(train line)	✓ (5 min)	
Liu et al. [44]						✓(Area)		
Luo, Peng, et al. [105]						✓(train line)		
Wu et al. [20]					✓	✓(stn)		
Meng et al. [42]				✓ (15 min)	✓			
Liu et al. [43]					✓			
Huang et al. [8]						✓ (train line)		
Tiong et al. [81]						✓(stn)		
Lapamonpinyo et al. [82]	FI				✓	✓(stn)		
Luo, Huang, et al. [80]						✓(stn)		
Wang [83]			RMSW		✓			
Klumpenhouwer and Shalaby [84]	FI				✓			
Kusonkhum et al. [85]						✓ (stn)		
Tiong et al. [86]						✓ (stn)		
Chen et al. [45]	FI				✓			
Bao et al. [46]	FI				✓	✓ (stn)		
Laifa et al. [87]					✓			
Rößler et al. [53]		SHAP			✓			
Shi et al. [88]	FI					✓(stn)		
Grandhi et al. [49]	FI				✓			
Wu et al. [63]					✓			
Zhang et al. [65]					✓			
Pradhan et al. [89]					✓			
Huang, Wen, Fu, Peng and Li [91]				✓ (5 min)		✓ (stn)		
Li et al. [6]	FI							
Li et al. [4]					✓			

Table 2 (continued)

Author	Interpretability	Practicality	Level of analysis					
	Explanation (1)	**Explanation (2)**	**Applicability**	**Consistency (level)**	**Overall**	**Spatial (level)**	**Temporal (interval)**	**Tran-specific (level)**
Huang, Wen, Fu, Lessan, et al. [90]	SA					✓ (train line, stn)	✓ (5 min)	
Taleongpong et al. [7]		SHAP		✓ (1 step)	✓			
Li et al. [10]	FI				✓	✓ (stn)		
Huang, Wen, Fu, Peng and Tang [92]						✓ (train line, stn)		
Gao et al. [58]			Prediction accuracy under different allowance errors		✓			
Oh et al. [35]					✓			
Shi and Xu [47]					✓	✓ (stn)		
Ji et al. [61]					✓			
Wen et al. [34]						✓ (stn)		
Nair et al. [70]								
Nabian et al. [50]	FI				✓			
Wang and Zhang [93]					✓			
Mou et al. [29]						✓ (stn)		
Jiang et al. [94]	FI				✓			
Barbour, Martinez Mori, et al. [22]	FI				✓	✓ (stn)		
Lulli et al. [68]						✓ (stn)	✓ (day)	✓ (train order, categories)
Oneto et al. [95]					✓	✓ (stn)		✓ (train order)
Ghaemi et al. [96]				✓ (1 time point)				
Barbour, Samal, et al. [104]					✓			
Jiang et al. [97]					✓			
Wen et al. [31]		LR			✓			
Oneto et al. [98]						✓ (stn)		✓ (train order)
Lee et al. [51]					✓			
Li et al. [99]		LR			✓			
Oneto et al. [100]						✓ (stn)		✓ (train order)
Marković et al. [27]					✓			
Kecman and Goverde [30]	FI	LR			✓	✓ (stn)		
Pongnumkul et al. [101]						✓ (stn)		✓ (train number)
Yaghini et al. [102]					✓		✓ (year)	
Gorman [3]		LR			✓	✓ (stn)		✓ (train priority)

Table 2 (continued)

Author	Interpretability	Practicality	Level of analysis					
	Explanation (1)	Explanation (2)	Applicability	Consistency (level)	Overall	Spatial (level)	Temporal (interval)	Tran-specific (level)
Peters et al. [103]						✓ (stn)		✓ (train number)

Interpretability: Expln: Explanation; FI = Feature Importance; LR = Linear Regression Coefficient; SHAP = SHapley Additive exPlanations; SUR: Seemingly Unrelated Regression; SA = Sensitivity Analysis

Level of analysis: Stn: Station

empty box signifies that the respective evaluation aspect is not addressed in the relevant studies. The subsequent subsections provide in-depth discussions of each evaluation aspect.

3.1 Accuracy

Model accuracy is evaluated by analysing residual errors, which quantify the closeness between observed and predicted values [13]. Hyndman and Koehler [14] classified accuracy measures into scale-dependent measures, percentage error-based measures, and relative error-based measures. In train delay prediction studies, the selection of accuracy measures is often subjective without adequate justification. However, Wallström [15] emphasised the need to consider factors such as differences in scale, resistance to outliers, and relevance to decision-making when choosing appropriate measures.

Scale-dependent measures are accuracy measures influenced by the scale of the data. As shown in Table 1, the commonly used measures in train delay prediction studies are scale-dependent measures, specifically Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Both RMSE and MAE are sensitive error measures, but RMSE penalises larger errors more heavily due to its squared error term. Median error measures like Root Median Squared Error (RMdSE) and Median Absolute Error (MedAE) are less affected by outliers and provide a more robust assessment of prediction accuracy compared to mean error measures [16]. For example, in a dataset with measurements of 6.57, 6.63, 6.59, 6.66, and an outlier of 65.60, the mean is 18.41, skewed by the outlier. In contrast, the median remains reasonable at 6.59. However, Armstrong and Collopy [17] found median error measures less informative due to their low sensitivity. Besides being easy to interpret, scale-dependent measures provide meaningful insight into the error magnitude, allowing decision-makers to understand the practical implications of prediction errors [17]. For instance, an RMSE of 2 min suggests a 2-min adjustment when rescheduling the trains.

Percentage error measures compute the prediction error as a percentage of the actual value, denoted by $p_k = \frac{100e_k}{y_k}$, where e_k is the prediction error and y_k is the actual value. These measures are scale-independent, facilitating comparison of prediction performance across datasets. The limitations of these measures include the potential for infinite values when actual values are zero and skewed distributions when actual errors approach zero. Mean Absolute Percentage Error (MAPE) is the widely used percentage error measure in data-driven train delay prediction studies (see Table 1). Armstrong and Collopy [17],Makridakis [18] criticised the asymmetric treatment of prediction errors in MAPE, whereby overestimations are penalised more heavily than underestimations. To address this limitation, modified measures like Symmetric Mean Absolute Percentage Error (SMAPE) are used by Wu et al. [19] and Wu et al. [20] for evaluating train delay prediction models. Due to the nature of the train delay predictions involving near-zero values, Taleongpong et al. [7] advised against using MAPE and SMAPE. Instead, they proposed the adoption of Mean Arctangent Absolute Percentage Error (MAAPE), introduced by Kim and Kim [21]. However, percentage error measures are less practical for decision-making. For instance, a 20% prediction error must be translated into minutes for actionable rescheduling since train operators require the exact magnitude of minutes to make adjustments to operational plans rather than percentage errors.

Relative error-based measures such as Geometric Mean Relative Absolute Error (GMRAE) and Median Relative Absolute Error (MdRAE) are scale-independent measures that compare the prediction errors of a specific model to those of a benchmark by computing the ratio, $r_k = \frac{e_k}{e_k^*}$, where e_k is the prediction error for the model and e_k^* is the prediction error from the benchmark method. Relative error-based measures are unit-free and resilient against outliers. However, they can become problematic when the benchmark model has zero prediction errors, resulting in infinite values, or when the prediction errors are close to zero, causing skewed distributions. Relative error-based measures are easy to interpret. For example, MdRAE > 1 indicates that the evaluated model outperforms the benchmark. Barbour, Martinez Mori, et al. [22] used relative measures to show the improvement of train delay prediction models over a baseline and uncover the advantages of incorporating network traffic state features over simple counts of network traffic. However, relative error-based measures are less practical for decision-making in train operations since they fail to quantify the necessary adjustments in train operations to rectify delays while considering inherent model errors.

As the number of studies on the development of advanced prediction models grows, adopting scaled-independent measures that provide standardised units of prediction error is crucial for enabling comparisons across studies with different datasets and scales. This approach ensures continuous improvement in prediction models over time, given the challenge of replicating prediction algorithms from different studies. However, it is worth noting that most existing train delay prediction studies primarily focus on developing new models and typically compare them against simple baseline models, wherein all models utilise the same dataset and unit of measurement. This justifies the continued preference for scale-dependent measures in the field.

3.2 Precision

Precision refers to the statistical variance [23] or the spread of data [24]. It reflects the uncertainty surrounding the predicted values [25]. Thus, precision is essential in assessing prediction models by quantifying the underlying prediction uncertainty.

Walther and Moore [23] state that measures of variability, such as the range, variance, or standard deviation, can be used as a precision measure. The interquartile range is considered a more robust precision measure as it is less sensitive to extreme values [26]. Marković et al. [27] validated ANOVA results by examining the confidence interval of R2 in prediction models, with a narrow confidence interval indicating greater reliability and precision. However, unscaled precision measures are often difficult to compare across studies. Scaled precision metrics, like the coefficient of variation (CV), recommended by Mohammadnazar et al. [28] for vehicle mobility prediction, face challenges in train delay prediction. This limitation arises from frequent near-zero delays, causing the CV to be highly sensitive to minor mean fluctuations, particularly when the mean of train delays in the denominator is zero or negative, misleading the CV results. Conversely, graphical methods, particularly boxplots, are commonly used to assess precision in train delay prediction studies. For instance, boxplots were utilised by Mou et al. [29] to compare the distribution of actual and predicted delays, by Huang, et al. [91]; Kecman and Goverde [30] to visualise the distribution of prediction errors,and by Marković et al. [27],Wen et al. [31] to display the distribution of MSE and R^2 respectively.

Precision reflects the level of uncertainty in the prediction errors, with narrower ranges signifying higher reliability of the predictions generated by the predictive model [25]. As demonstrated in Table 1, precision is often overlooked, with basic descriptive statistics such as minimum, maximum, and mean errors often being explored to provide a fundamental understanding of model precision. However, these metrics are sensitive and easily influenced by outliers. Therefore, a comprehensive evaluation that includes robust precision measures is necessary to ensure the credibility of the prediction.

3.3 Generalisability

Generalisation, according to Denyer and Pilbeam [11], is the ability of a predictive model to make accurate predictions on previously unseen data. It is impeded by two main factors: 1) data fitting, where the predictive model is closely tailored to the training data and performs poorly on new data,and 2) model validity, a mismatch between the complexity of the model's hypothesis and the data characteristics [32].

According to Van Calster et al. [33], train delay prediction models should undergo train delay prediction models both internal and external validation, using metrics similar to those applied in accuracy evaluation (e.g. RMSE, MAE). Internal validation entails validating the model on data that is used during its development. It ensures the model is not overfitting and remains valid under similar train operation conditions as those used during model development. Table 1 reveals that the majority of train delay prediction studies conduct internal validation, with the most commonly used techniques being train-test split and cross-validation. On the other hand, external validation refers to validating the model on data that is different from the data used for model development. It verifies the model's transportability across diverse train operation conditions. This can be accomplished through temporal validation, which involves using data collected at a different time point but at the same location, or through geographic validation, which involves using data collected at a different location. Li et al. [5] conducted temporal validation by training and validating their train delay prediction model using train operation data from March 2015 to November 2016, and testing the model on data from 2018. Similarly, Wen et al. [34] performed geographic validation by training the model on data from the Rotterdam Central to Dordrecht section of the Dutch railway system and evaluating its performance on the Rilland Bath-Vlissingen section.

Model validity testing is critical for evaluating how well a model aligns with the specific characteristics of the problem [106]. The key criterion for diagnosing the model's adherence to assumptions is that residuals should show a random pattern rather than a systematic trend. It is worth noting graphical methods are commonly employed to validate model validity in train delay prediction studies. Mou et al. [29],Oh et al. [35] plotted observed and predicted train delays, given the strong alignment of data points along the 45-degree diagonal line, suggesting the model can effectively capture underlying patterns and variability in the data. Huang, et al. [91], Wen et al. [34] used residual distribution plots to identify the adequate structured model that displays residuals close to zero. This characteristic is also known as white noise residuals, where the residuals are independently and identically distributed with a zero mean and constant variance. Huang, et al. [90]; Li et al. [10]; used cumulative error distribution plots to assess the distribution of prediction errors, with a high percentage of zero residuals and a significant portion of residuals falling within a small threshold indicating good model performance.

External validation is highly recommended before deploying prediction models for practical application, as it bridges the gap between model development and real-world implementation and ensures that decision-related railway traffic management is not based on incorrect prediction models in realistic scenarios. The evaluation of predictive models for train delays often emphasises prediction accuracy, neglecting the importance of assessing how well the model fits the data. Various error specification tests proposed by Washington [106] are worth considering, such as tests aimed at assessing the fit with respect to serial dependence, neglected non-linearity, serial independence, constancy of variance, and symmetry.

3.4 Robustness

Robustness refers to the system's ability to function correctly in the presence of invalid inputs or under stressful environmental conditions [36]. Since robustness emphasises the events that should *not* happen rather than how the system *should* function under normal circumstances [37], robustness testing is often neglected. Nevertheless, model performance must be systematically assessed in relation to robustness risks to mitigate potential negative impacts.

The evaluation of robustness is critical to ensure that train delay prediction models can function correctly even in the absence of high-quality input data. Real-time train delay prediction models are exposed to various perturbations, like noise, missing values, measurement errors, data drift, and data processing errors, which can result in faults during field application. Intelligent algorithms, particularly machine learning models, are susceptible to adversarial attacks, where small input perturbations can lead to inaccurate predictions [38, 39]. For instance, Wu et al. [19] showcased the superior performance of a train delay prediction model integrated with imputed datasets compared to non-imputed datasets, highlighting the efficacy of their imputation strategy in mitigating the impact of missing values. Table 1 indicates that many train delay prediction studies tend to emphasise the robustness of prediction models when exposed to variations in the size of training data. While the robustness of train delay prediction models in the presence of low-quality input has not been extensively studied, other fields have actively researched this area. For instance, Hines et al. [40] evaluated model robustness by assessing performance with input features corrupted by Gaussian noise, and Fernandez et al. [41] considered noise that resembles naturally occurring distortions.

Robustness also encompasses the model's ability to exhibit acceptable behaviour under exceptional or unforeseen operating conditions [41]. Evaluating the model's robustness involves examining its prediction accuracy in various scenarios, such as days with or without incidents [42], different initial delay lengths [43, 44], different weather conditions [45], as well as different delay durations, particularly extreme delays [46, 47]. This verification process ensures the model's robustness and reliability in providing dependable estimates of train delay times, particularly when non-recurring events occur. The annual Swedish timetable change at year-end, for instance, has the potential to influence the punctuality of train paths both individually (different allocation of running time margin, rolling stock allocation and rotation, stopping patterns, scheduled dwelling time, etc.) and in their interactions with the rest of the traffic (added/removed near-conflicts, different headway buffers at stations, station track allocations, prioritisation rules in dispatching, etc.), thus evaluating the model's performance before and after such timetable changes ensures the robustness of the train delay prediction model to significant system-wide discontinuities.

Shahrokni and Feldt [37] highlighted that many studies are evaluated in controlled settings, which raises concerns about their applicability, resulting in prediction models remaining purely academic contributions without real-world industrial use. To ensure their practical viability, it is crucial to evaluate the robustness of train delay prediction models using datasets with realistic perturbations representative of real-world application scenarios.

3.5 Interpretability

Interpretability in prediction models refers to how well the extent to which users can comprehend the prediction results [15]. Building upon this concept, Doshi-Velez et al. [48] formulated a governing principle for interpretability, emphasising the provision of useful explanations for users to assess the impact of inputs on the output. They suggested at least one of the following explanations should be provided to ensure interpretability in the prediction model:

(1) **Human-interpretable information about the factors used in a decision and their relative weight**. This includes the provision of a comprehensive list of factors considered in the prediction, ordered based on their significance to the output. Feature importance analysis is commonly used in train delay prediction studies, as it is a built-in function in statistical models like linear regression [49], machine learning models, especially tree-based models [4, 45, 50], and neural network models [49]. Evaluating feature importance helps gain insights into the prediction model [4, 45, 50] and enhance the predictive model by removing less important features and retaining those with higher importance [46]. Some studies have also employed sensitivity analysis to study their models at varying levels of complexity [22, 51].

(2) **An answer to a counterfactual question.** When a more comprehensive understanding of a decision is required, merely knowing whether a factor was considered is insufficient. In such cases, counterfactuals delve deeper into the impact of altering inputs on the output and identify the necessary modifications for desired outcomes. These can be achieved either through interpretable-by-design models or post-hoc interpretability methods. The former involves models that are by nature interpretable, such as linear regression, decision trees, and operational rules validated by experts. The simplicity of these interpretable algorithms allows us to understand how they learn from input data, converging towards a solution and identifying the types of relationships involved. For example, based on the coefficients of the linear regression models, Gorman [3] identified primary sources of interference (e.g., meets, passes, overtakes of trains) that have the greatest effect on congestion delays. Interpretable models struggle with high-dimensional data,complex black-box models offer better performance but are less understandable [52]. To address this trade-off, post-hoc interpretability methods where ad hoc techniques (like the Shapley Additive Explanations (SHAP)) are applied to the black-box machine learning model. For example, Taleongpong et al. [7] and Rößler et al. [53] used SHAP to explain complex data-driven train delay prediction models without compromising accuracy.

The interpretability of models is increasingly important for justifying the decisions these models make. With the growing use of machine learning models for train delay prediction, there is a high demand for understanding the mechanism by which the model operates, ensuring the decisions based on prediction models involved a holistic consideration. Furthermore, the explanations (2) derived from model interpretation using domain knowledge are essential for scientific discovery in train operations. More specifically, this type of explanation disentangles the underlying cause of train delays and supports the identification of effective train operation management policies. Despite the availability of various ad hoc techniques in the area of black-box model interpretability, such as Local Interpretable Model-agnostic Explanation (LIME) [54], Anchors [55], Local Rule-based Explanations (LORE) [56], and Model Agnostic Supervised Local Explanations (MAPLE) [57], their adoption in the train delay prediction field remains limited. Therefore, more effort should be dedicated to exploring post-hoc interpretability methods for generating explanations (2) from highly accurate but complex models in order to promote further advancements in the interpretability of train delay prediction models.

3.6 Practicality

Practicality is an application-oriented aspect, considering the task fulfilment capability of the prediction model from the end-user perspective (passengers or operators). The idea of incorporating application-oriented aspects into the evaluation framework is because the tolerance for prediction errors varies depending on the model's use case. For example, the overestimation of the delay prediction is less favoured by passengers since it may increase the risk of missing connections. The traffic control centres, especially in border regions, require accurate prediction for train conflict resolution and to effectively coordinate border crossings and local traffic. Thus, identifying the correct practical requirement is essential for the usefulness of train delay prediction in the real world. Mathematically, the model's practicality is measured using asymmetric prediction error measures by penalizing the end-user unfavourable prediction errors. For example, Weighted Mean Absolute Error (WMAE) can be defined to account for different penalties on prediction errors by incorporating a parameter w_k, as expressed in Eq. 1, respectively. More specifically, if the passenger penalizes the overestimation of the delay prediction, then w_k can be set to a value larger than 1.

$$\text{WMAE} = \frac{1}{N}\sum_{k=1}^{N}|w_k E_k| \tag{1}$$

where N is the sample size, and E_k is the prediction error derived from $y_k - \widehat{y}_k$ with y_k and $\widehat{y}_k$ respectively represent the actual and predicted arrival delays, recorded in minutes. w_k is weight, for instance, $w_k = 5$ if the E_k >5 and 1 otherwise. Another approach, as demonstrated by Gao et al. [58], involves evaluating prediction accuracy under various percentages of allowable errors, establishing a zone of acceptable prediction error before raising concerns about the effectiveness of the train delay prediction model. However, based on Table 2, this aspect of practicality is commonly overlooked in evaluation.

In real-time applications, train delay prediction models continuously generate predictions for downstream stations and update them at regular intervals. The stability of predictions assesses the consistency of the generated predictions at each prediction interval along the train's journey, ensuring the reliability of the train delay prediction model perceived by the end user. Passengers might perceive the predictions as unreliable even before the train reaches their station if they experience significant fluctuations in the information received, for instance, the expected delays of 30 min, then 25 min, followed by 35 min and 30 min. Such inconsistency may cause passengers to disregard the predictions altogether. This aligns with the results reported by Parbo et al. [59] indicating that passengers expressed dissatisfaction with the stress they experienced when the onward planned connections looked doubtful. This aversion may be attributed to the inherent disutility associated with the risk of arriving late, which can be perceived as an anxiety cost, or a decision cost associated with the need for contingency plans. This aspect of uncertainty is a common concern for numerous passengers, as highlighted by Börjesson et al. [60]. Recognising the importance of ahead-of-time states for proactive actions by train operators and preparing passengers for travel plan changes, Huang et al. [91]; Meng et al. [42],Taleongpong et al. [7] tested the train delay prediction model at various prediction intervals, including 5, 10, and 15 min.

Besides evaluating models from a practicality aspect, the technical implementation of these models often lacks practical consideration. For example, there is still no study that explores how to effectively translate train delay prediction results for end users. Without addressing how to communicate predictions effectively, the model's practical impact remains limited, regardless of its accuracy. Furthermore, these models operate within a dynamic environment where train schedules, routes, and external conditions frequently change, necessitating retraining or recalibration to maintain their reliability. Nevertheless, efforts to study standardised procedures that ensure the functionality and trustworthiness of train delay prediction models as operational conditions evolve—such as performance monitoring, periodic retraining, and adapting to changing circumstances—are still very limited. Thus, it is crucial to explore and expand the scope of practicality in this field by delving further research into how these models can be effectively integrated and sustained in real-world applications.

4 Level of analysis

To uncover the prediction pattern of the train delay prediction model in realistic operating scenarios, it is necessary to evaluate model performance across various levels of analysis while considering the aforementioned six aspects. Table 2 categorises the literature based on the levels of analysis they conducted. The overall performance evaluation is the most commonly presented dimension because it offers a fundamental understanding of the overall model quality for the given prediction task. This involves the highest level of data aggregation, that is, aggregating all observations in terms of prediction error measures to provide an overview of model performance. This level of evaluation is commonly used during the initial stage of model comparison to select the best performance model before more detailed analysis, as by Ji et al. [61], Wang and Work [62], Wu et al. [63], Zhang et al. [64].

Since a model with good overall performance may underperform in specific situations, detailed evaluation aids in investigating underlying performance patterns and uncovering circumstances where models excel or fall short. Because train delay prediction is a spatiotemporal problem that provides train movement information across different stations and time periods Zhang et al. [64], evaluating prediction performance must consider both spatial and temporal dimensions. The granularity of evaluation varies depending on the purpose of the evaluation, with spatial evaluation ranging from the network level to the station level and temporal evaluation ranging from weekly intervals to minute intervals. Wang et al. [66] compared the support vector regression model against the benchmark models at various stations and sections of the railway system, whereas Zhang et al. [64] examined five different models' prediction performance over a 12-h prediction horizon.

Besides spatial and temporal dimensions, the train-specific dimension, which involves examining the performance of prediction models for individual trains or train categories, should not be overlooked. The analysis entails aggregating data based on train characteristics such as train frequency, train type (e.g., commuter, regional, high-speed), priority, empty rolling stock movements, and maintenance fleet movements. By evaluating model performance on a train-by-train basis, it is possible to distinguish characteristics that make certain train categories more prone to prediction errors. This allows the identification of where intervention is necessary to enhance the model's performance. For example, freight train operation is typically subjected to considerably greater variability than passenger traffic [67], which is clearly a challenge to most prediction models. Lulli et al. [68] evaluated the accuracy of their train delay prediction model based on train category (e.g., regional, high speed, and freight train), whereas Boateng and Yang [69] assessed the models based on train ID, considering that different train IDs follow distinct routes.

5 Prediction model evaluation framework

Taking into account the aforementioned evaluation dimensions and six aspects, the AP-GRIP evaluation framework divides the train delay prediction model evaluation process into three stages: evaluation objective, evaluation analysis, and evaluation outputs, as illustrated in Fig. 3. The three-stage AP-GRIP framework is designed to provide a comprehensive view of model performance for diverse end-users in the development, evaluation, and deployment of train delay prediction models, including model developer, practitioners, and IT departments. The framework serves as a tool to reveal the strengths and weaknesses of a model in each dimension, helping evaluators become aware of the trade-offs between criteria. This allows stakeholders to make decisions with a full understanding of the trade-offs involved and identify models that best align with their specific goals. It is worth noting that the proposed AP-GRIP framework is flexible and adaptable to accommodate the needs of different decision makers (passengers, train operators, and infrastructure managers), recognising that there is no one-size-fits-all approach. For instance, some may favour simplicity, interpretability, and ease of implementation, even if it means sacrificing some accuracy, while others may prioritise highly accurate models despite their complexity. Rather than imposing a fixed ranking among the dimensions, the framework encourages decision-makers to weigh the importance of each dimension based on their specific context and goals. Overall, the purpose of the AP-GRIP framework is to reveal model performance comprehensively and make these trade-offs visible, fostering a transparent approach to model evaluation and selection.

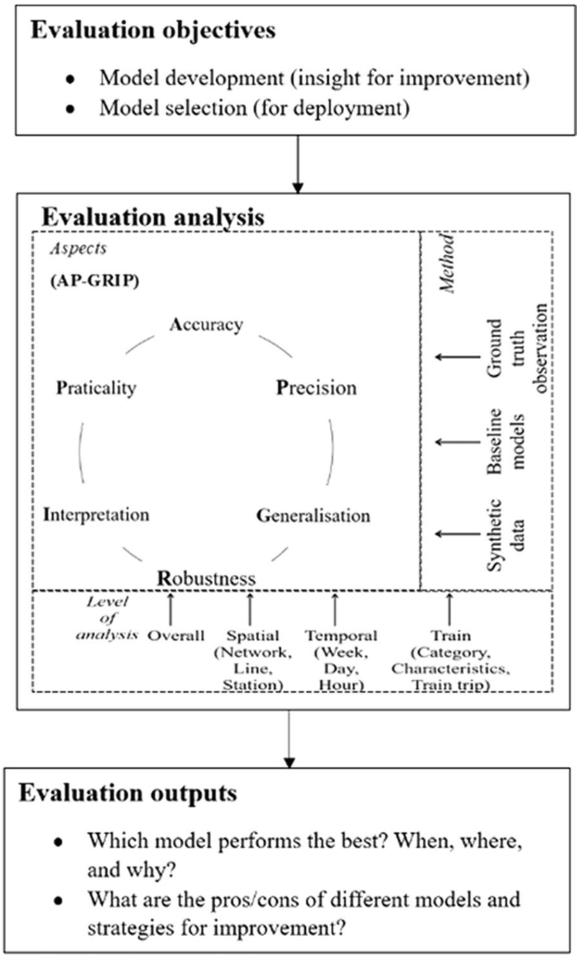
Evaluation objectives
• Model development (insight for improvement)
• Model selection (for deployment)
Evaluation analysis
Aspects
(AP-GRIP)
Accuracy
Praticality
Precision
Interpretation
Generalisation
Robustness
Method
Ground truth observation
Baseline models
Synthetic data
Level of analysis
Overall
Spatial (Network, Line, Station)
Temporal (Week, Day, Hour)
Train (Category, Characteristics, Train trip)
Evaluation outputs
• Which model performs the best? When, where, and why?
• What are the pros/cons of different models and strategies for improvement?

Fig. 3 Logical flow in evaluating train delay prediction model using AP-GRIP evaluation framework

5.1 Evaluation objectives

The first stage is to clearly define the evaluation objective since this is the key enabler for the adoption of the proposed framework. The framework is designed to accommodate two objectives: 1) Model development; and 2) Model selection. *Model development* study aims to build a model with strong predictive capability, seek insight for improvement, and comprehend the model's causal relationships. This type of study focuses on the entire model development process, including data preprocessing, feature engineering, model selection, and model training, with model evaluation included as part of the study, as in Barbour, Martinez Mori, et al. [22],Wang et al. [66],Wen et al. [34]. The model development study emphasises the incorporation of the same input variable into the models to ensure a fair comparison between the models. *Model selection*, on the other hand, aims to identify the most effective algorithms capable of generating accurate results, irrespective of the methods employed or resource accessibility, such as training data. This is particularly relevant for practitioners seeking to assess model efficiency without taking part in the model development process. External evaluators do not have access to the input data and focus on a series of tests performed on the output data generated by the prediction models to determine the best performance model, usually prior to model deployment.

5.2 Evaluation analysis.

The second stage, which involves implementing evaluation analysis, is the core of the proposed evaluation framework. To simplify the complex evaluation analysis process, this stage can be divided into three key components where critical decisions need to be made. The first component is the selection of *evaluation aspects*—AP-GRIP and their corresponding metrics, which must consider the predefined objective, model type, and data availability. For instance, the external evaluator lacks access to training and input data in the model selection study, making assessments like model generalisation (model fitting) and interpretability irrelevant. Similarly, these aspects are not feasible for event-driven models, as these models rely on train-event dependency structure assumptions.

The second component is the selection of the *level of analysis*. Most existing studies typically conduct an overall performance evaluation on the selected evaluation aspect to gain an overview of the models' quality, but they often overlook the importance of conducting detailed evaluations from multiple dimensions, including temporal, spatial, and train-specific dimensions. Since a model with good overall performance might fail to perform in certain circumstances, a detailed evaluation is necessary to reveal any deficiencies in the prediction patterns.

The third component is the determination of the appropriate *method* for assessing the prediction model based on the chosen aspects and the level of analysis. The most prevalent approach involves comparing the predicted values with the ground truth observations, specifically the actual train delays, which is the simplest and most direct way to measure the performance of the prediction model. In train delay prediction, the ground truth data is a long-tail distribution data, where the majority of actual train delays are either minimal or no delays [70]. Consequently, the assessment of model performance based solely on ground truth data may not adequately account for scenarios involving substantial delays, which is a matter of concern to various stakeholders. Therefore, testing with synthetic data generated through simulation that reflects the rare and infrequent scenarios generated through simulation may be essential for a comprehensive assessment of the model's capability to respond to these critical situations.

Methodologically, comparisons of prediction performance between dissimilar methods provide evidence of a model's superiority over others. Since there are no universally superior learning algorithms [71], comparisons with benchmark models should remain standard practice to facilitate the identification of the most effective models for train delay prediction.

In summary, the evaluation stage involves determining the evaluation aspects, level of analysis, and assessment methods. This process is flexible and subjective, often requiring multiple iterations based on the user's needs and the depth of model understanding desired. For example, if the chosen evaluation aspect is the model's accuracy, the first step might be to assess the overall accuracy at the highest level of analysis to gain a broad understanding of the model's performance. The assessment method could be to compare the model with various benchmark models to demonstrate its superiority. The next iteration might still focus on the same evaluation aspect (accuracy) and assessment method (comparison with benchmark models) but examines at different levels, such as the temporal dimension at an hourly level, to determine how the model performs throughout the day and identify any periods of low accuracy. Further iterations could focus on other timeframes (e.g., weekly or monthly) or move into a different dimension, such as spatial analysis (e.g., station-level accuracy). The evaluation iteration depends on the specific goals of the end user, balancing the need for in-depth analysis with the available resources and time constraints. For instance, some users may need detailed insights (e.g., minute-level accuracy or performance by train type), while others may be satisfied with a high-level overview.

5.3 Evaluation output

The overall model performance evaluation, which involves comparison across multiple models, enables *the identification of the best-performing model*. By considering various evaluation aspects, we can determine and *justify the strengths and limitations of the methods* used to develop the model. Furthermore, analysing models across multiple dimensions offers an in-depth understanding of their predictive capabilities in various circumstances, *answering when, where, and how certain models perform well or poorly in different times and spaces*. This comparison not only identifies the top-performing models but also highlights the critical model specifications that contribute to developing strong prediction models. Ultimately, the evaluation process enables us to draw concise conclusions and make well-informed suggestions for the continuous improvement of train delay prediction models, laying a solid foundation for building robust, real-time predictive algorithms.

6 Conclusion

The growing need for train movement forecasts within real-time ITS has driven the development of a large number of train delay prediction models. Nevertheless, the lack of clear standards for evaluating model performance hinders the deployment of these models due to uncertainties regarding their effectiveness in real-world scenarios. In this paper, we reviewed the data-driven train delay prediction literature and proposed the AP-GRIP evaluation framework. The framework aids practitioners in selecting models that best fit their operational needs by offering an in-depth understanding of their predictive capabilities, revealing model strengths and weaknesses, and recognising the trade-offs between various criteria. The framework also acts as a stepping stone for train delay prediction model development since it enables a deeper understanding of where models may need refinement and how their performance can be better balanced across diverse dimensions.

The framework outlines three key stages for evaluating train delay prediction models: defining the evaluation objective, executing the evaluation analysis, and deriving the desired evaluation outputs. Alongside predetermined objectives and desired outputs, critical characteristics such as model type and data availability significantly influence the evaluation analysis. For example, they restrict the selection of the aspects of evaluation (e.g. accuracy, precision, generalisation, robustness, interpretation, and practicality) and the level of analysis (e.g. overall, spatial, temporal, and train-specific dimensions) that can be conducted. The study provides a comprehensive discussion of current practices, limitations, and potential improvements for each evaluation aspect, advocating for a thorough evaluation that considers various levels of analysis and emphasises benchmark model comparisons. It is worth noting that the proposed framework applies to models built using methods other than data-driven approaches.

Based on our review and discussion, several key research directions are identified:

- Adopting a standardised evaluation framework like AP-GRIP for consistent comparisons across studies, which is crucial for continuous model improvement, Metrics like scale-independent prediction errors and robust precision measures should be included to ensure reliable evaluations, reducing the influence of outliers and data unit variations.
- Conducting external validation is essential for ensuring the applicability of models in real-world scenarios. Evaluating models using realistic datasets with perturbations can help bridge the gap between controlled research settings and practical railway operations.
- Evaluating model performance from the end-user perspective for meeting current operational demands. This involves penalising predictions based on end-user tolerance for errors and ensuring prediction consistency at different intervals, aligning model outputs with real-world expectations.
- Advancing post-hoc interpretability methods for train delay prediction models by exploring ad hoc techniques proven effective in other fields to improve transparency and trust in complex machine learning models.

The AP-GRIP framework is subjective in nature, where stakeholders need to make informed decisions after considering model performance across various aspects. Future research could focus on developing a more objective framework, incorporating specific ranking or prioritisation strategies that tailor model evaluation criteria to stakeholder roles and model utility. Another interesting direction for future research is to conduct a comprehensive case study, which would provide a real-world context to demonstrate the applicability of the AP-GRIP evaluation framework. Such an extension would not only validate the theoretical underpinnings of the AP-GRIP framework but also offer valuable insights into the practical utility of train delay prediction models in addressing complex challenges in train operation planning and management. Given that the metrics identified in this study primarily rely on prediction error, a potential extension for future studies could involve transforming the prediction error into a loss function associated with decision-making factors (e.g., economic justification, passenger convenience, computational complexity, etc.).

Acknowledgements
The anonymous reviewers of the European Transport Research Review are gratefully acknowledged for their eforts toward improving this paper. The authors wish to acknowledge the Innovation Programme RISE under grant agreement no. 823759.

Author contributions
Kah Yong Tiong: Conceptualization, Methodology, Formal analysis, Writing—Original Draft, Writing—Review & Editing, and Visualization. Zhenliang Ma: Conceptualization, Methodology, Validation, Writing—Original Draft, Writing—Review & Editing, and Supervision. Carl-William Palmqvist: Resources, Data Curation, Writing—Review & Editing, Supervision, Project administration, and Funding acquisition.

Funding
Open access funding provided by Lund University. This work was funded by the Swedish Transport Administration, grant numbers TRV2018/139443, TRV2020/119576, and TRV2022/32406.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that there are no conflicts of interest regarding the publication of this paper.

Received: 6 February 2024 Accepted: 29 December 2024
Published online: 10 March 2025

References

1. Spanninger, T., Trivella, A., Büchel, B., & Corman, F. (2022). A review of train delay prediction approaches. *Journal of Rail Transport Planning & Management, 22*, 100312. https://doi.org/10.1016/j.jrtpm.2022.100312
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*, 104027. https://doi.org/10.1016/j.trc.2023.104027
3. Gorman, M. F. (2009). Statistical estimation of railroad congestion delay. *Transportation Research Part E: Logistics and Transportation Review, 45*(3), 446–456. https://doi.org/10.1016/j.tre.2008.08.004
4. Li, Y., Xu, X., Li, J., & Shi, R. (2020). *A delay prediction model for high-speed railway: an extreme learning machine tuned via particle swarm optimization* 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), https://doi.org/10.1109/itsc45102.2020.9294457
5. 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
6. Li, Z., Wen, C., Hu, R., Xu, C., Huang, P., & Jiang, X. (2020). 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
7. Taleongpong, P., Hu, S., Jiang, Z., Wu, C., Popo-Ola, S., & Han, K. (2020). Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network. *Journal of Intelligent Transportation Systems, 26*(3), 311–329. https://doi.org/10.1080/15472450.2020.1858822
8. Huang, P., Spanninger, T., & Corman, F. (2022). Enhancing the understanding of train delays with delay evolution pattern discovery: A clustering and bayesian network approach. *IEEE Transactions on Intelligent Transportation Systems, 23*(9), 15367–15381. https://doi.org/10.1109/tits.2022.3140386
9. Li, J., Li, Z., Wen, C., Peng, Q., & Huang, P. (2022). Train operation conflict detection for high-speed railways: A naïve Bayes approach. *International Journal of Rail Transportation, 11*(2), 188–206. https://doi.org/10.1080/23248378.2022.2071346
10. Li, Z., Huang, P., Wen, C., Jiang, X., & Rodrigues, F. (2022). Prediction of train arrival delays considering route conflicts at multi-line stations. *Transportation Research Part C: Emerging Technologies, 138*, 103606. https://doi.org/10.1016/j.trc.2022.103606
11. Denyer, D., & Pilbeam, C. (2013). Doing a literature review in business and management *[Presentation to the British Academy of Management Doctoral Symposium]*. https://www.ifm.eng.cam.ac.uk/uploads/Research/RCDP/Resources/Working_with_literature_for_Cambridge.pdf
12. Wohlin, C. (2014). *Guidelines for snowballing in systematic literature studies and a replication in software engineering* Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. https://doi.org/10.1145/2601248.2601268
13. Loague, K., & Green, R. E. (1991). Statistical and graphical methods for evaluating solute transport models: Overview and application. *Journal of Contaminant Hydrology, 7*(1–2), 51–73. https://doi.org/10.1016/0169-7722(91)90038-3
14. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
15. Wallström, P. (2009). *Evaluation of forecasting techniques and forecast errors: with focus on intermittent demand* Luleå tekniska universitet]. Sweden. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A990519&dswid=-7452
16. Rousseeuw, P. J. (1991). Tutorial to robust statistics. *Journal of Chemometrics, 5*(1), 1–20. https://doi.org/10.1002/cem.1180050103
17. Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting, 8*(1), 69–80. https://doi.org/10.1016/0169-2070(92)90008-w
18. Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting, 9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3
19. Wu, J., Du, B., Gong, Z., Wu, Q., Shen, J., Zhou, L., & Cai, C. (2023). A GTFS data acquisition and processing framework and its application to train delay prediction. *International Journal of Transportation Science and Technology, 12*(1), 201–216. https://doi.org/10.1016/j.ijtst.2022.01.005
20. Wu, J., Wang, Y., Du, B., Wu, Q., Zhai, Y., Shen, J., Zhou, L., Cai, C., Wei, W., & Zhou, Q. (2022). the bounds of improvements toward real-time forecast of multi-scenario train delays. *IEEE Transactions on Intelligent Transportation Systems, 23*(3), 2445–2456. https://doi.org/10.1109/tits.2021.3099031
21. Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting, 32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003
22. Barbour, W., Martinez Mori, J. C., Kuppa, S., & Work, D. B. (2018). Prediction of arrival times of freight traffic on US railroads using support vector regression. *Transportation Research Part C: Emerging Technologies, 93*, 211–227. https://doi.org/10.1016/j.trc.2018.05.019
23. Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography, 28*(6), 815–829. https://doi.org/10.1111/j.2005.0906-7590.04112.x
24. Debanne, S. M. (2000). The planning of clinical studies: Bias and precision. *Gastrointestinal Endoscopy, 52*(6), 821–822. https://doi.org/10.1067/mge.2000.110757
25. Du, N., Budescu, D. V., Shelly, M. K., & Omer, T. C. (2011). The appeal of vague financial forecasts. *Organizational Behavior and Human Decision Processes, 114*(2), 179–189. https://doi.org/10.1016/j.obhdp.2010.10.005
26. Manikandan, S. (2011). Measures of dispersion. *Journal of Pharmacology and Pharmacotherapeutics*, *2*(4), 315. https://www.proquest.com/docview/898546209?pq-origsite=gscholar&fromopenview=true
27. 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*, 251–262. https://doi.org/10.1016/j.trc.2015.04.004
28. Mohammadnazar, A., Arvin, R., & Khattak, A. J. (2021). Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning. *Transportation Research Part C: Emerging Technologies, 122*, 102917. https://doi.org/10.1016/j.trc.2020.102917
29. Mou, W., Cheng, Z., & Wen, C. (2019). *Predictive model of train delays in a railway system* 8th International Conference on Railway Operations Modelling Analysis (RailNorrköping), https://ep.liu.se/ecp/069/059/ecp19069059.pdf
30. Kecman, P., & Goverde, R. M. P. (2015). Predictive modelling of running and dwell times in railway traffic. *Public Transport, 7*(3), 295–319. https://doi.org/10.1007/s12469-015-0106-7
31. Wen, C., Lessan, J., Fu, L., Huang, P., & Jiang, C. (2017, 2017/08). *Data-driven models for predicting delay recovery in high-speed rail*, in 2017 4th International Conference on Transportation Information and Safety (ICTIS), https://doi.org/10.1109/ictis.2017.8047758.
32. Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series, 1168*, 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

33. Van Calster, B., Steyerberg, E. W., Wynants, L., & van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC Medicine, 21*(1), 70–70. https://doi.org/10.1186/s12916-023-02779-w
34. Wen, C., Mou, W., Huang, P., & Li, Z. (2019). A predictive model of train delays on a railway line. *Journal of Forecasting, 39*(3), 470–488. https://doi.org/10.1002/for.2639
35. Oh, Y., Byon, Y.-J., Song, J. Y., Kwak, H.-C., & Kang, S. (2020). Dwell time estimation using real-time train operation and smart card-based passenger data: A case study in Seoul. *South Korea. Applied Sciences, 10*(2), 476. https://doi.org/10.3390/app10020476
36. IEEE. (1990). IEEE Standard Glossary of Software Engineering Terminology. https://doi.org/10.1109/ieeestd.1990.101064
37. Shahrokni, A., & Feldt, R. (2013). A systematic review of software robustness. *Information and Software Technology, 55*(1), 1–17. https://doi.org/10.1016/j.infsof.2012.06.002
38. Koçak, B., Cuocolo, R., & Dos Santos, D. P. S. (2023). Must-have qualities of clinical research on artificial intelligence and machine learning. *Balkan Medical Journal, 40*(1), 3.
39. Sharma, S., Henderson, J., & Ghosh, J. (2019). Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint*. https://doi.org/10.1145/3375627.3375812
40. Hines, A., Kendrick, P., Barri, A., Narwaria, M., & Redi, J. A. (2014). Robustness and prediction accuracy of machine learning for objective visual quality assessment. 22nd European Signal Processing Conference (EUSIPCO)
41. Fernandez, J. C., Mounier, L., & Pachon, C. (2005). A model-based approach for robustness testing. IFIP International Conference on Testing of Communicating Systems, Montreal, Canada.
42. Meng, M., Toan, T. D., Wong, Y. D., & Lam, S. H. (2022). Short-term travel-time prediction using support vector machine and nearest neighbor method. *Transportation Research Record: Journal of the Transportation Research Board, 2676*(6), 353–365. https://doi.org/10.1177/03611981221074371
43. Liu, Q., Wang, S., Li, Z., Li, L., Zhang, J., & Wen, C. (2022). Prediction of high-speed train delay propagation based on causal text information. *Railway Engineering Science, 31*(1), 89–106. https://doi.org/10.1007/s40534-022-00286-x
44. Liu, Z., Ma, Q., Tang, H., Li, J., Wang, P., & He, Q. (2022). Forecasting estimated times of arrival of US freight trains. *Transportation Planning and Technology, 45*(5), 427–448. https://doi.org/10.1080/03081060.2022.2115044
45. Chen, Z., Wang, Y., & Zhou, L. (2021). Predicting weather-induced delays of high-speed rail and aviation in China. *Transport Policy, 101*, 1–13. https://doi.org/10.1016/j.tranpol.2020.11.008
46. Bao, X., Li, Y., Li, J., Shi, R., & Ding, X. (2021). Prediction of train arrival delay using hybrid ELM-PSO approach. *Journal of Advanced Transportation, 2021*, 1–15. https://doi.org/10.1155/2021/7763126
47. Shi, R., & Xu, X. (2020, 2020/09/20). *A Train Arrival Delay Prediction model using XGBoost and Bayesian Optimization*, in 2020 IEEE 23rd international conference on intelligent transportation systems (ITSC), https://doi.org/10.1109/itsc45102.2020.9294365
48. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3064761
49. 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
50. Nabian, M. A., Alemazkoor, N., & Meidani, H. (2019). Predicting near-term train schedule performance and delay using bi-level random forests. *Transportation Research Record: Journal of the Transportation Research Board, 2673*(5), 564–573. https://doi.org/10.1177/0361198119840339
51. Lee, W.-H., Yen, L.-H., & Chou, C.-M. (2016). A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. *Transportation Research Part C: Emerging Technologies, 73*, 49–64. https://doi.org/10.1016/j.trc.2016.10.009
52. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems, 11*(3–4), 1–45. https://doi.org/10.1145/3387166
53. Rößler, D., Reisch, J., Hauck, F., & Kliewer, N. (2021). Discerning primary and secondary delays in railway networks using explainable AI. *Transportation Research Procedia, 52*, 171–178. https://doi.org/10.1016/j.trpro.2021.01.018
54. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, 2016/08/13). *"Why Should I Trust You?"* in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, https://doi.org/10.1145/2939672.2939778
55. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1), 11491. https://doi.org/10.1609/aaai.v32i1.11491
56. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv preprint*. https://doi.org/10.48550/arXiv.1805.10820
57. Plumb, G., Molitor, D., & Talwalkar, A. S. (2018). Model agnostic supervised local explanations. *Advances in neural information processing systems* https://proceedings.neurips.cc/paper_files/paper/2018/file/b495ce63ede0f4efc9eec62cb947c162-Paper.pdf
58. Gao, B., Ou, D., Dong, D., & Wu, Y. (2020). A data-driven two-stage prediction model for train primary-delay recovery time. *International Journal of Software Engineering and Knowledge Engineering, 30*(07), 921–940. https://doi.org/10.1142/s0218194020400124
59. Parbo, J., Nielsen, O. A., & Prato, C. G. (2016). Passenger perspectives in railway timetabling: A literature review. *Transport Reviews, 36*(4), 500–526. https://doi.org/10.1080/01441647.2015.1113574
60. Börjesson, M., Eliasson, J., & Franklin, J. P. (2012). Valuations of travel time variability in scheduling versus mean–variance models. *Transportation Research Part B: Methodological, 46*(7), 855–873. https://doi.org/10.1016/j.trb.2012.02.004
61. Ji, Y., Zheng, W., Dong, H., & Gao, P. (2020). *Train delays prediction based on feature selection and random forest* 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), https://doi.org/10.1109/itsc45102.2020.9294653
62. Wang, R., & Work, D. B. (2015, 2015/09). *Data driven approaches for passenger train delay estimation*, in 2015 IEEE 18th international conference on intelligent transportation systems, https://doi.org/10.1109/itsc.2015.94
63. Wu, J., Du, B., Wu, Q., Shen, J., Zhou, L., Cai, C., Zhai, Y., Wei, W., & Zhou, Q. (2021). A hybrid LSTM-CPS approach for long-term prediction of train delays in multivariate time series. *Future Transportation, 1*(3), 765–776. https://doi.org/10.3390/futuretransp1030042
64. Zhang, D., Peng, Y., Zhang, Y., Wu, D., Wang, H., & Zhang, H. (2021). 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
65. Zhang, Y. D., Liao, L., Yu, Q., Ma, W. G., & Li, K. H. (2021). Using the gradient boosting decision tree (GBDT) algorithm for a train delay prediction model considering the delay propagation feature. *Advances in Production Engineering & Management, 16*(3), 285–296. https://doi.org/10.14743/apem2021.3.400
66. Wang, Y., Wen, C., & Huang, P. (2021). Predicting the effectiveness of supplement time on delay recoveries: a support vector regression approach. *International Journal of Rail Transportation, 10*(3), 375–392. https://doi.org/10.1080/23248378.2021.1937355
67. Andersson, E., Peterson, A., & Törnquist Krasemann, J. (2011, February 16–18, 2011). *Robustness in Swedish railway traffic timetables*, in 4th international seminar on railway operations modelling and analysis, sapienza-university of rome. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A539662&dswid=-4552
68. Lulli, A., Oneto, L., Canepa, R., Petralli, S., & Anguita, D. (2018). *Large-Scale Railway Networks Train Movements: A Dynamic, Interpretable, and Robust Hybrid Data Analytics System* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), https://doi.org/10.1109/dsaa.2018.00048
69. Boateng, V., & 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

70. Nair, R., Hoang, T. L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., & Walter, T. (2019). An ensemble prediction model for train delays. *Transportation Research Part C: Emerging Technologies, 104*, 196–209. https://doi.org/10.1016/j.trc.2019.04.026
71. Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341
72. Vafaei, s., & Yaghini, M. (2024). Online Prediction of Arrival and Departure Times in Each Station for Passenger Trains Using Machine Learning Methods. In: Elsevier BV.
73. Wang, D., Guo, J., & Zhang, C. (2024). A novel hybrid deep learning model for complex systems: A case of train delay prediction. *Advances in Civil Engineering, 2024*, 1–14. https://doi.org/10.1155/2024/8163062
74. Pineda-Jaramillo, J., Bigi, F., Bosi, T., Viti, F., & D'ariano, A. (2023). Short-term arrival delay time prediction in freight rail operations using data-driven models. *IEEE Access, 11*, 46966–46978. https://doi.org/10.1109/access.2023.3275022
75. Tiong, K. Y., Ma, Z., & Palmqvist, C.-W. (2023a). Analyzing Factors Contributing to Real-time Train Arrival Delays using Seemingly Unrelated Regression Models. In: Research Square Platform LLC.
76. Gao, B., Zhang, L., Ou, D., & Dong, D. (2023). A novel deep learning model for short-term train delay prediction. *Information Sciences, 645*, 119270. https://doi.org/10.1016/j.ins.2023.119270
77. 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
78. Luo, J., Wen, C., Peng, Q., Qin, Y., & Huang, P. (2023). Forecasting the effect of traffic control strategies in railway systems: A hybrid machine learning method. *Physica A: Statistical Mechanics and its Applications, 621*, 128793. https://doi.org/10.1016/j.physa.2023.128793
79. Wiese, T. (2023). Predicting operating train delays into New York city using random forest regression and XGBoost regres-sion models. *International Journal of Engineering, Business and Management, 7*(1), 34–41. https://doi.org/10.22161/ijebm.7.1.5
80. Luo, J., Huang, P., & Peng, Q. (2022). A multi-output deep learning model based on Bayesian optimization for sequential train delays prediction. *International Journal of Rail Transportation, 11*(5), 705–731. https://doi.org/10.1080/23248378.2022.2094484
81. Tiong, K. Y., Ma, Z., & Palmqvist, C.-W. (2022b, 2022/10/08). *Real-time train arrival time prediction at multiple stations and arbitrary times*, in 2022 IEEE 25th international conference on intelligent transportation systems (ITSC), https://doi.org/10.1109/itsc55140.2022.9922299
82. Lapamonpinyo, P., Derrible, S., & Corman, F. (2022). Real-time passenger train delay prediction using machine learning: A case study with amtrak passenger train routes. *IEEE Open Journal of Intelligent Transportation Systems, 3*, 539–550. https://doi.org/10.1109/ojits.2022.3194879
83. Wang, H. (2022). A two-stage train delay prediction method based on data smoothing and multimodel fusion using asymmetry features in urban rail systems. *Wireless Communications and Mobile Computing, 2022*, 1–10. https://doi.org/10.1155/2022/5188105
84. Klumpenhouwer, W., & Shalaby, A. (2022). Using delay logs and machine learning to support passenger railway operations. *Transportation Research Record: Journal of the Transportation Research Board, 2676*(9), 134–147. https://doi.org/10.1177/03611981221085561
85. Kusonkhum, W., Srinavin, K., Leungbootnak, N., & Chaitongrat, T. (2022). Using a machine learning approach to predict the thailand underground train's passenger. *Journal of Advanced Transportation, 2022*, 1–15. https://doi.org/10.1155/2022/8789067
86. Tiong, K. Y., Ma, Z., & Palmqvist, C.-W. (2022). Prediction of real-time train arrival times along the Swedish southern mainline wit. *Transactions on The Built Environment*. https://doi.org/10.2495/cr220121
87. Laifa, H., Khcherif, R., & Ben Ghezalaa, H. H. (2021). Train delay prediction in Tunisian railway through LightGBM model. *Procedia Comput Sci, 192*, 981–990. https://doi.org/10.1016/j.procs.2021.08.101
88. Shi, R., Xu, X., Li, J., & Li, Y. (2021). Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization. *Applied Soft Computing, 109*, 107538. https://doi.org/10.1016/j.asoc.2021.107538
89. Pradhan, R., Kumar, A., Kumar, M., & Sharma, B. (2021). Simulating and analysing delay in indian railways. *IOP Conference Series: Materials Science and Engineering, 1116*(1), 012127. https://doi.org/10.1088/1757-899x/1116/1/012127
90. Huang, P., Wen, C., Fu, L., Lessan, J., Jiang, C., Peng, Q., & Xu, X. (2020). Modeling train operation as sequences: A study of delay prediction with operation and weather data. *Transportation Research Part E: Logistics and Transportation Review, 141*, 102022. https://doi.org/10.1016/j.tre.2020.102022
91. Huang, P., Wen, C., Fu, L., Peng, Q., & Li, Z. (2020). A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. *Safety Science, 122*, 104510. https://doi.org/10.1016/j.ssci.2019.104510
92. Huang, P., Wen, C., Fu, L., Peng, Q., & Tang, Y. (2020). A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems. *Information Sciences, 516*, 234–253. https://doi.org/10.1016/j.ins.2019.12.053
93. 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
94. Jiang, S., Persson, C., & Akesson, J. (2019, 2019/10). *Punctuality prediction: combined probability approach and random forest modelling with railway delay statistics in Sweden* 2019 IEEE Intelligent Transportation Systems Conference (ITSC), https://doi.org/10.1109/itsc.2019.8916892
95. Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., & Anguita, D. (2018). Train delay prediction systems: A big data analytics perspective. *Big Data Research, 11*, 54–64. https://doi.org/10.1016/j.bdr.2017.05.002
96. 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 & Management, 8*(2), 103–122. https://doi.org/10.1016/j.jrtpm.2018.02.002
97. Jiang, Z., Gu, J., Han, Y., Fan, W., & Chen, J. (2018). Modeling actual dwell time for rail transit using data analytics and support vector regression. *Journal of Transportation Engineering, Part A: Systems, 144*(11), 189. https://doi.org/10.1061/jtepbs.0000189
98. Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., & Anguita, D. (2017). Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 47*(10), 2754–2767. https://doi.org/10.1109/tsmc.2017.2693209
99. Li, D., Daamen, W., & Goverde, R. M. P. (2016). Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station. *Journal of Advanced Transportation, 50*(5), 877–896. https://doi.org/10.1002/atr.1380
100. Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., & Anguita, D. (2016). *Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data*, in 2016 IEEE international conference on data science and advanced analytics (DSAA), https://doi.org/10.1109/dsaa.2016.57
101. Pongnumkul, S., Pechprasarn, T., Kunaseth, N., & Chaipah, K. (2014, 2014/05). *Improving arrival time prediction of Thailand's passenger trains using historical travel times* 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), https://doi.org/10.1109/jcsse.2014.6841886
102. Yaghini, M., Khoshraftar, M. M., & Seyedabadi, M. (2012). Railway passenger train delay prediction via neural network model. *Journal of Advanced Transportation, 47*(3), 355–368. https://doi.org/10.1002/atr.193
103. Peters, J., Emig, B., Jung, M., & Schmidt, S. (2005). *Prediction of Delays in Public Transportation using Neural Networks* in international conference on computational intelligence for modelling, control and automation and international conference on intelligent agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), https://doi.org/10.1109/cimca.2005.1631451
104. Barbour, W., Samal, C., Kuppa, S., Dubey, A., & Work, D. B. (2018). *On the Data-Driven Prediction of Arrival Times for Freight Trains on U.S. Railroads* 2018 21st International Conference on Intelligent Transportation Systems (ITSC), https://doi.org/10.1109/itsc.2018.8569406
105. Luo, J., Peng, Q., Wen, C., Wen, W., & Huang, P. (2022). Data-driven decision support for rail traffic control: A predictive approach. *Expert*

Systems with Applications, 207, 118050. https://doi.org/10.1016/j.eswa.2022.118050

106. Washington, S. K., M. G. Mannering, F. Anastasopoulos, P. (2020). *Statistical and econometric methods for transportation data analysis*. CRC press.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.