Reliable Train Delay Forecasting with Conformal Prediction

Xu Feng(✉), Khuong An Nguyen, and Zhiyuan Luo

Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, UK
Xu.Feng@rhul.ac.uk

Abstract. Train delays cause significant economic and operational impacts. Accurate prediction of these delays is therefore essential for improving railway service reliability and supporting effective decision-making. Thus, this chapter introduces a novel approach combining tree-based machine learning (ML) models with Conformal Prediction (CP) to estimate train delays by predicting train travel times between consecutive stations. Using real-world data from over 12.8 million production service records collected over a period of 3 years and 8 months, the proposed approach achieved a prediction accuracy of 20 s, 90% of the time. Furthermore, CP delivers statistically valid prediction intervals with an average width of 19.3 s at a 90% confidence level, successfully meeting the target coverage as demonstrated by empirical results.

Keywords: Train delay prediction · Timetabling optimisation · Conformal Prediction

1 Introduction

The railway system is a key component of modern transportation infrastructure that plays a pivotal role in connecting people and goods over long distances [1,44,58,59]. One of the most critical aspects of the railway operations is punctuality, as it represents the reliability and integrity of the entire system [48,58–60,68]. Punctuality, in the context of railway systems, refers to the ability of trains to operate according to the timetable with minimal deviation [36]. However, high variability in real-world railway operations and unexpected disruptions such as weather conditions, train malfunctions and infrastructure issues like signal failures often result in train delays [75]. Train delays may result in economic losses, passenger inconvenience and unsatisfactory, and a cascade of congestion across the entire railway network [71].

Given the far-reaching impacts of train delays, predicting and mitigating them is of paramount importance. Over the past decades, there has been a concerted effort to employ various Machine Learning (ML) models and methods to train delay predictions. In [57], XGBoost combined with Bayesian Optimisation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
K. An Nguyen and Z. Luo (Eds.): Alexander Gammerman Festschrift, LNCS 16290, pp. 387–410, 2026.
https://doi.org/10.1007/978-3-032-15120-9_17

were utilised to predict train arrival delays based on spatio-temporal, operational, and infrastructure data. A long short-term memory (LSTM) model was proposed in [69], leveraging train operational features to predict train arrival delays. In [61], the researchers used artificial neural networks, random forest regression, gradient boosting regression with features from planned and actual train operation data to predict arrival and departure times at each main station.

However, train journeys do not happen in isolation. Thus, modelling the train delay prediction problem requires capturing the sequential relationships between stations along the route, which complicates dataset construction and model training. In addition, current train delay prediction models lack the ability to provide meaningful confidence measures for their predictions, making it difficult for railway operators to assess the reliability of the forecasts and potentially leading to suboptimal operational decisions.

To this end, this chapter proposes a novel train delay prediction model that leverages Machine Learning (ML) methods and Conformal Prediction (CP) to estimate train travel times between stations, thereby generating reliable delay forecasts at the following stations. In this way, the proposed approach preserves the dependencies between train arrival and departure events, as well as the sequential relationship between stations, while ensuring efficient dataset construction and model training. Furthermore, conformal prediction provides prediction intervals with user-specified confidence level for each delay estimate, enabling uncertainty quantification and more transparent and trustworthy decision-making. The contributions of this study are summarised as follows:

- We propose a novel machine learning-based approach for train delay prediction that leverages operational data to estimate the travel time between consecutive stations. This model preserves the dependencies between train events and the sequential relationships between stations, while also ensuring efficient dataset construction and model training.
- We leverage conformal prediction to generate rigorous, statistically valid prediction intervals for each train delay forecast, thereby enhancing the confidence, reliability, transparency, and trustworthiness of the decision-making process in railway systems.
- We validate the performance of the proposed machine learning model using real-world, large-scale train operational data, comprising 12,840,590 train service records collected over a period of 3 years and 8 months. Furthermore, we compare the performance of the most widely used machine learning models in the literature, offering an in-depth analysis of their reliability for train delay prediction.

The remainder of this paper is structured as follows: Sect. 2 overviews the most popular train delay prediction approaches. Section 3 provides a detailed description of the problem formulation and the system architecture. Section 4 offers in-depth introduction to the machine learning model used for train delay prediction and the conformal prediction framework. The empirical experiments and performances analysis are presented in Sect. 5. Finally, Sect. 6 concludes our work and outlines future work.

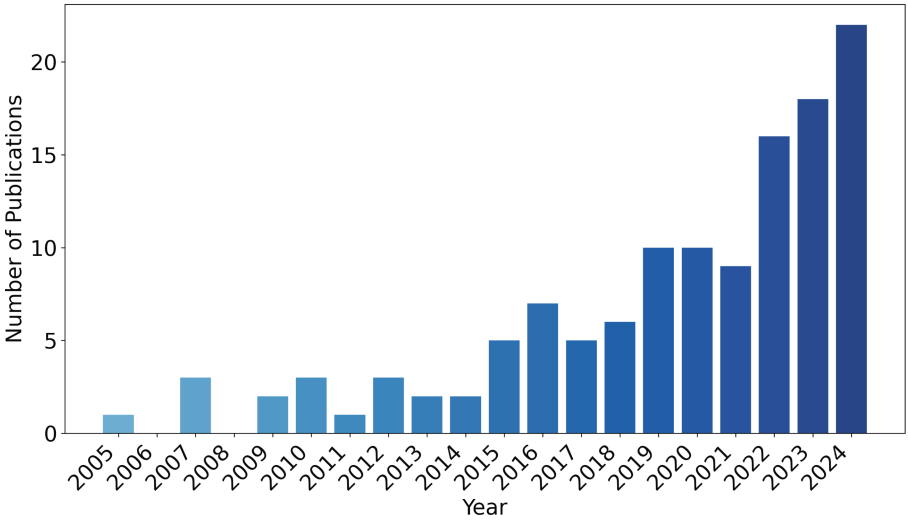

Fig. 1. Number of published papers on train delay prediction by year.

2 Literature Review

Over the past two decades, there has been a continuously growing trend in the number of train delay prediction contributions in the academic world, as discussed in Sect. 1 and illustrated in Fig. 1. The most popular train delay prediction approaches in the literature can be categorised into two groups, namely the event-driven models and data-driven models [58]. Event-driven approaches explicitly model the dependencies between the different train events (e.g., arrivals, departures, dwells, passing-by, etc.) and create a chain or network of events, capturing how delays propagate through the railway system. In contrast, data-driven approaches use historical data and machine learning or statistical methods to directly predict the delay at a given future point or station, based on available features (e.g., current delay, infrastructure status, weather information, etc.) in a single step. Event-driven approaches are generally more interpretable, as the operator can trace how delays propagate through the railway system, while data-driven methods typically rely more on large datasets and can capture complex hidden relationships in the data.

Consequently, typical event-driven train delay prediction approaches employ models such as Graph Models, Markov Chain (MC) models, Bayesian Networks (BN), and Equation Systems (EQS), among others, to highlight the dependencies between various train events during prediction. To explore the underlying patterns in the large-scale datasets, machine learning models like Neural Networks (NN), Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), Linear Regression (LNR) were employed for data-driven methods. The proportions of different prediction models among the included references in the literature are shown in Fig. 2. To provide a comprehensive comparison of the

approaches in the literature, we categorise and analyse them using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as shown in Table 1.

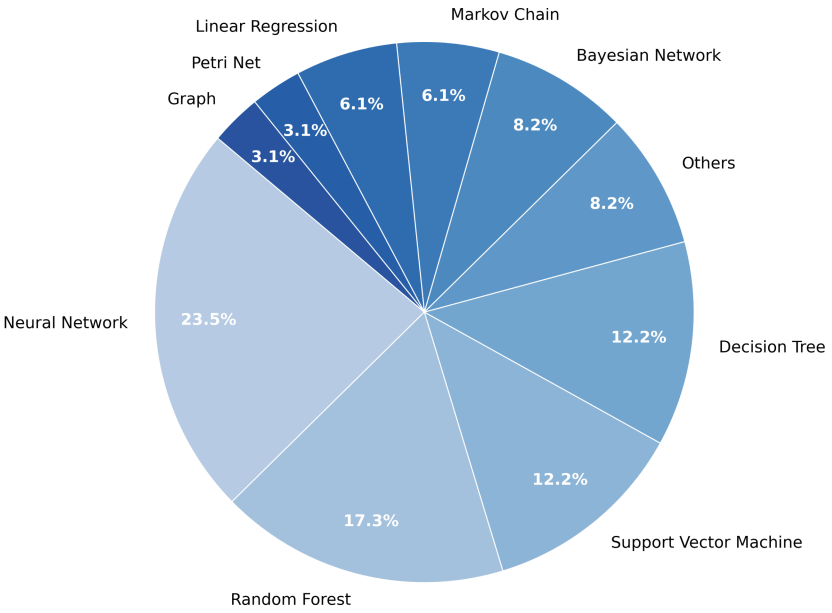

Fig. 2. Proportions of different prediction models in the literature.

Event-driven approaches heavily rely on the logic chains of the train events along the route, thus building and calibrating the dependency structure is time-consuming and challenging [13,42,58]. Upon using the highly specified data structure, event-driven methods may not fully utilise the richness of historical data beyond parameter fitting and may produce a less flexible prediction model that struggles with capturing complex, hidden relationships in the datasets. Data-driven methods can generate the train delay prediction directly with promising accuracies. However, the black-box training makes the final estimation less interpretable for real-world operators.

To address this issue, this chapter proposes a novel machine learning-based approach incorporating the most widely used tree-based models and conformal prediction [38,40,41]. The proposed approach leverages historical train operational data to predict the travel time between the current station and the next, rather than focusing solely on train delays. By estimating the train travel time between consecutive stations, the approach captures the dependencies between train departure and arrival events, which can then be used to more accurately estimate train delays at following stations along the route. At the same time, conformal prediction provides prediction intervals for estimated train travel time that helps train operators take proactive measures to stay on schedule with confidence. Leveraging the underlying patterns in large-scale, real-world datasets

Table 1. Comparisons of different train delay prediction approaches in the literature.

Paper	Model	Country	Data size	Metrics	Prediction Error	Prediction horizon
[17]	BN	CHN	98,982	MAE, RMSE	MAE 3.79 min, RMSE 9.03 min	Multiple
[28]	BN	CHN	378,510	MAE, MSE, RMSE	MAE 30 sec	Multiple
[9]	BN	SWE	486,000	MAE	1.40 min	Multiple
[33]	BN	N/A	N/A	N/A	N/A	One
[75]	BN	NLD	1,307	RMSE	100.15 min	One
[72]	EQS	N/A	N/A	MAE	33.16 sec	One
[15]	MC	IND	194,400	RMSE	11.75 min	One
[52]	MC	TUR	421	N/A	N/A	Multiple
[23]	MC	NLD	6,803	MAE	2.40 min	Multiple
[5]	MC	CHE	17,000	N/A	N/A	Multiple
[74]	PN	CHN	N/A	MAE	2 min	Multiple
[27]	Graph	DEU	2M	Acc	98.90%	One
[25]	Graph	DEU	223,873	N/A	N/A	One
[32]	Max-plus	CHN	N/A	N/A	N/A	Multiple
[24]	TEG	NLD	49,200	MAE	40 sec	Multiple
[3]	NN	CHN	400,000	MAE, RMSE	MAE 0.35 min, RMSE 1.04 min	One
[18]	NN	CHN	975,592	MAE	3.24 min	One
[73]	NN	CHN	2M	RMSE	0.50 min	One
[35]	NN	NLD	66 days	MAE, RMSE	30 sec, 87%	One
[69]	NN	NLD	66,178	MAE, RMSE	30 sec, 86%	One
[46]	NN	ITA	6 months	Avg Acc	1.70 min	Multiple
[31]	NN	CHN	1 day	RMSE	4.91 sec	One
[70]	NN	IRN	179,982	Acc	93.34%	One
[29]	NN	CHN	400,000	MAE, RMSE	MAE 0.89 min, RMSE 2.13 min	One
[20]	NN	CHN	229,320	MAE, RMSE	<0.6 min	One
[57]	DT	CHN	183,207	MAE, RMSE	MAE 0.85 min, RMSE 2.29 min	One

(*continued*)

Table 1. (*continued*)

Paper	Model	Country	Data size	Metrics	Prediction Error	Prediction horizon
[56]	DT	CHN	330,787	MAE, MSE, RMSE	MAE 0.46 min, RMSE 1.37 min	One
[65]	DT	CHN	2.7M	MAE	25 min	Multiple
[67]	DT	JPN	N/A	N/A	N/A	Multiple
[26]	DT	TUN	12,350	MAE, RMSE	MAE 9.88 min, RMSE 18.29 min	One
[8]	SVM	CHN	2,025	MAE, RMSE	MAE 1.68 min, RMSE 2.24 min	One
[66]	SVM	CHN	17,371	MAE, MSE	MAE 0.32 min, MSE 0.44 min	One
[19]	SVM	CHN	57,796	MAE	0.55 min	One
[53]	SVM	FRA	5 years	Acc	85.38%	One
[4]	SVM	USA	150,000	MAE	N/A	One
[2]	RF	IND	1,170	MAE, RMSE	MAE 49.28 min, RMSE 80.07 min	One
[14]	RF	CHN	1 year	MAE	<3 min	One
[30]	RF	NLD	5.6M	MAE, RMSE	MAE 0.69 min, RMSE 1.68 min	One
[37]	RF	DEU	3.3 years	RMSE	369.50 sec	One
[36]	RF	NLD	10M	Acc	93%	One
[21]	RF	CHN	29,662	MAE	1 min, 80.4%	One
[22]	RF	SWE	2.2M	RMSE	3.33 min	One
[50]	LNR	IND	1 year	N/A	N/A	One
[16]	LNR	DEU	N/A	MSE	<75 min	Multiple
[12]	Clustering	NLD	525,600	MSE	90 min	Multiple
[49]	TSA	THA	182 days	MAE	5.86 min	One, multiple
[11]	Simulation	RUS	N/A	Mean, STD	N/A	Multiple

Model abbreviations: BN=Bayesian Network; DT=Decision Tree; EQS=Equation System; LNR=Linear Regression; MC=Markov Chain; NN=Neural Network; PN=Petri Network; RF=Random Forest; SVM=Support Vector Machine; TEG=Timed Event Graph; TSA=Time Series Analysis.
Country codes: Standard ISO ALPHA-3 country codes (e.g. CHN=China, DEU=Germany).
Metrics: MAE=Mean Absolute Error; RMSE=Root Mean Square Error; MSE=Mean Square Error; Acc=Accuracy; Avg Acc=Average Accuracy; STD=Standard Deviation.
Units: min=minutes; sec=seconds; M=millions.
Prediction Horizon: One=one station ahead; Multiple=multiple stations ahead.

while providing well-calibrated confidence measures for each prediction, the proposed machine learning-based approach delivers accurate and reliable train delay predictions.

3 Problem Statement and System Architecture

This section begins by presenting the formulation of the train delay prediction problem that the proposed method aims to solve. Subsequently, a general overview of the proposed method's system architecture is provided.

3.1 Problem Formulation

To understand the train delay prediction problem in a systematic way, assume a targeted train service i is scheduled to arrive at station j at time t_s. Then at a current time $t \leq t_s$, the predicted train delay $\hat{\Phi}$ is defined as:

$$\hat{\Phi}_{i,j,t,t_s} = f(\Omega_{i,j}) \tag{1}$$

where $\Omega_{i,j}$ is historical train operational data of train service i at station j, namely the input to the model, and f maps the input information $\Omega_{i,j}$ to the predicted train delay $\hat{\Phi}_{i,j,t,t_s}$. This formulation naturally lends itself to a regression task.

In the proposed method, the predicted train delay $\hat{\Phi}$ is calculated from the scheduled train arrival time t_s and the difference in the train travel time $\Delta T_{i,j-1,j}$ from the previous station $j-1$ to the current station j. The estimated train travel time difference $\Delta\hat{T}_{i,j-1,j}$ is generated by the machine learning model M based on train operational data $\Omega_{i,j}$, defined as

$$\Delta\hat{T}_{i,j-1,j} = M(\Omega_{i,j}) - T_{j-1,j} \tag{2}$$

where $T_{i,j-1,j}$ is the scheduled train travel time between these consecutive stations $j-1$ and j of the train service i.

Therefore, the predicted train delay $\hat{\Phi}$ in the proposed method is defined as:

$$\hat{\Phi}_{i,j,t,t_s} = M(\Omega_{i,j}) - T_{i,j-1,j}. \tag{3}$$

3.2 System Architecture

To deliver accurate train delay estimates along with prediction intervals that provide meaningful confidence measures of each prediction $\hat{\Phi}_{i,j,t,t_s}$, the proposed system is structured into three main steps: data preparation, model training, and test sample prediction, as shown in Fig. 3. The brief description of each step is as follows:

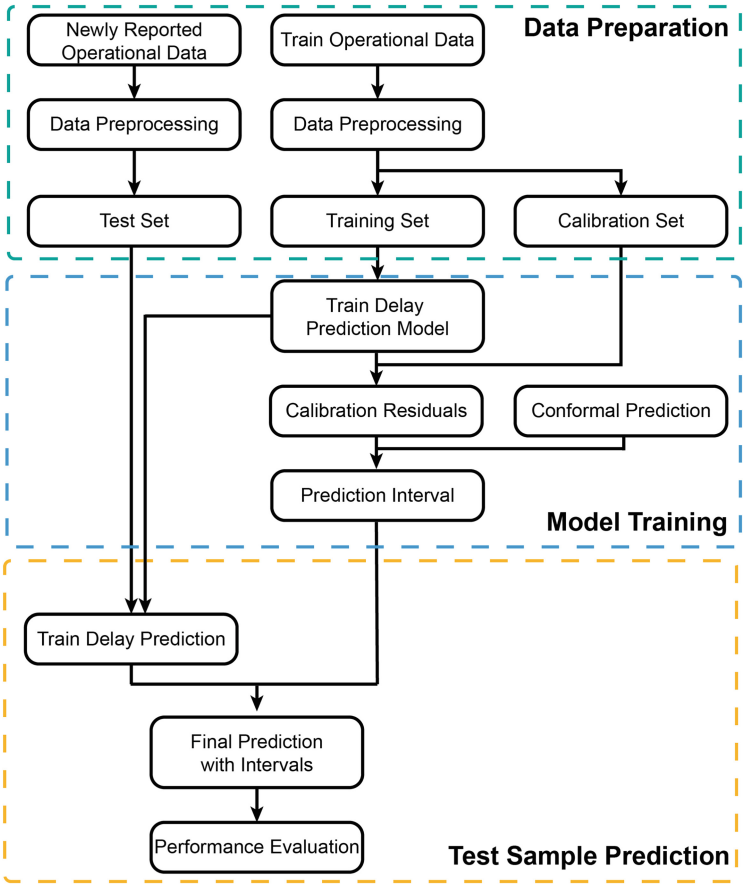

Fig. 3. The overview of the proposed train delay prediction approach.

Data Preparation: We begin by preprocessing the raw historical train operational records. Specifically, we first identify and remove any complete train services that contain missing values or outliers in their arrival time, departure time, dwell time, or train travel time at any station.

To facilitate numerical processing, ISO timestamps in the aforementioned operational data are converted to Unix timestamps. Next, categorical attributes, those identifying train services, engine models, and specific stations, are label-encoded to transform them into numeric form.

Finally, the dataset is divided into training, calibration, and test sets for model training, prediction interval generation, and final evaluation, respectively.

Model Training: In the model training step, a machine learning-based model for train delay prediction is developed, and conformal prediction is employed to generate statistically valid prediction intervals. In the proposed approach, the

output is the estimated travel time to the next station, which is subsequently used to generate the final train delay predictions (see Eq. 3).

First, tree-based models, identified in Sect. 2 as the most widely adopted algorithms in the literature, are trained on the training dataset. Next, the trained model is used to predict outcomes for the calibration set. The residuals from these predictions, calculated as the differences between the predicted and actual values in the calibration set, are then used within the conformal prediction framework to construct prediction intervals with user specified confidence level. Finally, the trained machine learning model is applied to the test set to generate train delay estimations.

Test Sample Prediction: In the final step, the accuracy of the machine learning train delay prediction model and the quality of the prediction intervals are evaluated.

Performance metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used to assess the predictive accuracy of the machine learning model. Subsequently, the model's performance is compared with other popular machine learning approaches for train delay prediction. Finally the quality of the prediction intervals is analysed.

4 Machine Learning Models and Conformal Prediction

This section offers a comprehensive description to the most widely used machine learning models in the literature for train delay prediction tasks and an in-depth introduction to conformal prediction that produces statistical valid prediction intervals with a predefined confidence level.

4.1 Machine Learning Models in Train Delay Prediction

To provide accurate and reliable train delay predictions, a large-scale real-world train operational dataset spanning 3 years and 8 months (see Sect. 5.1 for details) is leveraged in this study. To effectively uncover underlying patterns in this dataset, the data-driven approach is adopted in the proposed system.

As discussed in Sect. 2 and illustrated in Fig. 2, the most commonly used machine learning algorithms for data-driven approaches in the literature are Neural Networks (NN), Random Forests (RF), and Support Vector Regression (the regression extension of SVM). Given the dataset's size of 12,840,590 records, training NNs is extremely time-consuming and computationally expensive. Therefore, the more efficient RF model is selected as the primary machine learning algorithm in the proposed approach.

Random Forest (RF). Random Forest is a robust ensemble learning method that consists of multiple decision trees and combines their outputs to improve

prediction accuracy and control overfitting in train delay prediction [6]. A standard RF constructs numerous decision trees during training, each trained on a random subset of the data and a random subset of features at each split, and outputs the average of individual tree predictions.

A standard decision tree works by repeatedly splitting the feature space into smaller regions. Each internal node applies a rule based on one of the input features, and each leaf node gives a final prediction [7,34,45,51]. Thus, the ensemble of trees is defined as:

$$\{Tree_d(\mathbf{\Omega}_{i,j})\}_{d=1}^{D} \tag{4}$$

where $Tree_d$ denotes the individual decision tree, D is the total number of trees in the ensemble, and $\mathbf{\Omega}_{i,j}$ is a bootstrap sample of the input train operational data $\Omega_{i,j}$. A bootstrap sample is a sub-dataset created by randomly sampling with replacement from the original data, helping the model mitigate overfitting during training. For a tree node n_{tree} with $N_{n_{tree}}$ samples, the training objective is to minimise the MSE, defined as:

$$\text{MSE}(n_{tree}) = \frac{1}{N_{n_{tree}}} \sum_{k \in n_{tree}} \left(T_{i,j-1,j} - \bar{T}_n\right)^2 \tag{5}$$

where $T_{i,j-1,j}$ is the ground truth train travel time of the i-th train service at station j in the node, and $\bar{T}_n$ is the average of target train travel time values in the node. Subsequently, the final prediction $\hat{T}_{i,j-1,j}$ of the train travel time from the previous station $j-1$ to the current station j is calculated as:

$$\hat{T}_{i,j-1,j} = \frac{1}{D} \sum_{d=1}^{D} Tree_d(\mathbf{\Omega}_{i,j}). \tag{6}$$

In the proposed train delay prediction approach, the input train operational data $\Omega_{i,j}$ to the RF model contains service identifiers that uniquely distinguish the train service, as well as operational details that describe the train's arrival and departure at each station stop. After the estimation of the train travel time, the final train delay prediction $\hat{\Phi}_{i,j}$ at station j is derived.

To provide a comprehensive comparison of the performance of popular machine learning models for train delay prediction, we also implement distinguishing machine learning models such as Support Vector Regression (SVR) and Linear Regression (LNR).

Support Vector Regression (SVR). SVR, a regression extension of SVM, is a robust supervised learning models widely used in the literature to enhance the accuracy of train delay prediction [10,39,58,62]. A standard SVR aims to find an optimal hyperplane that best fits the data within a predefined margin. The hyperplane is determined by support vectors, which are the training samples nearest to the decision boundary and play a crucial role in its definition.

For the regression model is defined by the function:

$$f(\Omega_{i,j}) = \langle \mathbf{w}, \Omega_{i,j} \rangle + b. \tag{7}$$

The primal optimisation objective for SVR model for train delay prediction is given by:

$$\min_{\mathbf{w},b,\xi_k,\xi_k^*} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{N_{train}} \sum_{j=1}^{N_{station}} (\xi_{i,j} + \xi_{i,j}^*) \tag{8}$$

subject to:

$$\hat{T}_{i,j-1,j} - \langle \mathbf{w}, \Omega_{i,j} \rangle - b \leq \epsilon + \xi_{i,j}, \tag{9}$$

$$\langle \mathbf{w}, \Omega_{i,j} \rangle + b - \hat{T}_{i,j-1,j} \leq \epsilon + \xi_{i,j}^*, \tag{10}$$

$$\xi_{i,j} \geq 0, \quad \xi_{i,j}^* \geq 0 \tag{11}$$

where $\mathbf{w}$ is the weight vector defining the orientation of the regression hyperplane, b is the bias term shifting the hyperplane from the origin, $\xi_{i,j}$ and $\xi_{i,j}^*$ are slack variables measuring prediction errors above and below the margin tolerance ϵ respectively, C is the regularisation parameter controlling the trade-off between model simplicity and error tolerance, $\|\mathbf{w}\|^2$ is the squared L2-norm regularisation term promoting model flatness, N_{train} is the total number of train services, and $N_{station}$ is the total number train stations in the railway network.

Linear Regression (LNR). Linear Regression is a popular supervised learning algorithm for modelling the relationship between a the historical train operational data and the train delays [43,54,59]. The goal is to find a linear function that best predicts train delay times based on input features.

In the proposed approach, to predict the train travel time to the next station, the model assumes:

$$T_{i,j-1,j} = \mathbf{w}^T \Omega_{i,j} + b + \epsilon \tag{12}$$

where $\mathbf{w}$ is the weight vector, b is the bias term, and ϵ is the residual error. The LNR model is optimised by minimising the loss, defined as:

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b) = \min_{\mathbf{w},b} \frac{1}{N_{train} \times N_{station}} \sum_{i=1}^{N_{train}} \sum_{j=1}^{N_{station}} \left(T_{i,j-1,j} - (\mathbf{w}^T \Omega_{i,j} + b)\right)^2 \tag{13}$$

where N_{train} is the total number of train services, and $N_{station}$ is the total number train stations in the railway network.

Additionally, a range of tree-based models, including Decision Trees (DT), Gradient Boosting (GB), Histogram-based Gradient Boosting (histGB), and eXtreme Gradient Boosting (XGB), as well as regularised LNR-based models such as Ridge Regression (Ridge) and Lasso Regression (Lasso), are included in the empirical experiments.

4.2 Conformal Prediction

While machine learning models provide promising predictions of train delays, they do not guarantee the uncertainty associated with individual predictions. In high-stakes systems like railway networks, the ability to provide not only accurate but also interpretable and uncertainty-guaranteed predictions is crucial. Thus, conformal prediction is leveraged to quantify prediction confidence, allowing train operators to assess how often true train delays fall within a specified bounds, thus enabling more informed and reliable decision-making.

Conformal Prediction (CP) is a statistically rigorous framework for uncertainty quantification that produces prediction intervals with guaranteed coverage, thereby improving the reliability of machine learning model predictions [47,55,63,64].

Recall that a machine-learning model M is trained on train operational data $\Omega_{i,j}$ for train travel time prediction of train service i at station j, defined as:

$$\hat{T}_{i,j-1,j} = M(\Omega_{i,j}). \tag{14}$$

Given a user-specified confidence level α, CP constructs prediction intervals that are guaranteed to contain the true target value with probability at least $1-\alpha$, under the assumption of exchangeable data. In our experiments, we use a computationally efficient variant of CP, namely Inductive Conformal Prediction (ICP). To quantify uncertainty non-parametrically, ICP first uses a calibration dataset to compute nonconformity scores, which measures of how unusual or nonconforming a new observation is relative to the known data distribution. The nonconformity score is calculated by the calibration residuals defined as:

$$R_k^{\text{cal}} = |T_k - M(\Omega_k)|, \quad k = 1, \ldots, m \tag{15}$$

where R_k^{cal} is the residuals of the calibration set, m represents the number of samples in the calibration dataset, T_k and Ω_k is the ground truth label (i.e., train travel time) and input features of the k-th sample in the calibration set, respectively, M is the applied machine learning model. Subsequently, the critical quantile $\hat{q}$ which determines how wide the prediction interval should be to meet the specified confidence level α is calculated as:

$$\hat{q} = \text{quantile}\left(\left\{R_k^{\text{cal}}\right\}_{k=1}^{m} \cup \{\infty\}, \frac{\lceil (m+1)(1-\alpha) \rceil}{m}\right). \tag{16}$$

Once the critical quantile $\hat{q}$ is computed, a prediction interval for a new train operational data input Ω_{test} is defined as:

$$[M(\Omega_{test}) - \hat{q}, M(\Omega_{test}) + \hat{q}]. \tag{17}$$

Under the assumption that the training and test data are exchangeable, this prediction interval comes with a rigorous coverage guarantee [47,55]. Specifically, the interval will contain the true outcome with probability at least $1-\alpha$, formalised as:

$$\mathbb{P}\left(T_{\text{test}} \in [M(\Omega_{test}) - \hat{q}, M(\Omega_{test}) + \hat{q}]\right) \geq 1 - \alpha. \tag{18}$$

This implies that, across many predictions, the fraction of the prediction intervals that capture the true train travel time values will be at least $1 - \alpha$, therefore providing a statistically sound and transparent measure of prediction uncertainty.

5 Empirical Experiments

This section begins with a detailed introduction to the train operational dataset used in this study, followed by a brief overview of the evaluation metrics employed to assess the performance of the proposed train delay prediction approach. Finally, it presents the empirical results and a performance comparison with other popular machine learning models in the literature.

5.1 Dataset Description

To evaluate the accuracy of the proposed train delay prediction method and the quality of the prediction intervals, a large-scale real-world train operational dataset is utilised. This dataset encompasses approximately 3 years and 8 months of data, spanning from February 12, 2020, to October 8, 2023. It includes records for 12,621 unique train services operating between specific stations along a designated railway line in the southeast of the United Kingdom. Each of these services may have run multiple times during the recorded period. An example of a raw train service record is shown in Table 2. A brief description of the train operational record attributes is as follows:

`Headcode` Used nationally to determine a train service between specific stations and on a prespecified line.

`UnitNumber` Identifies which engine is working on this train service; this is the engine that "drives" the train.

`TrainModel` Represents the service family inside the railway system of the operator company.

`Stops` Consists of the train operational records at each station where it stops, including:

- `Name` The name of the station.
- `CRS` The Computer Reservation System code, used to identify railway stations for ticketing and passenger information systems.
- `Tiploc` The Timing Point Location code, used to identify specific timing points on the railway network, including stations, junctions, sidings, and other timing points (e.g., for arrival, departure, dwell).
- `BookedDeparture`, `ActualDeparture`, `DepartureDiff` Information about departure at this station.
- `BookedArrival`, `ActualArrival`, `ArrivalDiff` Information about train arrival at this station.

`DwellBooked`, `DwellActual`, `DwellDiff` Information about train dwell time at this station.
`UntilNextLocationBookedTime` Booked train travel time until the next stop.
`UntilNextLocationActualTime` Actual train travel time until the next stop.
`UntilNextLocationTimeDiff` Difference in train travel time until the next stop.
`Delayed` The delayed time identified for this stop.

Note that the time '0001-01-01T00:00:00.000Z' in the booked and actual arrival times indicates that this is the first station/stop in this train service, while the same time appears in the booked and actual arrival times represents the very last stop in the route.

To construct a training dataset suitable for machine learning models, each train service record is segmented into multiple data samples. Each sample includes key identifiers (e.g., 'Headcode', 'UnitNumber') to uniquely distinguish the train service, as well as operational details (e.g., 'Tiploc', 'BookedDeparture', 'ActualDeparture', 'BookedArrival', 'ActualArrival', 'DepartureDiff', 'ArrivalDiff', 'DwellBooked', 'DwellActual', 'DwellDiff', 'UntilNextLocationBookedTime', 'UntilNextLocationActualTime', 'UntilNextLocationTimeDiff') that describe the train's movements at each station stop. After preprocessing, the final dataset consists of 12,840,590 individual train service records.

5.2 Evaluation Metrics

To assess the prediction performance of the train delay prediction approach, popular evaluation metrics including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are utilised. To assess the quality of prediction intervals, Residual Coverage Score (RCS) is leveraged.

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) is a straightforward metric for evaluating the accuracy of a train delay prediction model. It calculates the mean of the absolute differences between predicted and actual train delays, as shown by:

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k|. \tag{19}$$

Here, n is the number of observations, y_k represents the actual delay, and $\hat{y}_k$ denotes the predicted delay for each observation. A lower MAE suggests better model accuracy, offering a direct measure of the average prediction error.

Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a widely used metric for evaluating the accuracy of train delay predictions. It is derived from the Mean Squared Error (MSE), which measures the average of the squared differences between the actual and predicted values:

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2. \tag{20}$$

Table 2. A snapshot of the raw train operational records.

Attribute	Value
Headcode	1V84BA
UnitNumber	376016
TrainModel	465
Stops	
Name	Charing Cross
Crs	CHX
Tiploc	CHRX
BookedDeparture	2020-03-17T23:37:00.000Z
BookedArrival	0001-01-01T00:00:00.000Z
ActualDeparture	2020-03-17T23:37:04.000Z
ActualArrival	0001-01-01T00:00:00.000Z
DwellBooked	0
DwellActual	0.0
DepartureDiff	4.0 s
ArrivalDiff	0.0 s
DwellDiff	0.0 s
UntilNextLocationActualTime	143.0 s
UntilNextLocationBookedTime	120.0 s
UntilNextLocationTimeDiff	23.0 s
Delayed	0
Stops	
Name	London Bridge
...	...

By taking the square root of MSE, RMSE expresses this error in the same units as the original data:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}. \tag{21}$$

RMSE penalises larger errors more heavily due to the squaring step, making it particularly sensitive to significant deviations between predicted and actual delays. Its interpretability—being in the same scale as the target variable—makes it especially valuable for assessing the magnitude of prediction errors in a practical and intuitive manner. A lower RMSE indicates better predictive performance and is therefore a central metric in evaluating the effectiveness of delay prediction models.

Residual Coverage Score (RCS)
To quantify how frequently the true train delay values fall within the predicted intervals, Residual Coverage Score (RCS) is introduced, defined as:

$$\mathrm{RCS} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\left(\hat{y}_k^{\mathrm{low}} \leq y_k \leq \hat{y}_k^{\mathrm{up}}\right) \tag{22}$$

where y_k is the true value for the k-th sample, $\hat{y}_k^{\mathrm{low}}$ and $\hat{y}_k^{\mathrm{up}}$ are the lower and upper bounds of the prediction interval, respectively. For each of the test samples, the indicator function $\mathbb{I}$ returns 1 if the interval contains the true value and 0 otherwise. The RCS is then the average of these outcomes across the entire dataset. In repeated sampling, the proportion of times that the interval contains the true train delay values should match the predefined confidence level $1 - \alpha$. Thus, RCS assesses whether the prediction intervals are statistically valid.

5.3 Empirical Results

To evaluate the performance of the proposed train delay prediction approach, Random Forest (RF) and several widely used machine learning models are implemented on the aforementioned train operational dataset. These include Support Vector Regression (SVR), Linear Regression (LNR), Decision Trees (DT), Gradient Boosting (GB), Histogram-based Gradient Boosting (HistGB), and eXtreme Gradient Boosting (XGB), along with regularised linear models such as Ridge Regression (Ridge) and Lasso Regression (Lasso).

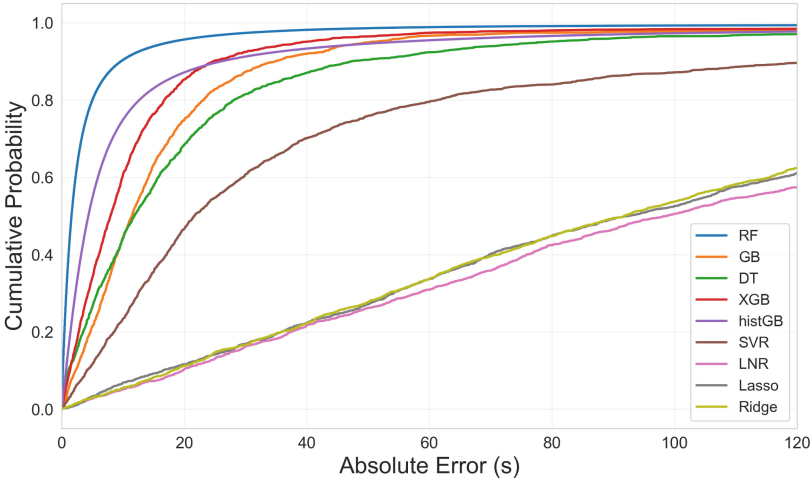

Fig. 4. The CDF curves of the absolute errors generated by each machine learning model. The RF model produces the most accurate and robust prediction, achieving an accuracy of 20 s, 95% of the time. However, SVR and LNR-based models fail to effectively map the correlation between the train operational data to the train travel time between stations.

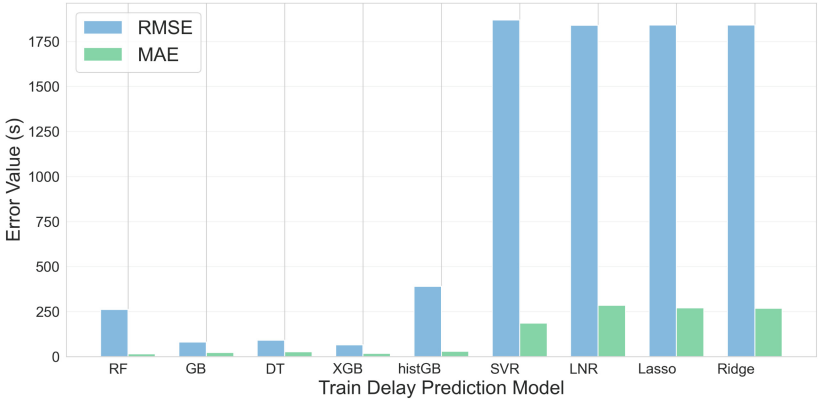

Fig. 5. The performance comparison of the popular machine learning models. The results show that tree-based models outperform SVR and LNR models in predicting train delays on large-scale operational data.

In order to predict the train travel time to the next station (i.e., 'UntilNextLocationActualTime'), 'Headcode', 'UnitNumber', 'Tiploc', 'BookedDeparture', 'ActualDeparture', 'BookedArrival', 'ActualArrival', 'DepartureDiff', 'ArrivalDiff', 'DwellBooked', 'DwellActual', 'DwellDiff', and 'UntilNextLocationBookedTime' are leveraged as the input features to the machine learning model. A desktop PC equipped with an Intel i9-12900k @ 4.90 GHz CPU and 32GB DDR4 4000MHz memory was used to analyse the results.

The cumulative distribution function (CDF) curves, along with the RMSE and MAE of the employed models, are presented in Table 3 and Figs. 4 and 5. It is observed that the RF model produces the most accurate and robust predictions, achieving an accuracy of 20 s, 95% of the time. SVR, another popular machine learning model in train delay prediction, however, struggles at a mean error of 185.83 s. Given the extremely time-consuming training process on real-world train operational data, SVR models are not considered as a suitable method for train delay prediction on large-scale datasets. Additionally, LNR-based models fail to effectively map the correlation between the train operational data to the train travel time between stations. This is because, in real-world railway systems, operational features often lack a clear linear relationship with the train travel time between stations. In contrast, most tree-based models offer accurate and robust estimations of train delays, although their performance can vary across models. Tree-based models excel in train delay prediction because they effectively capture complex, non-linear relationships between diverse factors such as train services identifier, train operational records and station information.

However, while machine learning models predict train delays effectively, they lack guarantees on individual uncertainty. Therefore, conformal prediction is leveraged to quantify uncertainty, helping train operators assess how often true delays fall within predicted bounds for more reliable decision-makings. By apply-

Table 3. Train delay prediction model performance comparison.

Model	RMSE (s)	MAE (s)
RF	**261.59**	**14.71**
GB	**261.38**	22.99
DT	265.69	27.14
XGB	289.50	17.09
histGB	389.96	29.78
SVR	1869.18	185.83
LNR	1839.20	284.96
Lasso	1840.17	271.07
Ridge	1841.20	268.53

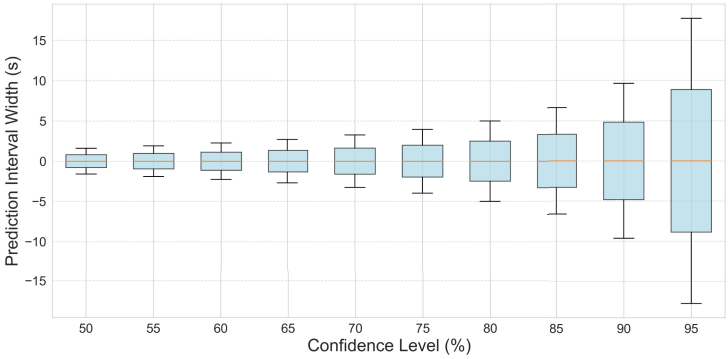

Fig. 6. The prediction interval widths (blue boxplot) at different confidence levels were evaluated within the conformal prediction framework. The orange line indicates the prediction generated by the ML model. At a 90% confidence level, the resulting interval width was 19.3 s, while the interval width increases to 35.5 s at the confidence level of 95%. This demonstrates the expected trade-off in conformal prediction, where higher confidence levels yield wider prediction intervals.(Color figure online)

ing conformal prediction with confidence levels of 90% and 95%, the resulting prediction intervals are 19.3 s and 35.5 s, respectively, as illustrated in Fig. 6. This demonstrates the expected trade-off in conformal prediction, where higher confidence levels yield wider prediction intervals. This means that for any train delay prediction $\hat{\Phi}_{test}$ made by the proposed approach, the probability that the true delay time falls within $[\hat{\Phi}_{test}$ - 9.65 s, $\hat{\Phi}_{test}$ + 9.65 s] and $[\hat{\Phi}_{test}$ - 17.75 s, $\hat{\Phi}_{test}$ + 17.75 s] is 90% and 95%, respectively.

To investigate the validity of the prediction intervals, the Residual Coverage Score is also analysed, as shown in Fig. 7. It illustrates the actual empirical coverage scores within each bin, compared to the target coverage rates of 90% (red dashed line) and 95% (yellow dashed line). It is observed that the conformal

prediction intervals achieve the expected coverage for lower predicted values below 263.4 s. However, they exhibit under-coverage in the highest prediction bin, where the predicted values exceed 425.4 s. This suggests that for larger and longer train delays, the model is unable to provide estimations with enough confidence.

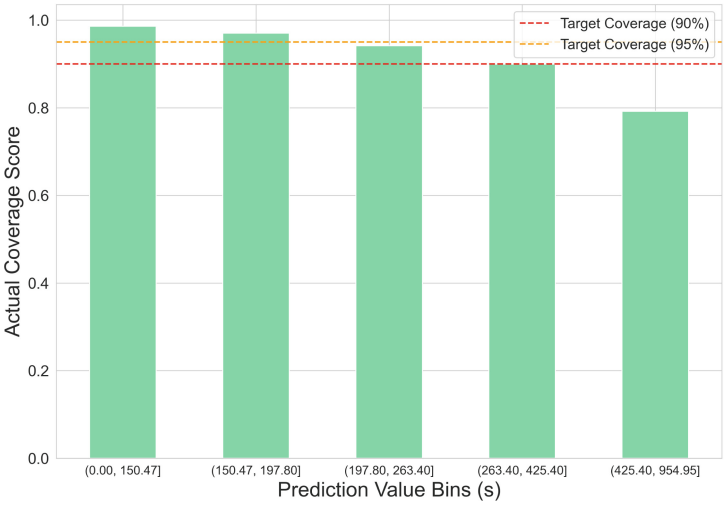

Fig. 7. The coverage scores of conformal prediction intervals across different predicted value bins. It is observed that the prediction interval achieve the desired coverage for lower predicted train delay values below 263.4 s, but under-cover for the predicted values above 425.4 s.

6 Conclusion

This study presents a robust machine learning approach for train delay prediction with quantified uncertainty. By modelling train travel times between consecutive stations using tree-based machine learning models, the proposed approach effectively captures sequential dependencies while ensuring efficient processing of large-scale datasets and facilitating the learning of underlying patterns within the dataset. The integration of Conformal Prediction provides statistically rigorous uncertainty quantification, delivering prediction intervals (e.g., 19.3 s at the 90% confidence level) that enhance decision-making transparency for railway operators. Validated on over 12.8 million real-world train production service records, the proposed method demonstrates an accuracy below 20 s, 90% of the time, outperforming widely used train delay prediction models such as LNR and SVR alternatives. While the framework achieves target coverage for typical delays below 263 s, extreme delays longer than 425 s exhibit under-coverage.

Further research could investigate methods for dynamically adjusting prediction intervals for individual test samples, particularly in relation to varying levels of delay severity.

Acknowledgments. The authors would like to thank Distributed Analytics Solutions Ltd for the financial support of this research.

References

1. Abdi, A., Amrit, C.: A review of travel and arrival-time prediction methods on road networks: classification, challenges and opportunities. PeerJ Comput. Sci. **7**, e689 (2021)
2. Arshad, M., Ahmed, M.: Train delay estimation in indian railways by including weather factors through machine learning techniques. Recent Adv. Comput. Sci. Commun. (Formerly: Recent Patents on Computer Science) **14**(4), 1300–1307 (2021)
3. Bao, X., Li, Y., Li, J., Shi, R., Ding, X.: Prediction of train arrival delay using hybrid elm-pso approach. J. Adv. Transp. **2021**, 1–15 (2021)
4. Barbour, W., Mori, J.C.M., Kuppa, S., Work, D.B.: Prediction of arrival times of freight traffic on us railroads using support vector regression. Transp. Res. Part C: Emerg. Technol. **93**, 211–227 (2018)
5. Barta, J., Rizzoli, A.E., Salani, M., Gambardella, L.M.: Statistical modelling of delays in a rail freight transportation network. In: Proceedings of the 2012 Winter Simulation Conference (WSC), pp. 1–12. IEEE (2012)
6. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
7. Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J.: Classification and Regression trees. Routledge (2017)
8. Chen, Z., Wang, Y., Zhou, L.: Predicting weather-induced delays of high-speed rail and aviation in china. Transp. Policy **101**, 1–13 (2021)
9. Corman, F., Kecman, P.: Stochastic prediction of train delays in real-time using bayesian networks. Transp. Res. Part C: Emerg. Technol. **95**, 599–615 (2018)
10. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
11. Davydov, B., Chebotarev, V., Kablukova, K.: Online train traffic adjustments: probabilistic modeling and estimating. In: Advanced Solutions of Transport Systems for Growing Mobility: 14th Scientific and Technical Conference Transport Systems. Theory & Practice 2017 Selected Papers, pp. 50–60. Springer (2018)
12. Dekker, M.M., Panja, D., Dijkstra, H.A., Dekker, S.C.: Predicting transitions across macroscopic states for railway systems. PLoS ONE **14**(6), e0217710 (2019)
13. Feng, X., Nguyen, K.A., Luo, Z.: A survey of deep learning approaches for wifi-based indoor positioning. J. Inf. Telecommun. **6**(2), 163–216 (2022)
14. Gao, B., Ou, D., Dong, D., Wu, Y.: A data-driven two-stage prediction model for train primary-delay recovery time. Int. J. Software Eng. Knowl. Eng. **30**(07), 921–940 (2020)
15. Gaurav, R., Srivastava, B.: Estimating train delays in a large rail network using a zero shot markov model. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 1221–1226. IEEE (2018)

16. Hauck, F., Kliewer, N.: Data analytics in railway operations: using machine learning to predict train delays. In: Operations Research Proceedings 2019: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Dresden, Germany, 4–6 September 2019, pp. 741–747. Springer (2020)
17. Huang, P., et al.: A bayesian network model to predict the effects of interruptions on train operations. Transp. Res. Part C: Emerg. Technol. **114**, 338–358 (2020)
18. Huang, P., et al.: Modeling train operation as sequences: a study of delay prediction with operation and weather data. Transp. Res. Part E: Logist. Transp. Rev. **141**, 102022 (2020)
19. Huang, P., Wen, C., Fu, L., Peng, Q., Li, Z.: A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. Saf. Sci. **122**, 104510 (2020)
20. Huang, P., Wen, C., Fu, L., Peng, Q., Tang, Y.: A deep learning approach for multi-attribute data: a study of train delay prediction in railway systems. Inf. Sci. **516**, 234–253 (2020)
21. Jiang, C., Huang, P., Lessan, J., Fu, L., Wen, C.: Forecasting primary delay recovery of high-speed railway using multiple linear regression, supporting vector machine, artificial neural network, and random forest regression. Can. J. Civ. Eng. **46**(5), 353–363 (2019)
22. Jiang, S., Persson, C., Akesson, J.: Punctuality prediction: combined probability approach and random forest modelling with railway delay statistics in sweden. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), pp. 2797–2802. IEEE (2019)
23. Kecman, P., Corman, F., Meng, L.: Train delay evolution as a stochastic process. In: 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015). IVT, ETH Zurich, Orange Labs (2015)
24. Kecman, P., Goverde, R.M.: Online data-driven adaptive prediction of train event times. IEEE Trans. Intell. Transp. Syst. **16**(1), 465–474 (2014)
25. Keyhani, M.H., Schnee, M., Weihe, K., Zorn, H.P.: Reliability and delay distributions of train connections. In: 12th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2012)
26. Laifa, H., Ghezalaa, H.H.B., et al.: Train delay prediction in tunisian railway through lightgbm model. Procedia Comput. Sci. **192**, 981–990 (2021)
27. Lemnian, M., Rückert, R., Rechner, S., Blendinger, C., Müller-Hannemann, M.: Timing of train disposition: towards early passenger rerouting in case of delays. In: 14th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2014)
28. Lessan, J., Fu, L., Wen, C.: A hybrid bayesian network model for predicting delays in train operations. Comput. Ind. Eng. **127**, 1214–1222 (2019)
29. Li, Y., Xu, X., Li, J., Shi, R.: A delay prediction model for high-speed railway: an extreme learning machine tuned via particle swarm optimization. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1–5. IEEE (2020)
30. Li, Z., Wen, C., Hu, R., Xu, C., Huang, P., Jiang, X.: Near-term train delay prediction in the dutch railways network. Int. J. Rail Transp. **9**(6), 520–539 (2021)
31. Liu, Y., Tang, T., Xun, J.: Prediction algorithms for train arrival time in urban rail transit. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. IEEE (2017)

32. Ma, H., Qin, Y., Han, G., Jia, L., zhu, T.: Forecast of train delay propagation based on max-plus algebra theory. In: Proceedings of the 2015 Chinese Intelligent Systems Conference, vol. 1, pp. 661–672. Springer (2016)
33. Martin, L.J.W.: Predictive reasoning and machine learning for the enhancement of reliability in railway systems. In: Lecomte, T., Pinger, R., Romanovsky, A. (eds.) RSSRail 2016. LNCS, vol. 9707, pp. 178–188. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-33951-1_13
34. Meister, J.A., Nguyen, K.A.: Conformalised data synthesis. Mach. Learn. **114**(3), 1–37 (2025)
35. Mou, W., Cheng, Z., Wen, C.: Predictive model of train delays in a railway system. In: Proceedings of the 8th International Conference on Railway Operations Modelling Analysis (RailNorrköping), pp. 913–929 (2019)
36. Nabian, M.A., Alemazkoor, N., Meidani, H.: Predicting near-term train schedule performance and delay using bi-level random forests. Transp. Res. Rec. **2673**(5), 564–573 (2019)
37. Nair, R., et al.: An ensemble prediction model for train delays. Transp. Res. Part C: Emerg. Technol. **104**, 196–209 (2019)
38. Nguyen, K., Luo, Z.: Conformal prediction for indoor localisation with fingerprinting method. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 214–223. Springer (2012)
39. Nguyen, K.A.: A performance guaranteed indoor positioning system using conformal prediction and the wifi signal strength. J. Inf. Telecommun. **1**(1), 41–65 (2017)
40. Nguyen, K.A., Luo, Z.: Enhanced conformal predictors for indoor localisation based on fingerprinting method. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 411–420. Springer (2013)
41. Nguyen, K.A., Luo, Z.: Reliable indoor location prediction using conformal prediction. Ann. Math. Artif. Intell. **74**(1), 133–153 (2015)
42. Nguyen, K.A., Luo, Z., Li, G., Watkins, C.: A review of smartphones-based indoor positioning: challenges and applications. IET Cyber-Syst. Robot. **3**(1), 1–30 (2021)
43. Nguyen, K.A., Luo, Z., Watkins, C.: Epidemic contact tracing with smartphone sensors. J. Location Based Serv. **14**(2), 92–128 (2020)
44. Nguyen, K.A., Watkins, C., Luo, Z.: Co-location epidemic tracking on london public transports using low power mobile magnetometer. In: 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–8. IEEE (2017)
45. Obayemi, A., Nguyen, K.A.: Uncertainty quantification of multimodal models. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 272–280. Springer (2025)
46. Oneto, L., et al.: Train delay prediction systems: a big data analytics perspective. Big Data Res. **11**, 54–64 (2018)
47. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. J. Artif. Intell. Res. **40**, 815–840 (2011)
48. Parbo, J., Nielsen, O.A., Prato, C.G.: Passenger perspectives in railway timetabling: a literature review. Transp. Rev. **36**(4), 500–526 (2016)
49. Pongnumkul, S., Pechprasarn, T., Kunaseth, N., Chaipah, K.: Improving arrival time prediction of thailand's passenger trains using historical travel times. In: 2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 307–312. IEEE (2014)
50. Pradhan, R., Kumar, A., Kumar, M., Sharma, B.: Simulating and analysing delay in indian railways. In: IOP Conference Series: Materials Science and Engineering, vol. 1116, p. 012127. IOP Publishing (2021)

51. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
52. Şahin, İ: Markov chain model for delay distribution in train schedules: assessing the effectiveness of time allowances. J. Rail Transp. Plann. Manage. **7**(3), 101–113 (2017)
53. Sara, L., Houda, J., Mohamed, A., et al.: Predict france trains delays using visualization and machine learning techniques. Procedia Comput. Sci. **175**, 700–705 (2020)
54. Seber, G.A., Lee, A.J.: Linear Regression Analysis. Wiley (2003)
55. Shafer, G., Vovk, V.: A tutorial on conformal prediction. J. Mach. Learn. Res. **9**(3) (2008)
56. Shi, R., Wang, J., Xu, X., Wang, M., Li, J.: Arrival train delays prediction based on gradient boosting regression tress. In: Proceedings of the 4th International Conference on Electrical and Information Technologies for Rail Transportation (EITRT) 2019: Rail Transportation Information Processing and Operational Management Technologies, pp. 307–315. Springer (2020)
57. Shi, R., Xu, X., Li, J., Li, Y.: Prediction and analysis of train arrival delay based on xgboost and bayesian optimization. Appl. Soft Comput. **109**, 107538 (2021)
58. Spanninger, T., Trivella, A., Büchel, B., Corman, F.: A review of train delay prediction approaches. J. Rail Transp. Plann. Manage. **22**, 100312 (2022)
59. Tiong, K.Y., Ma, Z., Palmqvist, C.W.: A review of data-driven approaches to predict train delays. Transp. Res. Part C: Emerg. Technol. **148**, 104027 (2023)
60. Tsolaki, K., Vafeiadis, T., Nizamis, A., Ioannidis, D., Tzovaras, D.: Utilizing machine learning on freight transportation and logistics applications: a review. ICT Express (2022)
61. Vafaei, S., Yaghini, M.: Online prediction of arrival and departure times in each station for passenger trains using machine learning methods. Transp. Eng. **16**, 100250 (2024)
62. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (2013)
63. Vovk, V.: Conditional validity of inductive conformal predictors. In: Asian Conference on Machine Learning, pp. 475–490. PMLR (2012)
64. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer Nature (2022)
65. Wang, P., Zhang, Q.p.: Train delay analysis and prediction based on big data fusion. Transp. Saf. Environ. **1**(1), 79–88 (2019)
66. Wang, Y., Wen, C., Huang, P.: Predicting the effectiveness of supplement time on delay recoveries: a support vector regression approach. Int. J. Rail Transp. **10**(3), 375–392 (2022)
67. Watanabe, S., Mori, Y., Takatori, Y., Yonemoto, K., Tomii, N.: Train traffic simulation algorithm based on historical train traffic records. Comput. Railways XVI pp. 285–32 (2018)
68. Wen, C., Huang, P., Li, Z., Lessan, J., Fu, L., Jiang, C., Xu, X.: Train dispatching management with data-driven approaches: a comprehensive review and appraisal. IEEE Access **7**, 114547–114571 (2019)
69. Wen, C., Mou, W., Huang, P., Li, Z.: A predictive model of train delays on a railway line. J. Forecast. **39**(3), 470–488 (2020)
70. Yaghini, M., Khoshraftar, M.M., Seyedabadi, M.: Railway passenger train delay prediction via neural network model. J. Adv. Transp. **47**(3), 355–368 (2013)
71. Zhang, D., Du, C., Peng, Y., Liu, J., Mohammed, S., Calvi, A.: A multi-source dynamic temporal point process model for train delay prediction. IEEE Trans. Intell. Transp. Syst. (2024)

72. Zhang, L., Feng, X., Ding, C., Liu, Y.: Mitigating errors of predicted delays of a train at neighbouring stops. IET Intel. Transport Syst. **14**(8), 873–879 (2020)
73. Zhou, P., Chen, L., Dai, X., Li, B., Chai, T.: Intelligent prediction of train delay changes and propagation using rvflns with improved transfer learning and ensemble learning. IEEE Trans. Intell. Transp. Syst. **22**(12), 7432–7444 (2020)
74. Zhuang, H., Feng, L., Wen, C., Peng, Q., Tang, Q.: High-speed railway train timetable conflict prediction based on fuzzy temporal knowledge reasoning. Engineering **2**(3), 366–373 (2016)
75. Zilko, A.A., Kurowicka, D., Goverde, R.M.: Modeling railway disruption lengths with copula bayesian networks. Transp. Res. Part C: Emerg. Technol. **68**, 350–368 (2016)