Transportation Research Part C 148 (2023) 104027

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A review of data-driven approaches to predict train delays ☆

Kah Yong Tiong [a], Zhenliang Ma [b,*], Carl-William Palmqvist [a,b]

[a] Department of Technology and Society, Lund University, Lund, 22100, Sweden
[b] Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, 114 28, Sweden

ARTICLE INFO

Keywords:
Train delay prediction
Data-driven prediction
Technical development
Railway operations and information

ABSTRACT

Accurate train delay prediction is vital for effective railway traffic planning and management as well as for providing satisfactory passenger service quality. Despite significant advances in data-driven train delay predictions, it lacks of a systematic review of studies and unified modelling development framework. The paper reviews existing studies with an explicit focus on synthesizing a structural framework that could guide effective data-driven train delay prediction model development. The framework consists of three stages including design concept, modelling and evaluation. The study synthesize and discusses six important modules of the framework: (1) Problem scope, (2) Model inputs, (3) Data quality, (4) Methodologies, (5) Model outputs, and (6) Evaluation techniques. For each module, the important problems and techniques reported are synthesized and research gaps are discussed. The review found that most studies focus on developing complex methodologies for the next stop delay predictions that have limited applications in practice. All studies validate the model accuracy, but very few consider other model performance aspects which makes it difficult to assess their usfulness in practical deployment. Future studies need a holistic view on defining the train delay prediction problem considering both application requirements and implementation challenges. Also, the modelling studies should place more attention to data quality and comprehensive model evaluations in representation power, explainability and validity.

1. Introduction

Following decades of increasing supply and demand, many railways are now operating at a high degree of capacity utilization. This increases the risk that the delay of one train will propagate to others, and that these delays will spread and grow over long distances. This can significantly degrade the performance and attractiveness of railways. Spanninger et al. (2022) categorized the train delay prediction approaches into event-driven and data-driven models based on their inherent modelling paradigms. Event-driven approaches have a train-event dependency structure, which entails the construction of a network of consecutive train events (departures, arrivals, and pass-through) over the prediction horizons. The event-driven train delay approach is an iterative process with a chain of prediction steps. Event-driven approaches are primarily based on an equation system (Medeossi et al., 2011) or a graph model such as Bayesian Networks (Corman and Kecman, 2018), Timed event graphs (Kecman and Goverde, 2014), Markov Chains (Schmidt et al., 2019), Petri Nets (Zhuang et al., 2016; Milinković et al., 2013), and Max-plus algebra (Goverde, 2007). On the other hand, data-driven approaches do not explicitly model train-event dependency structures nor intend to explicitly capture traffic flow dynamics. By mapping the input to output, data-driven approaches directly predict the delay at target stations or locations without intermediate predictions. In the review, we will only focus on data-driven approaches. This is primarily because along with

☆ This article belongs to the Virtual Special Issue on Advanced railway transpo.

* Corresponding author.
E-mail address: zhema@kth.se (Z. Ma).

https://doi.org/10.1016/j.trc.2023.104027
Received 30 April 2022; Received in revised form 10 January 2023; Accepted 13 January 2023
Available online 23 January 2023

the rapidly expanding volume of data in the railway industry and advancement of computing techniques, data-driven approaches have been increasingly adopted to develop train delay prediction models. This could be attributed to the significant opportunity offered by data-driven approaches to conduct in-depth analysis given their capability in handling large data sets and extracting valuable insights from ever-growing train operation databases.

Several review papers have been published focusing on AI and data-driven techniques in railway applications. For instance, Wen et al. (2019) provided a critical review on key issues in establishing different data-driven approaches for train dispatching management. Bešinović et al. (2021) introduced basic concepts and possible applications of AI to railway academics and practitioners, by presenting a structured taxonomy to understand AI techniques, research fields, disciplines, applications, ethics and explainability of AI. Tang et al. (2022) presented a systematic literature review of the current state-of-the-art of AI in railway transport from a holistic perspective, covering maintenance and inspection, planning and management, safety and security, autonomous driving and control, revenue management, transport policy, and passenger mobility. Spanninger et al. (2022) provided a synoptic review of diverse approaches (data-driven and event-driven) to predict train delays, data sources, the type of prediction (deterministic vs. stochastic), and prediction horizons under various modelling paradigms.

These literature reviews provide an overview of the application of AI in the railway sector and conceptual facets of data-driven approaches in train delay prediction. They basically provide information on what has been done using what approaches and data for what application contexts. However, it lacks a systematic review of train delay prediction from the model development perspective, including problem definition, inputs/outputs representations, modelling techniques, and model training and validation. In other words, we aim to review existing studies from the technical aspect and synthesize a unified model development framework with corresponding key problems and techniques for data-driven train delay prediction. Given the complicated train delay prediction process, we disaggregate it the into six aspects: (1) Scope determination (2) Model inputs (3) Data quality (4) Methodologies (5) Model outputs (6) Evaluation techniques. For each aspect, the important problems and techniques reported are synthesized and research gaps are discussed. This study will guide railway modellers and practitioners in developing data-driven prediction models step-by-step and be aware of important modelling aspects and technical options in different steps. For example, what are the techniques for validating a data-driven prediction model? Is the commonly used Mean Absolute Error (MAE) and Root Square Mean Error (RMSE) measures enough or do we need to explore further the prediction error distribution to assess the model bias and precision? What would be a valid problem definition for train delay prediction for passengers? Most studies predict the train delay for the next stop which may not be valid for information provision as passengers on board may go to different stops. Then, the valid problem definition should be to predict multi-stops train delays rather than the next-stop delay. In all, our review has a strong technical focus on streamlining the model development process used to 'correctly' and 'efficiently' develop a data-driven train delay prediction model, rather than simply reviewing conceptual facts as existing review papers do.

Note that the generic data science or data mining framework, e.g., CRISP-DM (Chapman et al., 2000), is a process model that serves as the base for guiding the design of a data driven project. It guides the development of the domain specific prediction framework for railway operation delays, which incorporates the domain knowledge (e.g., prediction problem definition, feature representations and selections) into data-driven modelling structure. It belongs to the Applied AI (a branch of AI community) – the art of using AI concepts in modelling and solving complex application problems. Compared to the generic data science framework, such a domain specific framework is necessary and valuable for predictive analytics to streamline the research findings and accelerate the uptake of data-driven approaches to advance practices in railways.

The remaining paper is organized as follows: Section 2 describes the literature review method. Section 3 discusses the application scope in terms of the level of implementation and types of prediction. Section 4 summarizes model inputs including model feature variables and feature selection methods. Section 5 discusses data quality issues in model training including noisy, missing and imbalanced data. Section 6 reviews prediction methodology studies and discuss them in four categories, i.e., statistical regression, machine learning, neural networks and hybrid models. Section 7 discusses two aspects of prediction outputs including output features and dimensions. Section 8 summarizes model evaluation techniques in assessing model accuracy, generalization, explainability and validity (i.e., model assumption test). Section 9 proposes a three-stage systematic framework for data-driven prediction model development based on the review and discusses future key research directions. Section 10 concludes the key findings and future research directions.

2. Literature review method

We conducted a literature search in March 2022 using Web of Science and Scopus databases. The search was restricted to academic journals and conference papers in English, with no restriction on the years of publication. In the first step, the keywords related to railway transport such as "train", "rail*", were used together with terms that capture train delay prediction such as "delay", "forecasting" and "prediction". To narrow it down toward data-driven approaches, phrases such as "data-driven", "machine learning", "regression", "artificial intelligence", "deep learning", "neural network", and "statistical regression" were used. All these strings were searched in title, abstract and keywords of the literature. The search results from both databases were combined, and duplicated results were filtered out, yielding 78 papers.

To ensure the quality and relevance of the reviewed papers, the second step involved a full-text review with a set of exclusion and inclusion criteria. Articles with no access to full-text versions or purely qualitative papers were excluded. Articles focused on other factors such as rail velocity instead of train events from the timetable perspective were excluded. Given the focus on data-driven train delay prediction, papers that emphasize purely mathematical perspectives, optimization, simulation, and queuing theory were also excluded. 37 papers remained in this step.

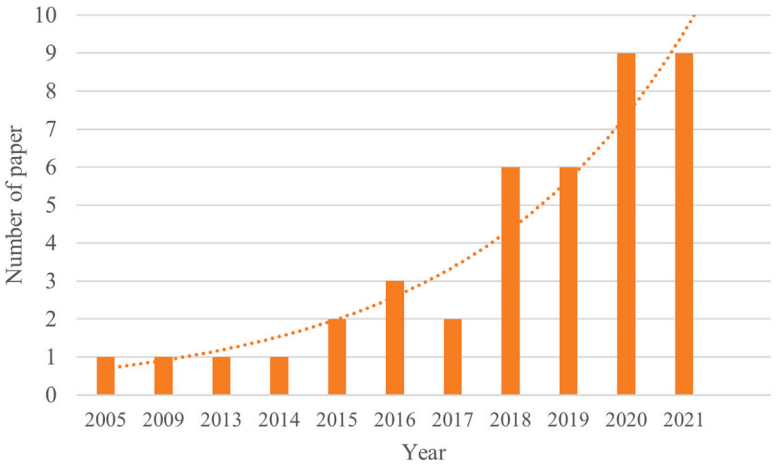

Fig. 1. Distribution of data-driven train delay prediction papers per year of publication.

Then, the forward and backward snowballing search strategies (see Wohlin, 2014) were applied. In backward snowballing, the reference list of each remaining paper was checked to identify additional papers, whereas in forward snowballing, new papers were identified based on those papers citing the paper being examined. To ensure only relevant articles were included, the titles, keywords, and abstracts were checked first, followed by a full text review. The backward snowballing process retrieved 7 new papers whereas forward snowballing retrieved 12 new papers, and both process ended at the first iteration when no new primary papers were found. Finally, a total of 56 papers were included in this study. Fig. 1 shows the distribution of the selected paper publications over year. It shows an continuously increasing interest in the data-driven train delay predictions. Table 1 categorizes literature based on their scope and methodology.

3. Application scope

The determination of application scope, which encompasses the model's level of implementation and type of prediction, is an prerequisite step to develop a data-driven train delay prediction model.

3.1. Level of implementation

Accurate train delay predictions have a well-established role in all levels of planning, control and management of railway traffic: strategic, tactical, and operational. At the strategic level, the train delay prediction model aids in infrastructure investment planning. The model assists planners in gaining a better understanding of the relationship between railway transportation efficiency and infrastructure investments or facility improvements, allowing planners to select the most cost-effective investment plan given budget constraints. Tactically, the accurate prediction of train movements is important for creating feasible timetables. At the operational level, train movements need to be predicted for real-time decision making and disruption management by dispatchers, especially with regard to timetable rescheduling, train re-sequencing, or rerouting, as well as to provide reliable passenger information to passengers for trip planning. Despite the potential of train delay predictions in driving strategic, tactical, and operational level applications, most studies focused on the operational level, with an emphasis on developing predictive models to assist decision-makers in developing effective management strategies.

3.2. Type of prediction

According to de Faverges et al. (2018), data-driven train delay prediction models can be categorized into two types based on prediction horizon: long-term and short-term delay prediction models. We further categorize the short-term prediction based on prediction horizons and updating methods as shown in Fig. 2.

Long-term delay prediction models are used at both the strategic and tactical levels. This type of model uses historical train operation data to derive insights that aid in the understanding the impact of endogenous and exogenous factors on railway system performance. The long-term delay prediction models use aggregated historical train operation data rather than real-time data to predict delays several days or even months in advance, giving train operators adequate time to develop train management plans. External factors such as weather, holidays, and seasons are often taken into account to examine the impact of these explanatory variables on train events, allowing train operators to adjust their plans accordingly when different scenarios are encountered.

Short-term prediction models are instead used mainly for the operational level and fed with real-time data. This type of research focuses on the applicability of prediction models in the real world. The prediction horizon is usually represented in space (e.g., next

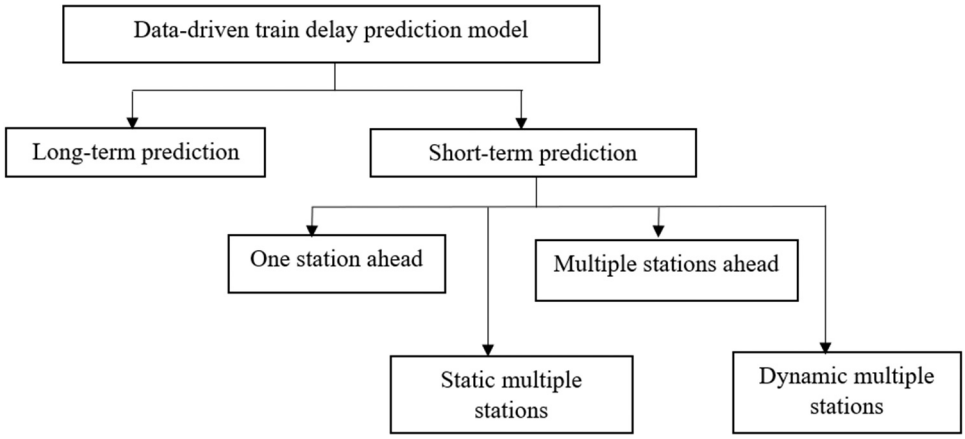

Fig. 2. Hierarchies of data-driven train delay prediction model in term of prediction horizon.

stop delay prediction) rather than time, due to the ease of handling the structure of the train operation data that is updated based on location.

In terms of space, the prediction horizon denotes the number of stations ahead to which the prediction refers. The short-term prediction models are further classified into four sub-groups based on prediction horizons and prediction updating methods, which are: (1) one station ahead prediction, (2) multiple stations ahead prediction, (3) static multiple stations prediction, and (4) dynamic multiple stations prediction. Suppose a train travels from an origin station, S_1, to a destination station, S_N. When it arrives at station S_i, $S_i \in \{S_1, S_2, \ldots, S_N\}$ at the prediction time, the one-station-ahead prediction problem is to predict the delay at the next station, S_{i+1}; multi-station-ahead prediction at multiple stations ahead, S_{i+q}, where $(q > 1)$; static and dynamic multiple stations predictions are to predict delays at all downstream stations, i.e., $S_{i+1}, S_{i+2}, \ldots, S_N$, by taking into account information at S_i and any previous station of S_i. The difference is that the static model makes 'one-shot' prediction with no update, while the dynamic model updates the predictions as railway traffic information evolves.

One of the limitations identified in recent studies is the emphasis on one station ahead prediction, which has limited applicability in practice. This gives rise to static and dynamic multiple stations predictions. Static multiple stations prediction is usually accomplished using a simulation approach in which the results from a data-driven model are fed into a simulation model to simulate train traffic at downstream stations (see Nair et al., 2019; Peters et al., 2005; Watanabe et al., 2018). Its goal is to reflect real-world conditions so that the proposed measures can be evaluated. Passengers require information for the station and train of interest given where they are now and update these predictions each time the train arrives at the next station, this is known as dynamic multiple stations prediction (Oneto et al., 2017, 2018). This dynamic models can act as a decision support tool to monitor delays in the rail network based on the decisions made in the current phase and allows both passengers and operators to have a clear picture of how long a train will take to complete the route.

It is worth noting that the prediction accuracy of a prediction model degrades as its prediction horizon is lengthened. In addition, short-term prediction models are usually developed to predict the train movements on a particular train line and limited in generalization to other train lines. After determining the application scope, Fig. 3 shows the model development workflow of train delay predictions. It includes inputs data, model building (data pre-processing and model selection), outputs data and model evaluations. The following sections review and discuss studies corresponding to these steps.

4. Model inputs

The study focuses on two aspects of model inputs, including feature variables and feature selection methods.

4.1. Feature variables

Table 2 summarizes the reported prediction variables and correponding data sources. Train operation data is the main data source for the delay prediction model. The train describer system is one of the operational data sources that store train movement information which records train arrival and departure timestamps at checkpoints by signalling systems or track-side sensors (Nair et al., 2019). Many train operating companies share real-time train operation data on websites (Ghofrani et al., 2018) and some studies scraped web data for train delay prediction (Pongnumkul et al., 2014; Wang and Zhang, 2019). However, the open-source data is still not fully utilized and explored, and more efforts are deserved for better use of this data since an abundance of open-source data can be expected in the future.

Table 1
Literature categorization based on scope and methodology.

Author	Scope				Methodologies		Model input
	Level	Purpose	Type	Prediction horizon	Method	Algorithm	Data size (train records)
Barbour et al. (2018a)	O	(a)	S	One stn	ML	SVR	4 200
Wen et al. (2017)	O	(f)	L		S; ML	LR; RF	29 662
Huang et al. (2019)	O	(a),(c)	S	Simultaneous time	ML	K-MC	6 006
Huang et al. (2020b)	O	(a)	S	One ,Multiple stn	HB	SVR; KF	57 796
Wen et al. (2020)	O	(a)	S	One stn	NN	LSTM	66 178
Lulli et al. (2018)	O	(a)	S	Dynamic prediction	HB	RT; RF	4 127 380
Marković et al. (2015)	T	(d),(f)	L		ML	SVR	727
Kecman and Goverde (2015)	S,T,O	(a),(c),(f)	L,S	One stn	S; ML	LR; RF; RT	145 807-101 481
Oneto et al. (2017)	O	(a)	S	Dynamic prediction	NN	SELM; DELM	More than 1-year
Li et al. (2021)	O	(a),(b)	S	20 min	ML	RF	433 402-5 491 362
Li et al. (2020b)	O	(a)	S	One stn	HB	ELM; PSO	More than 400000
Huang et al. (2020a)	O	(a)	S	One ,Multiple stn	HB	FCNN; LSTM	323 584-651,264
Taleongpong et al. (2020)	O	(a)	S	One ,Multiple stn	ML,NN	XGBOOST, ANN	39 911
Lee et al. (2016)	O	(a)	S	Dynamic prediction	HB	DT	4327
Li et al. (2020a)	O	(a)	S	One stn	HB	XGBOOST; SVR	3/2015 to 11/ 2016
Pongnumkul et al. (2014)	T	(e)	L	Dynamic prediction	S, ML	MA, K-NN	186
Peters et al. (2005)	O	(a)	S	Simultaneous stn	HB	S; NN	16 800
Yaghini et al. (2013)	T	(c)	L		NN	ANN	179 982
Nair et al. (2019)	O	(b)	S	Simultaneous stn	HB	S; KR; RF	50 millions
Nabian et al. (2019)	O	(a),(b)	S	20 min	HB	RF	180 000
Oneto et al. (2018)	O	(a)	S	Dynamic prediction	NN	SELM; DELM	6 months
Wang and Zhang (2019)	T	(b),(c),(d)	L		ML	" GBRT"	2 697 568
Huang et al. (2020c)	O	(a)	S	One ,Multiple stn	HB	CNN; LSTM; FCNN	27 825-229 320
Li et al. (2016)	O	(a)	S	One stn	S; ML	LR; K-NN	17 306
Gorman (2009)	S,T	(d),(f)	L	24H,12H,6H	S	LR	1/2001 to 8/2006
Watanabe et al. (2018)	S,T	(f), (d)	L	Simultaneous stn	HB	S; RT	Nils
Chen et al. (2021)	O	(a)	S	2H	ML,S	LR; DT; SVM and RF	2 025- 338 251
Li et al. (2022)	O	(a)	S	Dynamic prediction	HB	LSTM; FCNN	149 433-162 609
Gao et al. (2020)	O	(a),(b)	S	One stn	HB	RF	86 855
Bao et al. (2021)	O	(a),(c)	S	One stn	HB	ELM; PSO	400 000
Mou et al. (2019)	O	(a),(c)	S	One stn	NN	LSTM	66 working days
Laifa et al. (2021)	O	(b),(f)	S	One stn	ML	LightGBM	12 350
Huang et al. (2021)	O	(a)	S	One ,Multiple stn	HB	FCNN; CNN	242 680-346 112
Rößler et al. (2021)	S,T	(f)	L		ML	RF; MLP	1-year
Oneto et al. (2016)	O	(a)	S	Dynamic prediction	NN,ML	ELM; KM; RF	6 months
Ghaemi et al. (2018)	T	(e)	L		S,HB	GLM; LgR	23 000
Shi et al. (2021)	S,O	(a),(d)	S	One stn	HB	BO; XGBoost	13 507-16 9700
Barbour et al. (2018b)	O	(a)	S	One stn	NN,ML	SVR; DNN; RF	over 170 000
Grandhi et al. (2021)	T	(e),(f)	L		ML,NN,S	LR; XGBoost; GLM; ANN	6 693
Jiang et al. (2019)	S	(e),(f)	L		HB	HB	2 238 320
Oh et al. (2020)	O	(a),(c)	S	One stn	ML,S	SVR; RF; LR	247 000
Jiang et al. (2018)	T	(c), (f)	L		ML	SVR	1 450
Liu et al. (2022a)	T	(e)	L		ML	RF,KNN	more than 2 million
Luo et al. (2022b)	O	(a)	S	One stn	HB	DF	17 881- 47 115
Wu et al. (2021b)	O	(a)	S	One stn	ML	RF	161-day data
Meng et al. (2022)	O	(a),(b)	S	15 min	HB	SVM,KNN	31-day data
Shi and Xu (2020)	O	(a)	S	One stn	HB	XGBoost,BO	16 977
Wu et al. (2021a)	T	(a),(c)	L	1 day	HB	LSTM, CPS	161-day data
Zhang et al. (2021)	O	(a)	S	One stn	HB	KNN,GBDT	47 693
Liu et al. (2022b)	O	(a)	S	One stn	ML	XGBoost	516

(*continued on next page*)

Table 1 (*continued*).

Huang et al. (2022)	O	(a)	S	Dynamic prediction	HB	K-MC,BN	48 180 and 287 972
Tiong et al. (2022)	O	(a),(b)	S	Dynamic prediction	ML	LightGBM	2 240-6 414
Pradhan et al. (2021)	T	(b)	L		S	LR	data for yr 2016
Lapamonpinyo et al. (2022)	O	(b)	S	One stn	NN,ML,S	MLP,LR,RF,GBM	2008 to 2019
Luo et al. (2022a)	O	(a)	S	Dynamic prediction	HB	FCNN,LSTM	69 999
Ji et al. (2020)	O	(a),(b), (f)	S	One stn	ML	RF	639

Scope-Level: S = Strategic level; T = Tactical level; O = Operational level.
Scope-Purpose:(a) = Decision support for dispatchers; (b) = Inform the passengers, (c) = Timetable planning; (d) = Investment planning, (e)=train management plan; (f) = for better understanding railway system.
Scope-Type:L = Long term prediction, S = Short term prediction.
Scope-Horizon:stn = station, Dest = Destination, H = Hour.
Methodologies-Method: S = Statistical Regression Model; ML = Machine Learning Model; NN = Neural Networks Model; HB = Hybrid Model.
Methodologies-Algorithm: RF = Random Forest; ANN = Artificial Neural Networks; DT = Decision Tree; RT = Regression Tree; SVR = Support Vector Regression; GBR = Gradient Boosting Regression; GBRT = Gradient Boosted Regression Trees Model; XGBOOST = eXtreme Gradient Boosting;GBDT=Gradient Boosting Decision Tree; K-NN = K-Nearest Neighbours; LR = Linear Regression; ELM = Extreme Learning Machines; SELM = Shallow Extreme Learning Machines; DELM = Deep Extreme Learning Machines; LSTM = Long Short Term Memory; MA = Moving Average; CNN=Convolutional Neural Networks; FCNN = Fully Connected Neural Networks; KF = Kalmar Filter; PSO = Particle Swarm Optimization; KR = Kernel Regression; LightGBM = Light Gradient Boosting Machine; K-MC = K-Means Clustering Algorithm; MLP=Multi-Layer Perceptron; SVM = Support Vector Machine; S = Simulation; GLM=Generalized Linear Model; BO = Bayesian Optimization; DNN = Deep Neural Network; WIM = Weibull Intensity Model; BLR = Binomial Logistic Regression; SBLR = Semi-parametric Binomial Logistic Regression; CPS = Critical Point Search.

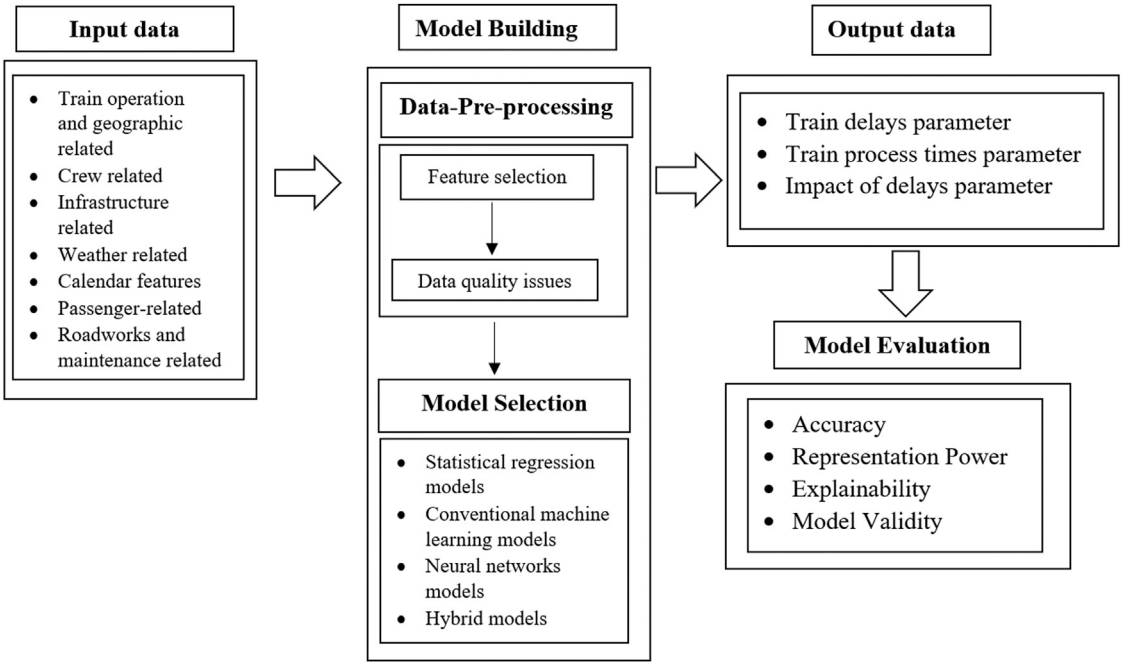

Fig. 3. The workflow for data-driven train delay prediction model.

Other influencing factors related to train operations are also reported contributing to train delay predictions, including crew schedule (Barbour et al., 2018a; Huang et al., 2020c; Nabian et al., 2019; Nair et al., 2019; Barbour et al., 2018b), rolling stock circulation (Huang et al., 2020c; Nair et al., 2019), infrastructure data (Barbour et al., 2018a; Nabian et al., 2019; Nair et al., 2019; Pongnumkul et al., 2014; Huang et al., 2020c; Barbour et al., 2018b) and passenger data (Lee et al., 2016; Nair et al., 2019; Oh et al., 2020). Train operations would also be affected by external factors, such as weather, peak hours, calendar features, roadworks and maintenance activities. For example, Oneto et al. (2017, 2016) report that incorporating the weather data improves the train delay prediction accuracy by about 10%. Li et al. (2016) show that dwell times are better estimated separately for peak and off-peak hours. Watanabe et al. (2018) argue the need to consider external factors of varied railway traffic conditions approximated by dummy variables such as months and day of the week (see Table 2).

According to Vlahogianni et al. (2004), treating traffic variables as a function of time and space is theoretically valid, given that temporal and spatial variables from previous locations can capture the dynamics of traffic and provide useful information about how it evolves. It is worth mentioning that train operation information at the station closest to the target prediction station provides

Table 2
Predictor variables from different data sources.

Variable category	Data source	Variables
Train operation and geographic related	Signalling systems (i.e., block systems: centralized traffic control (CTC), automatic block systems (ABS)), automatic train supervision (ATS), radio-frequency identification (RFID),planned train timetable,website (e.g. NSW's open data hub,General Transit Feed Specification (GTFS), Darwin's HSP API)	Train type, train length, train tonnage, train speed,train horsepower, train count, train direction, train priority and train order, arrival delays, departure delays, scheduled dwell time, actual dwell time and dwell time delays, scheduled run time, actual run time and run-time delays, travel time at individual sections between stations,buffer time, actual headways,scheduled headways, headway delays, train interaction(meet, pass, overtaking), station attributes,areas attributes, zone,railway checkpoint, railway section, the number of stops/distance to the destination station, distance travelled, per cent of the journey completed, and the number of stops/distance between consecutive stations"
Crew related	Crew timetable	Scheduled driver changes, crew time remaining, and on-duty time to departure
Infrastructure related	Automatic Vehicle Location (AVL) technology such as Global Positioning System (GPS)	The number of tracks, track segment occupancy, track occupation conflict indicators, track allocation, platform conflict indicators, designated platform, platform change status, the availability of sidings and the number of sidings
Weather related	Weather station, website(e.g. SMHI)	Temperature, wind speeds, snow depths, rainfall or precipitation
Calendar features	Planned train timetable	Times of day, days of the week, months of the year, holidays or working days, season, peak hours or off-peak hours
Passenger-related	Automatic Fare Collection, Automatic passenger counters, ticket sales, Load sensor	Total number of passengers, boarding passenger counts, and alighting passenger counts
Roadworks and maintenance related	Trackwork plan	Types of maintenance, time required for maintenance/roadwork

more useful information for predictions since the features used for prediction will change in time and space as trains travel along the route towards the destination (Barbour et al., 2018b).

The long-term prediction model adopts historical train operation data where all the information is observed. Thus, various types of explanatory variables are used to estimate train delay, with little emphasis on spatiotemporal data representation because the data is sorted at an aggregated level. The major goal is to understand the influence of each variable on train delays, providing theoretical support when making decisions for train schedule planning or adjustment as well as investing in infrastructure.

Spatial and temporal flow patterns of train operation data are frequently taken into account when developing short-term train delay prediction models. This is mainly because the incorporation of the spatiotemporal evolution of railway traffic information, for instance, the upstream or downstream train operation information, can be beneficial in boosting the prediction model's performance due to the interdependencies between train activities. To deliver an accurate prediction of the future railway traffic evolution on the network for real-time decision-making, short-term prediction models emphasize selecting variables that are beneficial from both spatial and temporal viewpoints. Thus, short-term prediction models are often fed with real-time data that can capture the latest actual railway traffic state as well as the inevitable disturbances. More specifically, real-time data that offers the most recent observations from the nearest station, S_i, or for a few stations ahead of S_i for the prediction at the subsequent stations S_{i+1} is one of the important data sources for short-term prediction.

4.2. Feature selection methods

Feature selection involves the selection of important input variables to eliminate redundant or irrelevant features. The majority of papers used subjective approaches, based on domain knowledge and common sense, to select/develop variables to incorporate in models that are believed to have predictive values. For instance, Barbour et al. (2018a) and Marković et al. (2015)determine the set of explanatory feature variables based on discussions with transportation operations experts.

The feature selection can also be accomplished using less subjective methods such as the filter method, wrapper method, and embedded method. The filter method selects features based on various statistical tests so that only input variables with statistically significant relationships to the target variable are selected. For instance, Pearson's Correlation (Wang and Zhang, 2019; Marković et al., 2015) and Pearson's chi-square test (Yaghini et al., 2013) are utilized to test the dependence between input variables and train delays. Using the correlation metric, Li et al. (2020b) select the top five variables that have statistically significant correlations with

the target variable, that is, the train arrival delay. To visualize the relationship between train delays and input variables, several studies perform exploratory analysis by plotting train delays with various input variables (Laifa et al., 2021; Oh et al., 2020; Grandhi et al., 2021).

In the wrapper method, the prediction models are selected based on the prediction performance. Generally, different prediction models are built with different subsets of input variables. Those input variables that contribute to the best performing model are selected. When dealing with a large dataset, the wrapper method is not suitable since it involves model training, which is computationally expensive. Some common wrapper methods are: forward feature selection, backward feature elimination, and recursive feature elimination methods. For example, Li et al. (2016) use forward feature selection, in which the first model is started with the fewest number of input variables, then one new input variable is added in each iteration until there is no improvement to the model's performance. Ghaemi et al. (2018) use a recursive feature elimination method, in which a greedy optimization algorithm continuously builds models by adding input variables into the feature subset or by deleting one from the current subset in each iteration with the aim of finding the best performing feature subset. Huang et al. (2020a) construct benchmark models with a subset of features based on categories of inputs (such as category for weather data, infrastructure data, etc.), whereas Barbour et al. (2018a) construct benchmark models by gradually adding categories of inputs into the feature subset.

In the embedded or intrinsic method, the feature selection process is embedded in the predictive model, enabling the model to automatically select input variables that maximize the model's accuracy. This includes tree- and rule-based models such as lasso, decision trees, and random forest. Feature importance is a built-in metric in tree-based models that is frequently used in train delay prediction studies to investigate the relevance of input features to the target variable (Li et al., 2021; Nabian et al., 2019; Chen et al., 2021). Shi et al. (2021) select final input variables with XGBoost feature importance greater than 5%. Instead of using tree-based models as the final prediction model, several studies only utilize its feature importance score to decide the final input variables fed into the non-tree based prediction model (Bao et al., 2021; Jiang et al., 2019) . It is worth noting that the correlation between each independent variable should be checked for multicollinearity (Grandhi et al., 2021; Wen et al., 2017).

5. Data quality

Data pre-processing is essential to ensure that a data-driven model is not trained on erroneous data. The data quality issues in train delay prediction include:(1) Noisy data (2) Missing data and (3) Imbalanced data. Table 3 listed the data quality issues reported in literature.

5.1. Noisy data

Noisy data are associated with outliers that lie far away from the main cluster of data. Abnormal data such as duplicate items, invalid data, and erroneous items are usually removed (Gorman, 2009; Huang et al., 2020a,b; Lee et al., 2016; Chen et al., 2021; Bao et al., 2021; Laifa et al., 2021; Rößler et al., 2021; Pongnumkul et al., 2014; Li et al., 2021). For instance, Li et al. (2021), Nabian et al. (2019) remove unusual delays when train cancellations should have occurred. Gao et al. (2020), on the other hand, correct faulty records by comparing them to train operation records from other days, whereas Huang et al. (2021) replace abnormal observations by using the mode of the respective variables.

5.2. Missing data

Empirical data often contains observations with missing values. Gorman (2009), Huang et al. (2020b), Lee et al. (2016), Chen et al. (2021) exclude the observations with missing values, since removing observations with some missing values does not impact the model's performance if they constitute a small proportion of the total data set (Rhys, 2020). Others proposed various techniques to impute the missing data. For instance, the missing data could be filled using weighted means (Li et al., 2021), unweighted means (Huang et al., 2020a, 2021), median (Laifa et al., 2021), or adjacent records (Gao et al., 2020; Bao et al., 2021). Taleongpong et al. (2020) utilized the imputation technique when there is only less than 10% missing data and discarded train journey data with more than 10% missing data.

5.3. Imbalanced data

Imbalanced or skewed data is due to the disproportionate ratio of observations in each class (Felix and Lee, 2019). This can lead to biased model training and evaluations, since it gives a higher probability of minority class mis-classification when compared to the majority class, causing the majority class to be overclassified (Johnson and Khoshgoftaar, 2019). Right-skewed running time data results in lower congestion delay estimation (Gorman, 2009) and makes the model unable to accurately predict unexpected situations (Huang et al., 2020b). Several studies state that train delay prediction is regarded as a challenging problem given the limited training sample for longer delays (Huang et al., 2020a; Shi et al., 2021).

Table 3
Categorization of literature based on quality of data.

Author	Noisy data	Missing data	Imbalanced data
Barbour et al. (2018a)	✓	✓	
Wen et al. (2017)			
Huang et al. (2019)			✓
Huang et al. (2020b)	✓	✓	✓
Wen et al. (2020)			
Lulli et al. (2018)			
Marković et al. (2015)			✓
Kecman and Goverde (2015)	✓	✓	
Oneto et al. (2017)			
Li et al. (2021)	✓	✓	
Li et al. (2020b)			
Huang et al. (2020a)	✓	✓	✓
Taleongpong et al. (2020)		✓	
Lee et al. (2016)	✓		
Li et al. (2020a)	✓		
Pongnumkul et al. (2014)	✓		
Peters et al. (2005)			
Yaghini et al. (2013)			
Nair et al. (2019)			✓
Nabian et al. (2019)			
Oneto et al. (2018)			
Wang and Zhang (2019)			
Huang et al. (2020c)	✓	✓	
Li et al. (2016)			
Gorman (2009)	✓	✓	✓
Watanabe et al. (2018)			
Chen et al. (2021)	✓	✓	
Li et al. (2022)	✓		
Gao et al. (2020)			
Bao et al. (2021)	✓	✓	
Mou et al. (2019)			
Laifa et al. (2021)	✓	✓	
Huang et al. (2021)	✓	✓	✓
Rößler et al. (2021)	✓	✓	
Oneto et al. (2016)			
Ghaemi et al. (2018)	✓		✓
Shi et al. (2021)		✓	✓
Barbour et al. (2018b)	✓		✓
Grandhi et al. (2021)	✓		✓
Jiang et al. (2019)	✓		
Oh et al. (2020)	✓	✓	
Jiang et al. (2018)	✓		
Liu et al. (2022a)	✓		
Luo et al. (2022b)	✓		✓
Wu et al. (2021b)			
Meng et al. (2022)			
Shi and Xu (2020)			
Wu et al. (2021a)			
Zhang et al. (2021)			
Liu et al. (2022b)	✓		

(*continued on next page*)

Table 3 (*continued*).

Author	Noisy data	Missing data	Imbalanced data
Huang et al. (2022)			✓
Tiong et al. (2022)	✓	✓	✓
Pradhan et al. (2021)			
Lapamonpinyo et al. (2022)			
Luo et al. (2022a)			
Ji et al. (2020)			

6. Methodologies

Methodologically, data-driven train delay prediction approaches can be categorized into four types, including statistical regression, conventional machine learning (ML), neural networks (NN), and hybrid methods. The following are observed from Table 1:

1. From 2005 to 2022, the prediction type shifts from long-term prediction models to short-term prediction models. This is reasonable since, at the initial stage, emphasis has been placed on the theoretical aspect, that is, understanding the importance of each explanatory factor on train delays. With the maturity of data-driven approaches, the recent development of technology and the advancement of computing system, the attention has shifted to the application of prediction models in real-world emphasizing the accuracy and robustness of the models.
2. Long-term prediction is progressively moving from a statistical approach to machine learning methods. Aside from being more accurate and not making assumptions about the distribution of underlying data as in statistical regression models, machine learning models, particularly tree-based models, have the advantage in providing insights into individual decisions based on underlying data by computing feature importance measures and ranks for each predictor variable.
3. There is a substantial boost in the number of publications on the short-term prediction, with a trendy preference over hybrid-based models. This is as expected, especially when dealing with large datasets with varying structures and different data-driven methodologies having advantages over different data types. The hybrid model can adapt to the complex information of railway traffic conditions through the combination of multiple models, resulting in a more robust train delay prediction model.

6.1. Statistical regression models

Statistical regression models, such as linear regressions, are mainly used to understand the variable impact on train delays, but also useful to predict them in real time. Gorman (2009) predicts the train's running time using linear regression and identifies primary congestion-related factors have the largest impact on congestion delay, such as train meeting, passing, and overtaking . The statistical regression model is simple to implement, and the result easy to interpret. However, it is limited in modelling the non-linear relationship between inputs and outputs (Osarogiagbon et al., 2021). Li et al. (2016) use non-parametric regression models for dwell time prediction in off-peak hours to better model the non-linearity compared to parametric regression models. Chen et al. (2021) state that the linear regression model has advantages over other data-driven methods when the sample size is relatively small and it also has less concern about overfitting. One drawback of statistical regression is that it frequently overlooks the complex relationship between different variables (Brnabic and Hess, 2021).

6.2. Conventional machine learning models

ML is used to uncover hidden knowledge by learning from relationships in historical data to produce a reliable and repeatable prediction (Wen et al., 2020). Supervised regression is the most common ML task in literature; commonly models used include support vector regression, random forest regression, and regression trees. There is no definitive evidence that certain algorithms in ML always outperform others, and the optimal delay prediction model is ultimately determined through algorithm comparison (Li et al., 2021). However, it is worth noting that the random forest regression is the most widely used ML model.

The capability of the ML models in capturing the nonlinear relationship between independent and dependent variables enables them to provide a better prediction results but less interpretability than statistical regression models. ML models are capable of handling high-dimensional and noisy railway data and can therefore provide more accurate decision support. Barbour et al. (2018a) develop a support vector regression model to predict individual freight train arrival times by taking inputs of the train properties, network characteristics, and potentially conflicting traffic in the network.

ML requires human-engineered spatiotemporal features to capture the spatial and temporal flow pattern of train operation when predicting train delays. On the other hand, NN can automatically learn spatiotemporal representations from the raw train operation data. According to Wang et al. (2020), spatial proximity and temporal correlations in train operation data can be learnt automatically from the raw data directly with the multi-layer convolution operation in Convolution Neural Networks (CNNs) and the recurrent structure of Recurrent Neural Networks (RNNs) respectively. Model framework transformations for ML models are also not as flexible

as the NN. This can be seen when both static and dynamic multiple stations train delay prediction of ML models necessitates a combination with other methods to form a hybrid model. For example, Nair et al. (2019) proposed a large-scale ensemble prediction model constituting random forest, kernel regression, and mesoscopic simulation to predict train delays for the nationwide passenger service network of Deutsche Bahn.

6.3. Neural networks models

The design flexibility of NN enables the development of a train delay prediction model based on the modeller's requirements. For instance, Li et al. (2020b), Oneto et al. (2018, 2017) propose the use of Extreme learning machines (ELM) consisting of a single hidden layer feedforward neural network to ensure high training speed and avoid overfitting problems when predicting train delays for a large railway network. NN enables the design of multivariate and multiple step-ahead train delay prediction models with the input data from multiple lagged times and spaces (Huang et al., 2021; Taleongpong et al., 2020).

Different NN algorithms are found to be effective in handling different types of data . For instance, LSTM or RNN is well-suited for analysing sequential data and time-series data; CNN for spatial data or image data; fully connected neural networks (FCNN) for cross-sectional data. Another advantage of NN is its flexibility in integrating different architectures to form a hybrid model which is efficient in the handling of heterogeneous and multi-attribute data in dynamical railway systems. NN is also capable of modelling highly non-linear relationships in a multivariate setting (Vlahogianni et al., 2004). It fits well the delay prediction in railway system in which inevitable interactions exist among stations and trains given the interlocking equipments, train operations (overtaking, crossing, and passing) and infrastructure sharing. The capability of the NN in capturing these interactions can model delay propagation patterns and ensure good performance in delay prediction tasks (Huang et al., 2021, 2020a,c; Wen et al., 2020). These successes also indicate that NN is better at handling complex temporal and spatial relationships, compared to typical ML models.

6.4. Hybrid models

Hybrid models consist of a mixture of different models. A hybrid model involves the combination of two or more base algorithms to build a more robust standalone algorithm. The interest in hybrid models has become more avid in recent years due to their ability to make best use of the complementary performance of original methods. For instance, support vector regression that is trained using offline data fails to capture real-time traffic variation, leading to undermined prediction during railway disruptions. To address this problem, Huang et al. (2020b) propose a hybrid model consisting a SVR model trained offline and a Kalman filter model to update and correct the SVR results using real-time information. The predictive performance of a hybrid model is also more robust, since the result is derived from multiple base models, with uncorrelated prediction errors, and all predictions are unlikely to fail at the same time. For instance, the performance of the hybrid model proposed by Oneto et al. (2020) consisting of a regression tree for predicting train running time and dwell time as well as random forest regression for predicting train delays and penalty costs associated with a delay, is better than each of its component regression models. Huang et al. (2020c) propose CLF-Net constituting 3D-CNN, LSTM, and FCNN to process spatio-temporal features, time-series variables, and non-time-series data, respectively, in predicting train delays. Each algorithm in CLF-Net complements and rectifies the weaknesses of the others when processing the multi-attribute data.

7. Model outputs

We discuss two aspects of prediction outputs that are closely related to the types of prediction models (Fig. 2), including output feature variables and output dimensions (single or multiple aspects of train delays). Table 4 summarized model outputs of reviewed studies.

7.1. Output feature variables

The most commonly used output variables in train delay prediction studies can be grouped into parameters related to train delays, train process times, and the impact of delays. Arrival delays dominate the field of train delay prediction. This can be due to the prediction of train arrival delays being the most direct way to capture the disturbances in the scheduled timetables. Due to the sharing of infrastructure between trains, accurate prediction of train process times such as arrival time, departure time, dwell time, and running time is important for timetable planning and adjustment, resolving conflicts between train paths, and providing reliable passenger information. Pongnumkul et al. (2014) propose two algorithms to predict passenger train arrival times at downstream train stations in order to provide more information to rail commuters and improve the service's usability. Apart from the variables related to delays and train process times, studies also predict variables that quantify the impacts of delays in other forms, such as the number of affected trains, total delayed time, total time of affected trains, recovery time, and penalty costs. These provide useful insights into the consequences of delays on operations and supports informed real-time train dispatching, especially under disturbed situations. For instance, Lulli et al. (2018) suggest the exploitation of a prediction system to select the best dispatching solution that leads to minimum train delays and penalty costs.

Table 4
Categorization of Literature based on outputs and evaluation techniques.

Author	Outputs		Evaluation techniques			
	Variable	Specification	Accuracy	Representational power	Explainability	Model validity
Barbour et al. (2018a)	PT	Arrival times	MAE		FI	
Wen et al. (2017)	ID	Delay recovery	RMSE	R2, F-test, and t-tests	LR	RD
Huang et al. (2019)	ID	NAT, TDT		K-S ,GF		
Huang et al. (2020b)	PT	Running times	MAE, MAPE, actual vs predicted plot			
Wen et al. (2020)	TD	Train delays	MAE, RMSE	GF		RD
Lulli et al. (2018)	TD, PT, ID	Running times, dwell times, train delays, and penalty costs	Average Accuracy			
Marković et al. (2015)	TD	Train delays		ANOVA, R2		RD
Kecman and Goverde (2015)	PT	Running times, dwell times	MSE	ANOVA,R2	FI,LR	
Oneto et al. (2017)	TD	Train delays	Average Accuracy			
Li et al. (2021)	TD	Train delays	MAE, RMSE, LESSTHAN		FI	
Li et al. (2020b)	TD	Train delays	RMSE, MAE	R2		
Huang et al. (2020a)	TD	Train delays	MAE, MAPE, ROC, AUC		SA	RD, CDF
Taleongpong et al. (2020)	TD, PT, ID	KPIs, train delays, dwell times and travel times	RMSE, MAE ,MAPE	GF,R2	SHAP	
Lee et al. (2016)	TD, PT, ID	Delay root cause	Accuracy		SA	
Li et al. (2020a)	ID	NAT, TTAT	Accuracy, ROC, AUC, MAE, MAPE, Lessthan i		FI	
Pongnumkul et al. (2014)	PT	Arrival times	MAE			
Peters et al. (2005)	TD	Train delays	MSE			
Yaghini et al. (2013)	TD	Train delays	Accuracy,C			
Nair et al. (2019)	TD	Train delays	RMSE, accuracy,C			
Nabian et al. (2019)	TD	Train's delay category, Train delays	F-score, RMSE, Accuracy score		FI	
Oneto et al. (2018)	TD	Train delays	Average Accuracy			
Wang and Zhang (2019)	TD	Train delays	Actual vs predicted plot			
Huang et al. (2020c)	TD	Train delays	RMSE, MAE			
Li et al. (2016)	PT	Dwell times	MAPE, RMSE	R2,GF	LR	RD
Gorman (2009)	TD	Train delays	MSE, MAPE, MAE,ME	R2	LR	
Watanabe et al. (2018)	PT	Dwell times,train delays	Actual vs predicted plot			
Chen et al. (2021)	TD	Train delays	MSE,RMSE,MAE	R2,GF	FI	
Li et al. (2022)	TD	Train delays	RMSE,MAE	R2		RD, CDF
Gao et al. (2020)	PT, ID	Buffer times, delay recovery	MAE	R2		
Bao et al. (2021)	TD	Train delays	RMSE, MAE	R2		FT, WSRT
Mou et al. (2019)	TD	Train delays	RMSE, MAE	GF		RD,CDF
Laifa et al. (2021)	TD	Train delays	RMSE, MAE	R2		
Huang et al. (2021)	TD	Train delays	precision, recall, F1 score, and Jaccard score			

(*continued on next page*)

Table 4 (*continued*).

Rößler et al. (2021)	TD	Additional delays	MAE	R2	SHAP	
Oneto et al. (2016)	TD	Train delays	Average Accuracy			
Ghaemi et al. (2018)	TD	Train delays	ROC	GF,AIC		CDF
Shi et al. (2021)	TD	Train delays	MAE,RMSE, C	R2,GF		RD,FT, WSRT
Barbour et al. (2018b)	PT	Arrival times	MAE			
Grandhi et al. (2021)	ID	TDT	RMSE	R2	FI	
Jiang et al. (2019)	TD	Punctuality	RMSE	R2		
Oh et al. (2020)	PT	Dwell times	Accuracy	GF		
Jiang et al. (2018)	PT	Dwell times	MSE	R2 ,GF		RD
Liu et al. (2022a)	PT	Arrival times	MAE,MAPE			
Luo et al. (2022b)	ID	ToA,NoC	MAE,RMSE	R2,GF		FT, WSRT
Wu et al. (2021b)	PT	Running Times, dwell time, train delays	RMSE,ME,MAE,SMAPE,RRSE			
Meng et al. (2022)	PT	Travel times	MAPE,RMSE,APE			
Shi and Xu (2020)	TD	Train delays	MAE,RMSE	R2,GF		
Wu et al. (2021a)	TD, PT	Train delays, running times and dwell times	MAE,RMSE,MAPE	R2		
Zhang et al. (2021)	TD	Train delays	MAE,RMSE	R2		RD
Liu et al. (2022b)	ID	Delay recovery	MAE			RD
Huang et al. (2022)	TD	Delay jumps categories, Train's delay category	MAE,MAPE,RMSE			CDF
Tiong et al. (2022)	PT	Arrival times	RMSE,MAE	R2		
Pradhan et al. (2021)	TD	Train delays		R2		
Lapamonpinyo et al. (2022)	ID	TDT	RMSE,MAE	R2	FI	
Luo et al. (2022a)	TD	Train delays	MAE,RMSE			RD
Ji et al. (2020)	TD	Train delays	MSE,RMSE	R2		

Output-Category: TD = Train delays; ID = impact of TD;PT=Train process times.
Output-Specification: NAT = Number of affected trains, TDT = Total delayed times, TTAT = Total time of affected trains, KPIs = Key performance indicators.
Evaluation techniques: Accuracy: MAE = Mean Absolute Percentage 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; RRSE = Root Relative Squared Error; APE = Absolute Percentage of Error.
Evaluation techniques-Representational power: ANOVA= Two-way analysis of variance, K- = Kolmogorov–Smirnov; AIC = Akaike Information Criterion; H-L = Hosmer-Lemeshow test; GF = Goodness-of-fit plots .
Evaluation techniques-Explainability: FI = Feature Importance; LR = Linear Regression Coefficient; SHAP = SHapley Additive exPlanations, SA = Sensitivity Analysis.
Evaluation techniques-Validity: RD = Residual Distribution Plot; CDF = Cumulative Distribution Functions; FT = Friedman test; WSRT = Wilcoxon Signed-Ranks test.

7.2. Output dimensions

A majority of the articles only consider a single aspect of the train movements, giving a single value output. Attention must be given to the attempts made to predict multiple output variables. NN algorithms such as SELM and DELM, as well as hybrid models, are often used to simultaneously predict multiple output variables of the same type, e.g., delay times. However, simultaneously predict multiple outputs with multiple types is rarely reported. Kecman and Goverde (2015) develop global and local predictive models for both running and dwell times separately by training the models with relevant predictor variables for each process type. Li et al. (2020a) use two different algorithms since the results indicate that the eXtreme Gradient Boosting(XGBOOST) algorithm has the best predictive performance for predicting the number of affected trains, whereas the support vector regression algorithm was best for predicting the total delay time of affected trains. Overall, separate models are utilized when different types of output variables are to be predicted to ensure optimal model performance.

8. Model performance evaluation

The model performance includes model training and testing performance (internal validation), and prediction performance in terms of prediction accuracy, robustness and generalizability (external validation). Regardless of the type of model performance, indicators such as MAE and RMSE are commonly used given they are comparable among models. Other indicators are also reported, such as R^2.

8.1. Internal and external validation

Since the internal validation is part of the model development, almost all existing literature performed internal validation by splitting the dataset (from which the model was derived) into training and testing datasets. The commonly used internal validation strategies include split-sample (Li et al., 2016, 2021; Oneto et al., 2017; Wen et al., 2017), cross-validation (Barbour et al., 2018a; Huang et al., 2020a,b; Taleongpong et al., 2020), and bootstrapping (Jiang et al., 2019).

After a prediction model is developed, the external validation further validates the model performance using data different from the data used for model development. This is to provide evidence of the model's generalizability to different circumstances. However, very few existing studies have conducted external validation, possibly because the researchers overlooked its significance or misunderstood that randomly splitting a single dataset into model training and testing datasets is a form of external validation. For example, Huang et al. (2020c) and Bao et al. (2021) investigate the model's robustness to data size and compare their models with the benchmark models by training and testing the model using the development dataset but with varying sizes. This in fact is an inefficient internal validation since not all available data is used for model development. Another inefficient evaluation method is the supplementary validation utilizing development datasets with different settings rather than carrying out an external validation. For example, Shi et al. (2021) select data related to unusual events from the model development datasets validate the model's robustness to long delays, including the railway catenary failure, automatic train protection (ATP) system failure, and object intrusion. This might result in a biased model evaluation with an optimistic estimate of performance since the prediction model is exposed to the data during training and testing.

If the available development data set is sufficiently large, validation of the prediction model using a dataset at a different time point (temporal validation) or different place (geographical validation) might be a preferable option, since it provides some insight into the generalizability of a model. For instance, Li et al. (2020a) utilized train operation data for the period from March 2015 to November 2016 to train and validate the model, whereas data from 2018 were used to test the application of model.

8.2. Evaluation techniques

Evaluation techniques are used to assess the model accuracy, representational power (generalization), explainability, and model validity (e.g., model assumption test). Table 4 summarizes the evaluation techniques reported in literature.

8.2.1. Accuracy

Most studies focus purely on prediction accuracy to assess the model performance. To evaluate the model accuracy, the prediction error is calculated as the difference between the predicted value and the actual value. MAE is the most commonly used prediction error measurement, followed by the RMSE, the Mean Absolute Percentage Error (MAPE), and the Mean Square Error (MSE). The closer these values are to zero, the better the performance of the model. It is not surprising that MAE and RMSE are popular since they are both on the same scale as the dependent variable, making them easy to compare between models. One key difference between them is that RMSE gives a high weight to large prediction errors. The MAE and RMSE can be used together to represent the variation in prediction errors; that is, the greater the difference between the MAE and the RMSE, the greater the variation. The overall quality of a regression model cannot be determined based on the aggregated measures of MAPE, MSE, RMSE, or MAE since they all have $+\infty$ as the upper bound. Thus, benchmark models are used to evaluate the performance of the proposed model against other predictive models.

8.2.2. Representational power

A model demonstrating good prediction accuracy might not generalize well, which can render future predictions. Thus, statistical diagnostics for testing a model's goodness of fit are needed to ensure the model structure is adequate. The presence of "strong" statistical properties (e.g., serial correlation, changing variance, etc.) in the error term indicates the presence of bias due to omitted explanatory variables or that the functional form is incorrect (Thomaidis and Dounias, 2012). The most common tests for representational power are R^2. The R^2 quantify how well the variability in the dependent variable is explained by the independent variable in the model. The closer R^2 is to 1, the better the performance of the model is likely to be.

8.2.3. Explainability

The desire to explain certain phenomena in justifying the right model for the use case renders the evaluation metrics no longer sufficient to characterize the model. To explain a model, domain knowledge is required to analyse and comprehend results generated by the model. In the case of linear regression, the regression coefficients in the model equation are sufficient to explain a given prediction. Other data-driven models are often treated as blackboxes because of the less understandable model structure designed to adapt to the complicated data relationships. These models with stronger predictive capabilities have unexplainable outcomes, leaving researchers and practitioners unsure of the actual reasoning behind a prediction.

The importance of features is commonly computed to explore the impact of each explanatory feature on the predictability of the target variable and to understand the rationale behind the model's decisions. However, this approach is unable to assist decision-makers in gaining more in-depth insight or developing an effective response when confronted with different events. Tools such as the Local Interpretable Model-agnostic Explanations (LIME) algorithm (Ribeiro et al., 2016) and SHAP framework (Lundberg et al., 2018) are developed to interpret the black-box model without sacrificing the predictive accuracy of the model. By using the SHAP framework to explore the contribution of each feature to delay propagation mechanisms, Taleongpong et al. (2020) conclude that the greater the departure delay of the train from its previous station, the greater the impact on the predicted arrival delay of the train at the following station.

8.2.4. Model validity

Model validity test is critical to assess the degree of matching of the modelling framework assumption and the characteristics of the problem or data of interest. For example, to assess if a linear model would be fit for a specific data modelling, the assumption test on regularity and linearity of the data is required. Data-driven approaches, except for statistical regression models, often disregard the importance of model validity tests. Various error specification tests, such as serial independence, constancy of variance and symmetry, and neglected non-linearity, used by Vlahogianni and Karlaftis (2013), can be adopted as diagnostic checking to safeguard against systematic model bias, such as residual distribution and scatter plots of predicted values and errors. The residual distribution is plotted to ensure that the residuals are mostly near zero since a model with adequate structure should have white noise residuals, that is, the residuals are independently identically distributed with zero mean and constant variance (Washington et al., 2020)

8.3. Performance comparison

Performance comparison between various models is not explored in this study since the basis of comparison between different data-driven approaches is hardly established. The main reason is that researchers tend to develop their own prediction algorithms using their own datasets, which makes their work difficult to verify and duplicate. As a result, it is challenging to establish a fair comparison between different models, especially when the empirical results is derived from specific confidential datasets with specific size and complexity, and when prediction algorithms are difficult to replicate. Furthermore, if a single study attempts to employ a wide range of prediction algorithms to compare the performance of models, their research will deviate from the main objective. Lastly, model performance is not comparable between studies since not all studies have the same basis of comparison, i.e., no standard units of prediction error are employed, whether in terms of model accuracy or representation power.

While increasing efforts have been devoted to developing more advanced and complex prediction models, there is a lack of benchmarking datasets for comparing the efficiency of different models. Benchmarking datasets are critical for giving fair model performance comparisons. When researchers are developing prediction models using datasets of varying size and complexity, their results are hard to validate and replicate. Even while the development of potentially superior train delay prediction models has increased in recent years, there is always doubt regarding the applicability of such results, especially when there are no standard datasets available to assure that the prediction models improve with time. Apart from that, the availability of shared benchmarking datasets would enable researchers to undertake more tests and advance the train delay prediction studies, as the preparation of datasets is a tedious work requiring a significant amount of effort and time.

9. Discussion

Inspired by Vlahogianni et al. (2004), we develop a data-driven train delay prediction framework with three stages: design concept, modelling, and evaluation, considering the aforementioned six aspects (Fig. 4).

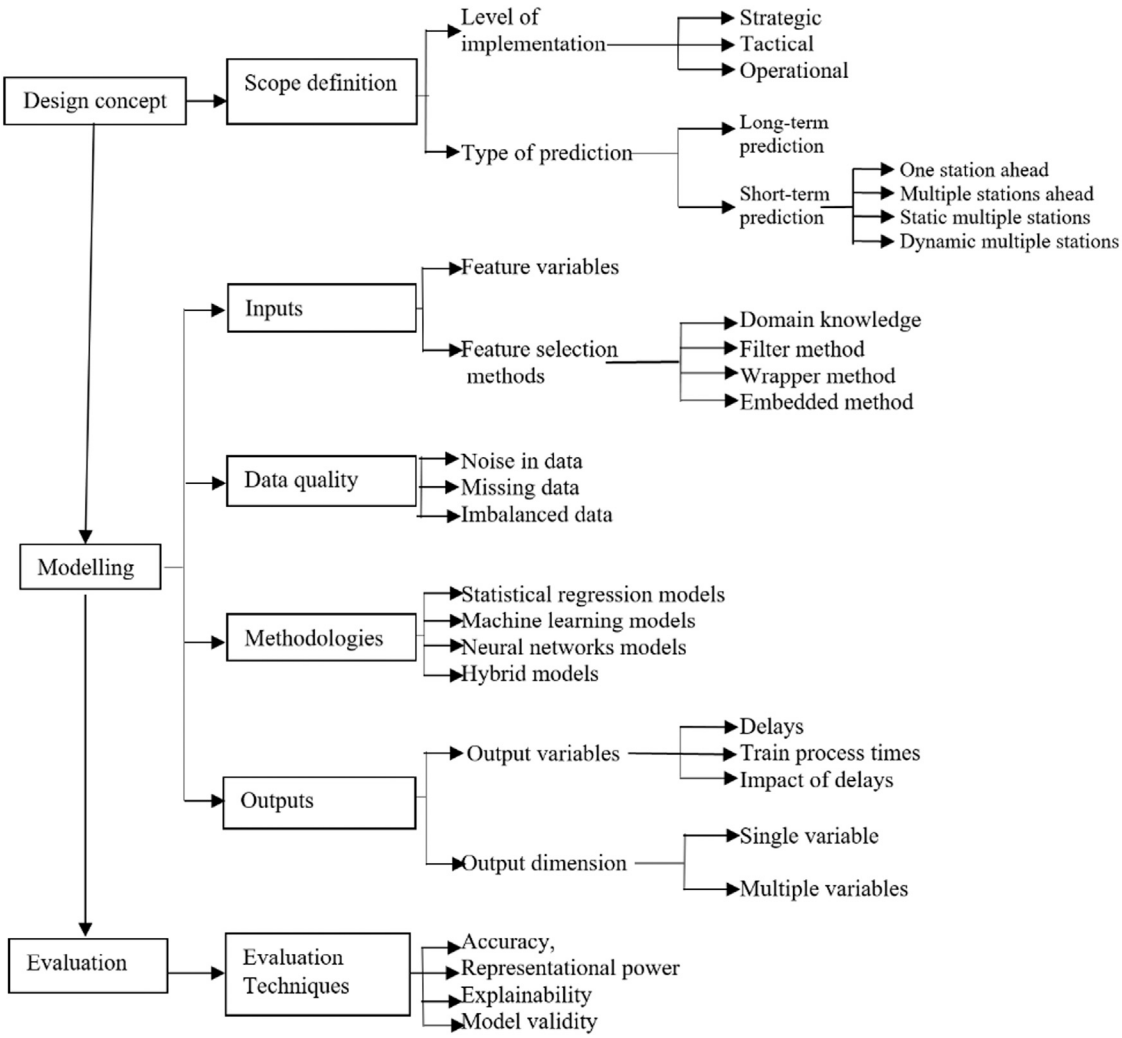

Fig. 4. Train delay prediction model development framework.
Source: Adapted from Vlahogianni et al. (2004).

9.1. Concept design

The design concept stage establishes the fundamental characteristics of the prediction model that will govern the subsequent modelling process. It determines the scope of modelling in terms of level of implementation and types of prediction.

The selection of implementation level and type of prediction interacts highly with the determination of the model's outputs. For instance, predicting variables related to delays and train process times is conceptually more useful at the strategic and tactical levels, whereas variables related to delays and the impact of delays at operational levels. At the strategic and tactical levels, accurate prediction of train process times through a long-term prediction approach enables the development of robust timetables, whereas accurate prediction of train delays enables effective investment plan determinations. On the operational level, the short-term prediction of delays and impacts of delays help operators to understand the impact of interruptions on train services, enabling informed timetable rescheduling. The short-term prediction model is fed with real-time data to generate predictions and is constantly updated based on railway traffic evolution. However, existing studies mainly focus on predicting the next station's train events. One interesting research direction is to predict simultaneously train events at multiple stations at arbitrary times regardless of where the train is.

9.2. Modelling process

The modelling stage develops the prediction model by taking inputs of the designed concepts. It includes model inputs, data quality, methodologies and model outputs.

In term of inputs, many studies only consider train operation related variables as inputs. Though they reported good prediction performance, other factors affecting railway operations (e.g., drivers' behaviour, passengers' volumes, infrastructure characteristics, route topography) and exogenous information (e.g., weather, holidays, etc.) from should be considered for train delay predictions to further improve the prediction accuracy and generalization under varied scenarios. However, most of these data are not yet publicly available, thus are not yet fully exploited for the training of data-driven prediction models.

In terms of data quality, merging different data to extract insights is often a challenging task. Data-quality related issues such as class imbalance, missing data, and noise in data are often encountered. Thus, data pre-processing is essential to ensure that models are fed with useful information. Data imbalance problems are discussed by some studies but remain unsolved. The utilization of imbalanced data or data with long-tailed distribution has led to considerable challenges, such as poor performance for long-delayed prediction or necessitating the use of significantly more data to train the predictive model. More work needs to be done in dealing with data quality issues in data-driven model training.

In terms of methodologies, NN and hybrid models are key trends for train delay prediction given their capabilities in capturing the temporal and spatial evolution of railway traffic characteristics, especially the interactions between different stations, sections, and trains. The flexibility in adjusting the architecture of these models to adapt to various types of input and output data facilitate the development of dynamic train delay prediction models. However, NN models may have a risk of overfitting when the datasets are not sufficiently large enough leading to poor model generalization. Essentially, a complex model is not necessarily the best solution depending on the matching degree between model assumptions and data problem characteristics. Instead of 'blindly' pursuing more advanced and complex models, the logic behind selecting the appropriate methodology for modelling train delay prediction models is also worth exploring.

9.3. Evaluation

Assessment of the model's performance solely on the datasets where it was developed is insufficient since the model is tailored to the development datasets. External validation, which entails validating the prediction model using a dataset that differs from the datasets used for model development is recommended to ensure the prediction model is transportable to a scenario with different railway traffic conditions. For instances, Wen et al. (2020) developed the train delay prediction model using data from the Rotterdam Central to Dordrecht section of the Dutch railway system. To verify the generalization and robustness of the model on different railway sections, Wen et al. (2020) then tested the model performance on the Rilland Bath-Vlissingen section of the railway system. Although many models are proposed in the literature, very few are externally validated. Testing the prediction model on new datasets, the so-called external validation, is imperative to avoid decision-related railway traffic management based on incorrect prediction models. For instance, if train operators make real-time rescheduling based on a prediction model that underpredicts delay risk, lower punctuality of trains can result. Thus, external validation is advised prior to the prediction model's deployment for practical use, either for passenger information or real-time traffic management, since it bridges the gap between the development and implementation of the train delay prediction models in the real world.

Most studies focus on evaluating the model accuracy in model development. Future studies should also consider model representation power, explainability and model validity. The efforts to handle high-dimensional data and to capture the spatio-temporal evolution of railway traffic characteristics have encouraged the application of more advanced data-driven approaches (NN and hybrid models). The structure of these models is adjusted to adapt to the complicated data relationships, forming the complex "black-box" model to achieve high accuracy. Developing algorithms to facilitate the interpretation of these models may be fruitful in generating insights, understand the relationships between inputs and outputs as well as to investigate the causalities rationale behind the model's decisions.

10. Conclusion

Predictive analytics is valuable for improving productivity in a wide range of railway applications, from train operation planning, control to management and passenger information provision. In this paper, we first provide a comprehensive survey of the applications of predictive analytics in railway operation delays (model development and techniques), to highlight the value and modelling challenges of these techniques. The papers surveyed generally offer point solutions applying specific predictive modelling techniques for particular application scenarios, and the modelling framework and reported findings are inconsistent across studies which highly hampered the knowledge transfer and advancement in the railway area. To address these, following the generic data science process model, we propose a domain-specific model development framework for predictive analytics of railway train delay data.

The goal for this framework is to integrate the problems and techniques for predictive analytics with a domain-specific modelling environment that makes problem specification and model development easier for railway domain experts. Different from existing survey studies (conceptual facts), we have a strong technical focus on streamlining the model development process used to 'correctly' and 'efficiently' develop a data-driven train delay prediction model. We also discussed the key modelling components and their corresponding problem definitions, challenges, and up-to-date techniques that would serve as guiding maps for model development.

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

- Developing dynamic multiple outputs train delay prediction models at arbitrary prediction times through data-driven approaches for practical applications like real-time traffic management or passenger information systems.

- Adopting multiple-source data for the development of a multi-regime train delay prediction model. Extension of the train delay prediction model by incorporating variables other than train operation-related variables is crucial to account for the impacts of unusual or sporadic occurrences such as accidents or adverse weather conditions on the prediction.
- Developing systematic data pre-processing techniques to handle data quality issues, especially noise, missing and imbalance to ensure that models are fed with useful information. All researchers must systematically disclose their data cleaning process to allow the evaluation of the effectiveness of each technique with regard to different data quality issues.
- Evaluating the prediction models from different perspectives, including accuracy, representation power, explainability, and model validity, to ensure the reliability of future predictions and to assess the input–output relationship. To ensure the satisfactory performance of the train delay prediction models in real practice, external validation is recommended, especially when the sufficiency of internal validation is debatable.

CRediT authorship contribution statement

Kah Yong Tiong: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Zhenliang Ma:** Conceptualization, Methodology, Formal analysis, Supervision, Writing – review & editing. **Carl-William Palmqvist:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was funded by the Swedish Transport Administration, Sweden, grant number TRV2018/139443.

References

Bao, X., Li, Y., Li, J., Shi, R., Ding, X., 2021. Prediction of train arrival delay using hybrid ELM-PSO approach. J. Adv. Transp. 2021.
Barbour, W., Mori, J.C.M., Kuppa, S., Work, D.B., 2018a. Prediction of arrival times of freight traffic on US railroads using support vector regression. Transp. Res. C 93, 211–227.
Barbour, W., Samal, C., Kuppa, S., Dubey, A., Work, D.B., 2018b. On the data-driven prediction of arrival times for freight trains on us railroads. In: 2018 21st International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 2289–2296.
Bešinović, N., De Donato, L., Flammini, F., Goverde, R.M., Lin, Z., Liu, R., Marrone, S., Nardone, R., Tang, T., Vittorini, V., 2021. Artificial intelligence in railway transport: taxonomy, regulations and applications. IEEE Trans. Intell. Transp. Syst.
Brnabic, A., Hess, L.M., 2021. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Med. Inform Decis Mak. 21 (1), 1–19.
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al., 2000. CRISP-DM 1.0: Step-by-Step Data Mining Guide, Vol. 9, No. 13. SPSS Inc, pp. 1–73.
Chen, Z., Wang, Y., Zhou, L., 2021. Predicting weather-induced delays of high-speed rail and aviation in China. Transp. Policy 101, 1–13.
Corman, F., Kecman, P., 2018. Stochastic prediction of train delays in real-time using Bayesian networks. Transp. Res. C 95, 599–615.
de Faverges, M.M., Russolillo, G., Picouleau, C., Merabet, B., Houzel, B., 2018. Estimating long-term delay risk with generalized linear models. In: 2018 21st International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 2911–2916.
Felix, E.A., Lee, S.P., 2019. Systematic literature review of preprocessing techniques for imbalanced data. IET Softw. 13 (6), 479–496.
Gao, B., Ou, D., Dong, D., Wu, Y., 2020. A data-driven two-stage prediction model for train primary-delay recovery time. Int. J. Softw. Eng. Knowl. Eng. 30 (07), 921–940.
Ghaemi, N., Zilko, A.A., Yan, F., Cats, O., Kurowicka, D., Goverde, R.M., 2018. Impact of railway disruption predictions and rescheduling on passenger delays. J. Rail Transp. Plan. Manag. 8 (2), 103–122.
Ghofrani, F., He, Q., Goverde, R.M., Liu, X., 2018. Recent applications of big data analytics in railway transportation systems: A survey. Transp. Res. C 90, 226–246.
Gorman, M.F., 2009. Statistical estimation of railroad congestion delay. Transp. Res. E Logist. Transp. Rev. 45 (3), 446–456.
Goverde, R.M., 2007. Railway timetable stability analysis using max-plus system theory. Transp. Res. B 41 (2), 179–201.
Grandhi, B.S., Chaniotakis, E., Thomann, S., Laube, F., Antoniou, C., 2021. An estimation framework to quantify railway disruption parameters. IET Intell. Transp. Syst.
Huang, P., Li, Z., Wen, C., Lessan, J., Corman, F., Fu, L., 2021. Modeling train timetables as images: A cost-sensitive deep learning framework for delay propagation pattern recognition. Expert Syst. Appl. 177, 114996.
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 Trans. Intell. Transp. Syst.
Huang, P., Wen, C., Fu, L., Lessan, J., Jiang, C., Peng, Q., Xu, X., 2020a. Modeling train operation as sequences: A study of delay prediction with operation and weather data. Transp. Res. E Logist. Transp. Rev. 141, 102022.
Huang, P., Wen, C., Fu, L., Peng, Q., Li, Z., 2020b. A hybrid model to improve the train running time prediction ability during high-speed railway disruptions. Saf. Sci. 122, 104510.
Huang, P., Wen, C., Fu, L., Peng, Q., Tang, Y., 2020c. A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems. Inform. Sci. 516, 234–253.
Huang, P., Wen, C., Peng, Q., Jiang, C., Yang, Y., Fu, Z., 2019. Modeling the influence of disturbances in high-speed railway systems. J. Adv. Transp. 2019.
Ji, Y., Zheng, W., Dong, H., Gao, P., 2020. Train delays prediction based on feature selection and random forest. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 1–6.
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. J. Transp. Eng. A Syst. 144 (11), 04018071.

Jiang, S., Persson, C., Akesson, J., 2019. Punctuality prediction: combined probability approach and random forest modelling with railway delay statistics in Sweden. In: 2019 IEEE Intelligent Transportation Systems Conference. ITSC, IEEE, pp. 2797–2802.
Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 1–54.
Kecman, P., Goverde, R.M., 2014. Online data-driven adaptive prediction of train event times. IEEE Trans. Intell. Transp. Syst. 16 (1), 465–474.
Kecman, P., Goverde, R.M., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transp. 7 (3), 295–319.
Laifa, H., Ghezalaa, H.H.B., et al., 2021. Train delay prediction in Tunisian railway through LightGBM model. Procedia Comput. Sci. 192, 981–990.
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 J. Intell. Transp. Syst. 3, 539–550.
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. Transp. Res. C 73, 49–64.
Li, D., Daamen, W., Goverde, R.M., 2016. Estimation of train dwell time at short stops based on track occupation event data: A study at a dutch railway station. J. Adv. Transp. 50 (5), 877–896.
Li, Z., Huang, P., Wen, C., Jiang, X., Rodrigues, F., 2022. Prediction of train arrival delays considering route conflicts at multi-line stations. Transp. Res. C 138, 103606.
Li, Z., Huang, P., Wen, C., Tang, Y., Jiang, X., 2020a. Predictive models for influence of primary delays using high-speed train operation records. J. Forecast. 39 (8), 1198–1212.
Li, Z., Wen, C., Hu, R., Xu, C., Huang, P., Jiang, X., 2021. Near-term train delay prediction in the Dutch railways network. Int. J. Rail Transp. 9 (6), 520–539.
Li, Y., Xu, X., Li, J., Shi, R., 2020b. 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, IEEE, pp. 1–5.
Liu, Z., Ma, Q., Tang, H., Li, J., Wang, P., He, Q., 2022a. Forecasting estimated times of arrival of US freight trains. Transp. Plan. Technol. 1–22.
Liu, Q., Wang, S., Li, Z., Li, L., Zhang, J., Wen, C., 2022b. Prediction of high-speed train delay propagation based on causal text information. Railw. Eng. Sci. 1–18.
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. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 371–380.
Lundberg, S.M., Erion, G.G., Lee, S.I., 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.
Luo, J., Huang, P., Peng, Q., 2022a. A multi-output deep learning model based on Bayesian optimization for sequential train delays prediction. Int. J. Rail Transp. 1–27.
Luo, J., Peng, Q., Wen, C., Wen, W., Huang, P., 2022b. Data-driven decision support for rail traffic control: A predictive approach. Expert Syst. Appl. 207, 118050.
Marković, N., Milinković, S., Tikhonov, K.S., Schonfeld, P., 2015. Analyzing passenger train arrival delays with support vector regression. Transp. Res. C 56, 251–262.
Medeossi, G., Longo, G., de Fabris, S., 2011. A method for using stochastic blocking times to improve timetable planning. J. Rail Transp. Plan. Manag. 1 (1), 1–13.
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. Transp. Res. Rec. 03611981221074371.
Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N., 2013. A fuzzy Petri net model to estimate train delays. Simul. Model. Pract. Theory 33, 144–157.
Mou, W., Cheng, Z., Wen, C., 2019. Predictive model of train delays in a railway system. In: RailNorrköping 2019. 8th International Conference on Railway Operations Modelling and Analysis, No. 069. ICROMA, Norrköping, Sweden, June 17th–20th, 2019, Linköping University Electronic Press, pp. 913–929.
Nabian, M.A., Alemazkoor, N., Meidani, H., 2019. Predicting near-term train schedule performance and delay using bi-level random forests. Transp. Res. Rec. 2673 (5), 564–573.
Nair, R., Hoang, T.L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., Walter, T., 2019. An ensemble prediction model for train delays. Transp. Res. C 104, 196–209.
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. Appl. Sci. 10 (2), 476.
Oneto, L., Buselli, I., Lulli, A., Canepa, R., Petralli, S., Anguita, D., 2020. A dynamic, interpretable, and robust hybrid data analytics system for train movements in large-scale railway networks. Int. J. Data Sci. Anal. 9 (1), 95–111.
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, IEEE, pp. 458–467.
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 Trans. Syst. Man Cybern. A 47 (10), 2754–2767.
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 Res. 11, 54–64.
Osarogiagbon, A.U., Khan, F., Venkatesan, R., Gillard, P., 2021. Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. Process Saf. Environ. Prot. 147, 367–384.
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, Vol. 2. CIMCA-IAWTIC'06, IEEE, pp. 92–97.
Pongnumkul, S., Pechprasarn, T., Kunaseth, N., Chaipah, K., 2014. 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, IEEE, pp. 307–312.
Pradhan, R., Kumar, A., Kumar, M., Sharma, B., 2021. Simulating and analysing delay in Indian railways. In: IOP Conference Series: Materials Science and Engineering, Vol. 1116, No. 1. IOP Publishing, 012127.
Rhys, H., 2020. Machine Learning with R, the Tidyverse, and Mlr. Simon and Schuster.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.
Rößler, D., Reisch, J., Hauck, F., Kliewer, N., 2021. Discerning primary and secondary delays in railway networks using explainable AI. Transp. Res. Procedia 52, 171–178.
Schmidt, M., Weik, N., Zieger, S., Schmeink, A., Nießen, N., 2019. A generalized stochastic Petri net model for performance analysis of trackside infrastructure in Railway Station Areas under uncertainty. In: 2019 IEEE Intelligent Transportation Systems Conference. ITSC, IEEE, pp. 3732–3737.
Shi, R., Xu, X., 2020. A train arrival delay prediction model using XGBoost and Bayesian optimization. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. ITSC, IEEE, pp. 1–6.
Shi, R., Xu, X., Li, J., Li, Y., 2021. Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization. Appl. Soft Comput. 109, 107538.
Spanninger, T., Trivella, A., Büchel, B., Corman, F., 2022. A review of train delay prediction approaches. J. Rail Transp. Plan. Manag. 22, 100312.
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. J. Intell. Transp. Syst. 1–28.
Tang, R., De Donato, L., Besinović, N., Flammini, F., Goverde, R.M., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z., 2022. A literature review of artificial intelligence applications in railway systems. Transp. Res. C 140, 103679.

Thomaidis, N.S., Dounias, G.D., 2012. A comparison of statistical tests for the adequacy of a neural network regression model. Quant. Finance 12 (3), 437–449.
Tiong, K., Ma, Z., Palmqvist, C.-W., 2022. Real-time train arrival time prediction at multiple stations and arbitrary times. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). pp. 793–798. http://dx.doi.org/10.1109/ITSC55140.2022.9922299.
Vlahogianni, E.I., Golias, J.C., Karlaftis, M.G., 2004. Short-term traffic forecasting: Overview of objectives and methods. Transp. Rev. 24 (5), 533–557.
Vlahogianni, E.I., Karlaftis, M.G., 2013. Testing and comparing neural network and statistical approaches for predicting transportation time series. Transp. Res. Rec. 2399 (1), 9–22.
Wang, S., Cao, J., Yu, P., 2020. Deep learning for spatio-temporal data mining: A survey. IEEE Trans. Knowl. Data Eng.
Wang, P., Zhang, Q.p., 2019. Train delay analysis and prediction based on big data fusion. Transp. Saf. Environ. 1 (1), 79–88.
Washington, S., Karlaftis, M., Mannering, F., Anastasopoulos, P., 2020. Statistical and Econometric Methods for Transportation Data Analysis. Chapman and Hall/CRC.
Watanabe, S., Mori, Y., Takatori, Y., Yonemoto, K., Tomii, N., 2018. Train traffic simulation algorithm based on historical train traffic records. Comput. Railw. XVI 285–293.
Wen, C., Huang, P., Li, Z., Lessan, J., Fu, L., Jiang, C., Xu, X., 2019. Train dispatching management with data-driven approaches: a comprehensive review and appraisal. IEEE Access 7, 114547–114571.
Wen, C., Lessan, J., Fu, L., Huang, P., Jiang, C., 2017. Data-driven models for predicting delay recovery in high-speed rail. In: 2017 4th International Conference on Transportation Information and Safety. ICTIS, IEEE, pp. 144–151.
Wen, C., Mou, W., Huang, P., Li, Z., 2020. A predictive model of train delays on a railway line. J. Forecast. 39 (3), 470–488.
Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. pp. 1–10.
Wu, J., Du, B., Wu, Q., Shen, J., Zhou, L., Cai, C., Zhai, Y., Wei, W., Zhou, Q., 2021a. A hybrid LSTM-CPS approach for long-term prediction of train delays in multivariate time series. Future Transp. 1 (3), 765–776.
Wu, J., Wang, Y., Du, B., Wu, Q., Zhai, Y., Shen, J., Zhou, L., Cai, C., Wei, W., Zhou, Q., 2021b. The bounds of improvements toward real-time forecast of multi-scenario train delays. IEEE Trans. Intell. Transp. Syst. 23 (3), 2445–2456.
Yaghini, M., Khoshraftar, M.M., Seyedabadi, M., 2013. Railway passenger train delay prediction via neural network model. J. Adv. Transp. 47 (3), 355–368.
Zhang, Y., Liao, L., Yu, Q., Ma, W., Li, K., 2021. Using the gradient boosting decision tree (GBDT) algorithm for a train delay prediction model considering the delay propagation feature. Adv. Prod. Eng. Manag. 16 (3), 285–296.
Zhuang, H., Feng, L., Wen, C., Peng, Q., Tang, Q., 2016. High-speed railway train timetable conflict prediction based on fuzzy temporal knowledge reasoning. Engineering 2 (3), 366–373.