Journal of Rail Transport Planning & Management 30 (2024) 100448

Contents lists available at ScienceDirect

Journal of Rail Transport Planning & Management

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

Forecasting train arrival delays on the Ankara – Eskişehir high-speed line in Turkey

Özgül Ardıç

Turkish State Railways (TCDD), Ankara, Turkey

ARTICLE INFO

Keywords:
Train arrival delays
Train delay predictions
Machine learning
Random forest

ABSTRACT

Railway operations may experience delays due to technical issues or weather conditions. Accurate prediction of such delays can enhance the quality of rail transport services and the effectiveness of railway operations. The study has developed the arrival delay prediction model using random forest regression based on the train operation data from the Ankara - Eskişehir high-speed train line in Turkey. The model can simultaneously predict arrival delays at all downstream stations on this line and continuously update these predictions as new information about train movements becomes available. The accuracy rates of the model vary from 76% to 99% under a 1-min prediction error. The results show that incorporating variables related to weather conditions and technical problems related to train control systems into the model improves prediction performance. The contribution of these variables to the model performance increases as the prediction horizon widens. The model results suggest that the model predictions may assist network managers in making better decisions about train operations. In order to evaluate the model's performance from the passengers' point of view, the study has proposed two methods: the proportion of late predictions and the stability of forecasts. The findings indicate that most trains (between 96.7% and 99%) have stable arrival delay predictions at target stations. The proportion of 2-min (or greater) late predictions, which means that the predicted delay exceeds the actual delay by 2 min or more, fluctuates from 14% to 0.5%, depending on the prediction horizon. Although the ratio for the short horizons (one station ahead) becomes relatively low, it is necessary to be cautious when using the model predictions to inform passengers because a prediction of more than 1 min late for short horizons might have negative consequences (e.g., misleading passengers to leave stations).

1. Introduction

Train delays can occur due to various technical issues, such as infrastructure problems or locomotive failures, as well as external factors. Such delays can be inconvenient for railway passengers and network managers responsible for ensuring smooth operations. Therefore, it is essential to make accurate predictions about train delays to enhance the efficiency of railway operations and improve the quality of rail transport services.

In the field of railway research, various modeling approaches are used to predict train delays. According to studies by Spanninger et al. (2022) and Tiong et al. (2023a), these approaches can be classified as "event-driven" and "data-driven". Event-driven approaches focus on the dependencies between train events (such as departures, arrivals, and pass-throughs) and the dynamics of railway

E-mail address: ozgulardic@hotmail.com.

https://doi.org/10.1016/j.jrtpm.2024.100448
Received 15 August 2023; Received in revised form 1 April 2024; Accepted 8 April 2024
Available online 20 April 2024

operations. On the other hand, data-driven approaches rely on historical train movement data and actual delays to predict delays at specific locations without considering the dependencies of train events. They typically use supervised machine learning (ML) methods. As digital technologies become increasingly prevalent in the rail sector, more historical data related to railway operations has become available, making supervised ML methods a popular tool for analyzing this data to predict train delays (see Spanninger et al. (2022) and Tiong et al. (2023a) for the extensive review of these studies).

However, most studies that use data-driven techniques with ML methods typically focus on predicting delays only one station ahead (e.g., Hauck and Kliewer, 2019; Lapamonpinyo et al., 2022; Laifa and Ghezalaa, 2021). This approach has limited practical use in train operations because passengers and network managers require more detailed information about train delays. Passengers need constantly updated information about the train's arrival time at their station of interest as the train approaches. Network managers additionally need to know the arrival times of trains at several downstream stations and receive updates each time the train arrives at the next station to make better decisions in train operations (Spanninger et al., 2022; Tiong et al., 2023a). Some studies achieve such dynamic multiple-station predictions. Nevertheless, to accomplish this, they combine ML models with simulation methods (see, for instance, Nair et al. (2019) and Peters et al. (2005)). Studies by Huang et al. (2020, 2021) and Taleongpong et al. (2022) introduce state-of-the-art ML methods to predict train delays at multiple stations. Still, their research does not forecast train delays at downstream stations along the route simultaneously when trains move along the way. Oneto et al. (2016, 2017, 2018) and Lulli et al. (2018) introduce hybrid ML methods to predict train delays at several downstream stations on trains' routes simultaneously and dynamically update delay predictions whenever trains depart from the next station. However, they build a different model for each train. Using their methods on a rail network where hundreds of trains run daily is not feasible because training these models and integrating them into IT systems to predict delays during daily train operations are too complex and time-consuming. Similarly, to achieve this same objective, Tiong et al. (2022) adopts a relatively more efficient approach and develops eight models for a train line with eight stations. Still, this approach might be impractical when predicting delays along several train lines with hundreds of stations. Unlike these, the main goal of this study is to develop a single ML model, which can be utilized for all trains and stations, to simultaneously predict arrival delays at all downstream stations on the Ankara-Eskişehir high-speed train line in Turkey and dynamically update delay predictions as trains move along the way. In addition, the study aims to explore three relatively overlooked topics in the literature.

Firstly, it has been shown that various weather conditions, such as wind, snow, fluctuations in temperature, and precipitation, have a significant impact on train arrival punctuality (Xia et al., 2013; Økland and Olsson, 2021). However, only a few studies using data-driven ML methods consider these aspects when developing models for predicting train delays. Some of the studies, such as those conducted by Nair et al. (2019), Huang et al. (2021), Lulli et al. (2018), Li et al. (2022), and Ji et al. (2020), incorporate weather-related variables into their prediction models. However, they do not report how these variables affect the accuracy of the prediction and how this effect changes over time. Huang et al. (2020), Hauck and Kliewer (2019), and Lapamonpinyo et al. (2022) find that including these variables in train delay prediction models improves their accuracy. Nevertheless, they do not analyze how the effect of weather-related variables on accuracy changes over different prediction horizons. Finally, Oneto et al. (2016, 2017) develop prediction models using weather-related variables and show that the effect of weather on model performance slightly increases as the prediction horizon broadens. However, they use only the data from the Italian rail network for six months, which does not show seasonal variation. Moreover, both studies have focused on only one train line in Italy. The effect of similar weather conditions might lead to different delay patterns depending on the type and geographic position of rail infrastructure. As a result, the impact of weather-related variables on the accuracy of prediction models across various prediction horizons needs to be explored in greater detail.

Secondly, studies show that train control system problems, such as issues with the pantograph, catenary, and signal systems, are responsible for about 20% of the causes of train delays (Huang et al., 2020; Økland and Olsson, 2021). Therefore, incorporating information about delays caused by these issues into prediction models may increase the accuracy of arrival delay predictions. However, none of the studies mentioned above have directly integrated this feature into their models and analyzed its impact on prediction accuracy.

Finally, studies that aim to inform passengers about train delays have some shortcomings in evaluations of model predictions (see, for instance, Tiong et al. (2022), Nair et al. (2019), and Laifa and Ghezalaa (2021)). They mainly use metrics like Mean Squared Error (MAE) and Root Mean Square Error (RMSE) to assess the accuracy of predictions. However, some additional assessment methods are required for the following reasons: (1) the same absolute prediction error can affect their behavior and emotions differently based on the direction of the prediction error (whether it is earlier or later than the actual delay). Early predictions can cause dissatisfaction and discomfort among passengers, while late predictions can mislead passengers into leaving the station, assuming the train will arrive much later. Therefore, it is essential to evaluate the rate of late predictions. (2) It is also crucial to measure the stability of predictions at a station during continuous updates whenever a train departs from the next station on its route, as the fluctuations may lead passengers to perceive the information as unreliable.

This work aims to address the research gaps listed above and make the following contributions to the scientific literature: Firstly, the study is the first to develop a single ML model to simultaneously predict arrival delays at all downstream stations and dynamically update these predictions as new information about train movements becomes available. Such model will benefit both passengers and network managers as it predicts arrival delays at all downstream stations simultaneously and continuously updates these predictions. Furthermore, using a single model will allow us to use the model in practice as a model training and evaluation process of one model, and its integration into other rail IT systems (e.g., passenger information and interfaces for dispatchers) is less complex and less time-consuming. To illustrate this further, let's consider the training and evaluation process of ML models. For large datasets, such as train movement data encompassing millions of records, it might take a few hours or more to train an ML model. To ensure that the model is not overfitted or underfitted and achieves high prediction performance, the training process must be repeated several times with

Fig. 1. The Ankara - Eskişehir high-speed Line.

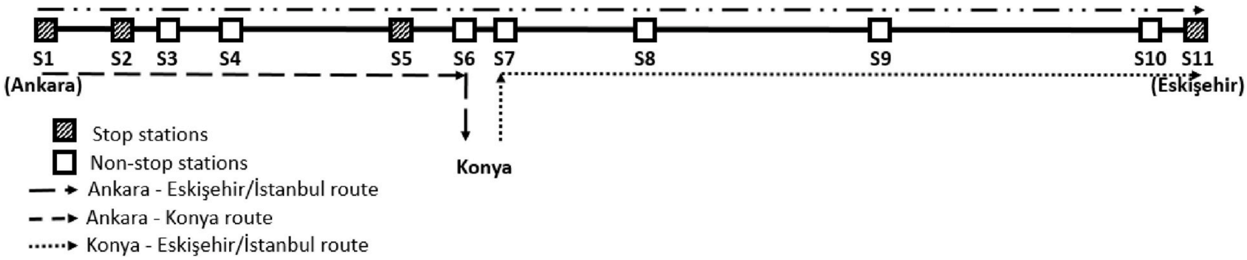

Fig. 2. Stations and train routes on the Ankara - Eskişehir high-speed Line.

different sets of parameters. Additionally, once implemented in a real-life setting, the model needs to be retrained regularly (e.g., once a day or week) with the latest data. If different models are required for each train and station, then this would necessitate at least hundreds, if not thousands, of models. Building and maintaining such a vast number of models over time can be a daunting task and almost impossible to achieve in a real-life scenario. Secondly, it is the first study that integrates problems regarding train control systems (e.g., signaling systems) in the model and analyzes its impact on delay prediction accuracy. Thirdly, the study sheds additional light on the effects of weather-related variables on the accuracy of the prediction models across different prediction horizons as the data inherits sessional variability (more than two years) and belongs to a Turkish high-speed line, which naturally has a different geographic location and rail infrastructure (e.g., the number of culverts, stations) than the Italian train line analyzed by the studies of Oneto et al. (2016, 2017). Finally, this study proposes two additional methods to evaluate delay prediction models based on the proportion of late predictions and the stability of forecasts (in addition to other accuracy metrics), which is of great importance for the models that aim to inform passengers.

The rest of the paper is divided into four sections. Section 2 discusses methodological issues, like problem formulation, data description, types of ML methods used, and model structures. Section 3 presents the results of the models and their prediction accuracies. Lastly, Section 4 summarizes the main findings and provides insights for future research directions.

2. Methodology

2.1. Problem formulation and data description

In a rail network, a train line contains a series of stations $\{i = 1, 2, 3, \ldots, n\}$, and a timetable is designed for each train, specifying the planned arrival and departure times at each station along the train line. Along this train line, We can label the originating station as S_1, the consecutive stations as S_2, $S_3 \cdots S_n$, and S_n being the last one, and define an arrival delay of a specific train at an individual station as a difference between the planned and actual arrival times. Accordingly, considering that a train of interest departs from $S_x \in \{S_1, S_2, \ldots, S_n\}$, the goal of this study is to simultaneously predict the expected arrival time of a train of interest at all downstream stations (e.g., S_{x+1}, S_{x+2}, …, S_n) and update the predicted arrival delays at each of these downstream stations whenever the train departs from the consecutive station on the line.

To achieve this goal, this research uses train movement data from the Ankara - Eskişehir high-speed train line in Turkey. The data was collected over a period of time, starting from July 10, 2021, and ending on October 31, 2023. This line is a double-track line and is dedicated solely to high-speed trains. The total distance of the line is approximately 245 km, and the maximum speed limit allowed on the line is 250 km/h. Fig. 1 shows the line's geographic position, while Fig. 2 displays the relative positions of the 11 stations on this line. These stations range from S1 (Ankara) to S11 (Eskişehir), and among them, S1, S2, S5, and S11 are stop stations where passengers alight and board the train. The remaining stations are non-stop, but the trains' scheduled departure and arrival times at these stations are included in their itineraries. This study only considers trains running in the direction from Ankara to Eskişehir. There are three

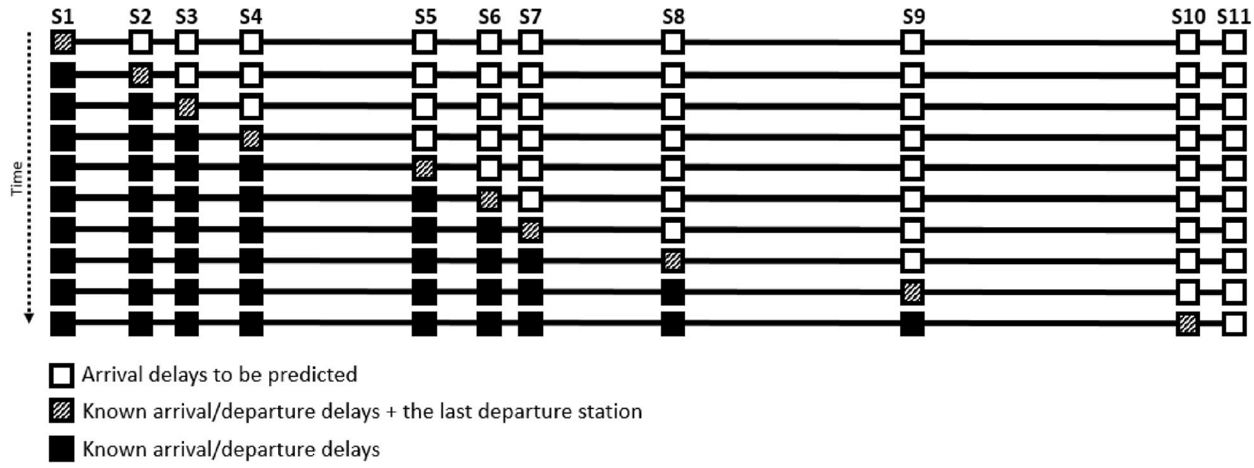

Fig. 3. The Arrival delay prediction framework on the Ankara - Eskişehir high-speed Line.

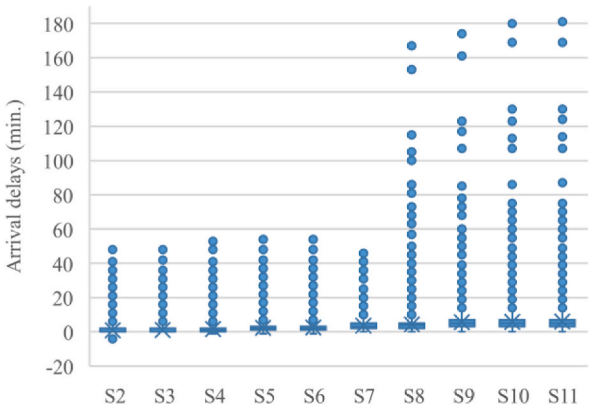

Fig. 4. The Distribution of arrival delays across stations.

Table 1
The Descriptive statistics of arrival delays across stations.

Stations	Mean	Std. Dev.
S2	0.95	2.23
S3	1.06	2.34
S4	1.57	2.41
S5	2.40	2.62
S6	2.47	2.74
S7	3.65	2.94
S8	4.40	5.00
S9	5.56	5.40
S10	5.85	5.78
S11	5.93	5.84

different routes in this direction (see Fig. 2).

(1) Trains running from Ankara to Eskişehir (and then İstanbul) pass over S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, and S_{10}, and arrive at Eskişehir
(2) Trains running from Ankara to Konya pass over S_2, S_3, S_4, S_5 and S_6 and then turn to the direction of Konya
(3) Trains running from Konya to Eskişehir (and then İstanbul) pass over S7, S8, S9, and S10 and arrive at Eskişehir.

The total number of trains that ran during the study period was 11613, 5208, and 3953 for each route, respectively. On average, 14, 6, and 5 trains run per day on each route (no variation between weekday and weekend), respectively.

Fig. 3 depicts the framework for predicting arrival delays on the Ankara-Eskişehir High-speed Line. The figure indicates that when a

Table 2
Description of train operation-related variables.

Variables	Type	Unit	Description	The dataset used for the calculation
Departure delay	continuous	minutes	The difference between planned and actual departure time at the departure station	(1) (2)
Recovery time	continuous	minutes	The extra time allowed for a train between the departure and the target station to compensate for losing time at earlier parts of its journey	(1)
Temporary speed limit	continuous	minutes	The total amount of planned delay applied on the way between the departure and the target station because of maintenance activities	(3)
The delay of the previous train	continuous	minutes	The delay of the previous train, which arrived at the target station earlier than the train concerned, between the departure and the target station	(1) (2)
The number of passengers	continuous	person	The number of passengers get on the train between the departure and the target station	(4)
Distance	continuous	meter	The length of way between the departure and the target station	(1)
Technical problems related to train control systems	dummy	0 or 1	whether there exist any problems regarding train control systems (e.g., signaling system, catenary) between the departure and the target station	(2)

Table 3
An example of records of a train in the dataset.[a]

Departure station	Target Station	Departure Delay	Arrival Delay	Temporary Speed limit	Recovery time	The delay (previous train)	Technical problems related to train control systems	The Number of Passengers	Distance (km)
S1	S2	0	0	0	−3	0	0	385	21
S1	S3	0	1	0	−5	1	0	465	23
S1	S4	0	2	1	−7	3	0	465	41
S1	S5	0	3	1	−8	4	0	465	83
S1	S6	0	4	1	−10	4	0	495	98
S1	S7	0	4	1	−10	4	0	495	100
S1	S8	0	4	1	−10	4	0	495	131
S1	S9	0	7	5	−11	7	0	495	184
S1	S10	0	7	6	−11	7	0	495	240
S1	S11	0	7	6	−11	7	0	495	245
S2	S3	1	1	0	−1	1	0	465	4
S2	S4	1	2	1	−3	3	0	465	20
S2	S5	1	3	1	−4	4	0	465	65
|	|	|	|	|	|	|	|	|	
|	|	|	|	|	|	|	|	|	
|	|	|	|	|	|	|	|	|	
S7	S8	4	4	0	0	0	0	495	30
S7	S9	4	7	4	−1	3	0	495	82
S7	S10	4	7	5	−1	3	0	495	136
S7	S11	4	7	5	−1	3	0	495	140
S8	S9	4	7	4	−1	3	0	495	51
S8	S10	4	7	5	−1	3	0	495	105
S8	S11	4	7	5	−1	3	0	495	110
S9	S10	7	7	1	0	3	0	495	58
S9	S11	7	7	1	0	3	0	495	60
S10	S11	7	7	1	0	0	0	495	4

[a] The original records have been slightly changed due to privacy reasons, and the records provided are only illustrations of the original ones. Some records in the middle of the table are omitted to save some space.

train departs from station S1, the expected arrival delays at stations S2, S3, S4, S5, S6, S7, S8, S9, S10 and S11 will be predicted simultaneously. Similarly, when the train departs from station S2, the expected arrival delays at stations S3, S4, S5, S6, S7, S8, S9, S10 and S11 will be updated. This process will continue until the train reaches station S10, where the expected arrival delay at station S11 will be updated.

Fig. 4 displays the arrival delay distribution at each station on the Ankara-Eskişehir High-speed Line. The figure indicates that the arrival delays at the stations follow a long-tailed distribution, with an increase in variability after S7. This increase in variability can be attributed to the fact that the risk of delay increases with distance. Additionally, it is likely that the probability of delays at downstream stations (S8, S9, S10, and S11) increases due to possible conflicting train operations at S6 and S7 (as seen in Fig. 2). Table 1 also demonstrates that the mean and standard deviation of arrival delays at a station increase as the distance from S1 increases. Train operation-related and weather-related variables are used as model predictors. Sections 2.1.1 and 2.1.2 explain train operation-related and weather-related variables, respectively.

Table 4
The Description of weather-related variables.[a]

Variables	Type	Unit	Description
Temperature	continuous	Celsius	The perceived feels-like temperature combining the wind chill factor
wind speed	continuous	kilometers per hour	Wind speed at 10 m above the ground
precipitation (rain + snow)	continuous	millimeters	Total precipitation (rain, showers, snow) sum of the preceding hour

[a] The description of variables is adopted from https://open-meteo.com/.

Table 5
The Descriptive statistics of weather-related variables.

Stations	Distance (km)[a]	Temperature				Precipitation				Wind Speed			
		Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
S1	14.4	11.19	11.11	−17.10	39.40	0.04	0.29	0.00	12.00	7.92	4.62	0.00	46.90
S2	4.52	11.60	11.13	−16.80	39.80	0.04	0.29	0.00	12.00	7.92	4.62	0.00	46.90
S3	2.39	11.61	11.14	−15.90	39.90	0.05	0.30	0.00	9.30	7.48	4.17	0.00	41.10
S4	7.16	11.92	11.21	−16.40	40.00	0.05	0.31	0.00	13.80	7.24	3.89	0.00	38.20
S5	9.38	11.38	10.96	−20.00	40.70	0.05	0.28	0.00	8.50	9.36	5.05	0.00	44.20
S6	3.67	11.82	11.25	−20.10	41.20	0.04	0.25	0.00	10.60	9.02	5.15	0.00	43.40
S7	6.6	12.13	11.26	−20.70	41.90	0.04	0.26	0.00	7.80	8.45	5.21	0.00	42.70
S9	4.91	12.00	10.94	−19.20	41.00	0.05	0.30	0.00	9.50	8.22	4.68	0.00	44.60
S8	5.68	12.52	11.07	−17.10	41.40	0.04	0.28	0.00	9.80	7.54	4.92	0.00	42.70
S10	10.81	10.93	10.50	−19.50	40.40	0.05	0.30	0.00	11.20	10.06	5.40	0.00	38.10
S11	4.31	11.11	10.68	−21.80	40.20	0.05	0.29	0.00	9.70	9.04	5.02	0.00	40.10

[a] The distance between stations and weather measurement points.

2.1.1. *Predictors: train operation-related variables*

Train delays and punctuality are influenced by various factors related to train operation, such as distance and temporary speed limits (Tiong et al., 2023b; Olsson et al., 2004). Studies that aim to predict train delays use different combinations of these variables according to the scope of the research and the availability of data (see Tiong et al. (2023a) for a detailed review of these studies). This research uses seven train operation-related variables listed in Table 2 as predictors. Variables related to infrastructure and rolling stock features are not included since the research focuses on a single high-speed line where these variables remain almost constant. The selected variables are obtained from four datasets provided by the Turkish State Railways (TCDD).

(1) Train timetables which contain planned departure, arrival, and recovery times of trains
(2) Actual train arrival and departure times of trains along the stations on their timetables, and delay reason codes if trains could not comply with the scheduled arrival and departure times
(3) Planned delays at specific segments of train lines because of planned maintenance activities
(4) Train ticket sales

Multiple records are created for each train comprising variables presented in Table 2 based on the four datasets listed above. The number of records for each train is equal to the number of departure and target station pairs in its route. Table 3 presents records for a train that travels from S_1 to S_{11} to illustrate the data structure. For each pair of departure and target station, the departure delay reflects the departure delay of the train at the relevant departure station, and the arrival delay reflects the arrival delay of the train at the relevant target station. Other variables, such as temporary speed limit, recovery time, delay (previous train), technical problems related to train control systems, number of passengers, and distance, are calculated between this departure and target station. For instance, in the first record of the example train in Table 3, the departure delay is the departure delay of the train at S1, and the arrival delay is the arrival delay of the train at S2. The values for the remaining variables are calculated between S1 and S2. After the record generation process is complete, the total number of records in the dataset amounts to 756,302.[1]

2.1.2. *Predictors: weather-related variables*

Each record in the dataset (see Table 3) is matched with the weather variables of the departure station based on the departure hour of the train. The dataset of hourly weather conditions during the research period is downloaded from the website https://open-meteo.com/. Table 5 presents the distance between stations and weather measurement points. The dataset includes three weather variables: temperature, wind speed, and precipitation. Table 4 and Table 5 provide descriptions and descriptive statistics of these variables. Table 5 shows that there is a very high variation in weather-related variables. Such variation is not surprising since the geographic position of this train line experiences frigid temperatures with frequent show falls in winters and very hot summers.

[1] Records of 10 trains that have abnormal data related to the number of passengers and departure/arrival times are eliminated.

Table 6
Predictor variables of NoW-NoT, NoW-T, W-NoT and W-T models.

	Predictor Variables	Models (RF)			
		NoW-NoT	NoW-T	W-NoT	W-T
Train operation-related variables	Departure delay	X	X	X	X
	Recovery time	X	X	X	X
	Temporary speed limit	X	X	X	X
	The delay (previous train)	X	X	X	X
	The number of passengers	X	X	X	X
	Distance	X	X	X	X
	Technical problems related to train control systems		X		X
Weather-related variables	Temperature			X	X
	Wind Speed			X	X
	Precipitation			X	X

2.2. Modelling

Four distinct models, each with different sets of predictors, are developed to explore the effect of weather conditions and technical problems related to train control systems on the accuracy of arrival delay predictions. All four models can predict arrival delays at all downstream stations simultaneously after the train departure from any station on this line. They are listed below. Table 6 presents the predictor variables of these models.

(1) The model without weather-related variables and the variable about technical problems related to train control systems (NoW-NoT)
(2) The model without weather-related variables but with the variable about technical problems related to train control systems (NoW-T)
(3) The model with weather-related variables but without the variable about technical problems related to train control systems (W-NoT)
(4) The model with both weather-related variables and the variable about technical problems related to train control systems (W-T)

Table 6 displays predictor variables included in NoW-NoT, NoW-T, W-NoT, and W-T models with a sign “X”, whereas excluded variables have no sign.

The implementation of the models in a real-world setting is intended to proceed as follows: During train operations, the values of the predictor variables are fed into the prediction model which calculates trains’ arrival delays. The process requires the integration of the prediction model with two existing applications: (1) The Passenger Information Application (PIA), which displays train departure and arrival times on information screens at stations to inform passengers (2) The Train Tracking Application (TTA), which displays real-time train locations, departure, and arrival times on screens to assist network managers in making decisions about train operations. The integration of the prediction model with these two applications is expected to function in the following manner: Whenever a train departs from a station, the PIA and TTA trigger the prediction model by sending values of all predictor variables related to each target station for which the train’s arrival delay is to be calculated. The proposed modeling framework for predicting arrival delays on the Ankara-Eskişehir High-speed Line is depicted in Fig. 3, in Section 2.1. Accordingly, for example, when a train leaves from station S1, 10 sets of predictor values are sent to the prediction model, one set for each of the 10 target stations (S2, S3, …, S11) as shown in Table 3. The prediction model is triggered 10 times, once for each target station. The prediction model then calculates the expected arrival delays for each target station and sends the results back to the PIA and TTA. The values of the following predictors, such as recovery time, temporary speed limit, distance, and the number of passengers, are known a day prior to the train’s departure. These values are then inserted into the prediction dataset for each departure and target pair, as shown in Table 3, during the night. The values of predictors related to the delay (previous train), technical problems related to train control systems, and departure delay are updated in the dataset shortly before the PIA and TTA trigger the prediction model as soon as the concerned data is available. In some cases, technical problems related to train control systems may occur after the train has left the station, which can cause changes in the technical problem variable. In such situations, the PIA and TTA rerun the prediction model with the updated variable value. The PIA and TTA obtain values of weather-related variables around the departure station at the time of departure from the website https://open-meteo.com/.

2.3. Model training and evaluation: random forest regression (RF)

Before the model training process starts, the dataset is randomly divided into two samples: 80% for training and 20% for testing. The models are trained using the training sample and then tested with the testing sample to ensure there are no underfitting or overfitting issues. MAE and RMSE values are then used to evaluate and compare the predictive accuracy of the models. If yi is the actual value, yp is the predicted value, and n is the number of observations, the formulas of MAE and RMSE are as follows:

Table 7
The Optimal model parameters.

	Trees_num	Try_num	Max_depth
NoW-NoT	400	6	30
NoW-T	400	7	30
W-NoT	400	7	30
W-T	500	8	30

$$\mathrm{MAE} = \frac{\sum \left| y_i - y_p \right|}{n}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum (Y_i - Y_P)^2}{n}}$$

The random forest regression (RF) method is used to build the prediction models. It is the most widely used ML method to forecast train delays (Spanninger et al., 2022; Tiong et al., 2023a). RF uses a group of decision trees to map the relationship between vectors of the predictor and dependent variables. The model's output is the mean prediction of the individual trees (Kecman et al., 2015). Three parameters of RF are optimized: (1) trees_num, the number of trees to grow; (2) max_depth, the maximum depth of a tree; and (3) try_num, the number of randomly selected variables for splitting. During the training process, the following options of three parameters are considered: Trees_num values range from 100 to 500, with increments of 100. Max_depth values are set to 5, 10, 15, 20, 25 and 30. The values of try_num depend on the number of predictors used in the model. Try_num values are 2, 4, and 6 for the NoW-NoT model; 1, 3, 5, and 7 for the NoW-T model; 1, 3, 5, 7, and 9 for the W-NoT model; and finally 2, 4, 6, 8, and 10 for the W-T model.

These four models are trained with all possible combinations of these parameter options, which takes around 6 h. To evaluate the performance of RF models and find the optimal parameters, an "out-of-bag" (OOB) error is employed. Random Forest (RF) uses bagging and random feature selection, where each tree in RF is developed on a sub-sample drawn with a replacement from the original training set. The samples not used for training a specific tree are used for computing the OOB error. The average OOB errors of all the trees in RF indicate the model's accuracy (Kecman et al., 2015). The OOB error, namely mean square error, is formulated as below.

$$\mathrm{MSE} = \frac{\sum (Y_i - Y_P)^2}{n}$$

Table 7 presents the parameters that provide the best performance for each model in terms of OOB errors. Appendixes 3, 4, 5, and 6 present the models' OBB errors. It is worth noting that the models are trained with scaled predictor variables, too, but the models trained with the original dataset are preferred because their performance is better than those of the models with scaled variables.

The study uses three benchmark models to compare the performance of the RF models. The first model is multiple linear regression (MLR), commonly used to forecast train delays (Spanninger et al., 2022; Tiong et al., 2023a). MLR models the output variable as a linear combination of the input features (Palmer et al., 2009). The alternating direction method of multipliers (ADMM) is applied to solve elastic net regularized multiple linear regression. Alpha and lambda parameters are adjusted. Alpha values range from 0 to 1, where 0 indicates ridge regularization and 1 indicates lasso regularization. The lambda parameter adjusts the coefficient shrinkage amount and takes values greater than 0. The training process includes Lambda values of 0, 0.1, 0.001, 0.01, 0.02, 0.3, 0.5, 0.7, 0.9, 1, 2, 3, 5, 7, and 9, and Alpha values of 0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.9, and 1. The grid search method with five-fold cross-validation is used to determine the optimal parameters of the MLR. Grid search evaluates every possible combination of hyper-parameter values. Cross-validation divides the dataset randomly into almost equally sized folds (two or more, depending on the sample size). One fold is reserved as a validation fold, and the rest is used to train the model. The training process is repeated until each fold becomes a validation fold once. The mean error of all steps indicates the model's performance (Neunhoeffer and Sternberg, 2019). The training process indicates that the optimal values of Lamda and Alpha for the four models above are 0. It is worth noting that the training process is repeated after scaling the predictor variables. However, the models trained with the original dataset are preferred because scaling does not improve model performance.

In addition to the MLR model, the study includes two baseline models for comparison: (1) the departure delay-based prediction (DDBP) and (2) the mean prediction (MP). These baseline models have quite straightforward algorithms and do not use any machine learning methods. DDBP uses departure delays of trains at the previous station (the last known delays of trains) as the arrival delays at the target station, as applied by Hauck and Kliewer (2019) and Pongnumkul et al. (2014). It should be noted that this model does not take into account various factors in predicting arrival delays, such as recovery time or previous train delays at the target station. MP is the mean value of arrival delays at the target station during the research period (see Meng et al. (2022) for a similar application). Similar to DDBP, MP might have biased arrival delay predictions as arrival delays have a long-tailed distribution. Nevertheless, comparing predictions of these baseline models, as well as predictions of MLR, with those of RF models can help us gain additional insights into the performance of RF models, provided that we keep their limitations in mind. Furthermore, it might be useful to consider DDBP as a baseline model for comparison, as it simply emulates how passengers can intuitively predict arrival delay from the last known delay as they do not have any other information.

Table 8
The MAE and RMSE values of the RF model (NoW-NoT) for training and testing datasets (in minutes).

RF-Model	MAE		RMSE	
	Training	Testing	Training	Testing
NoW-NoT	0.73	0.84	1.43	1.66

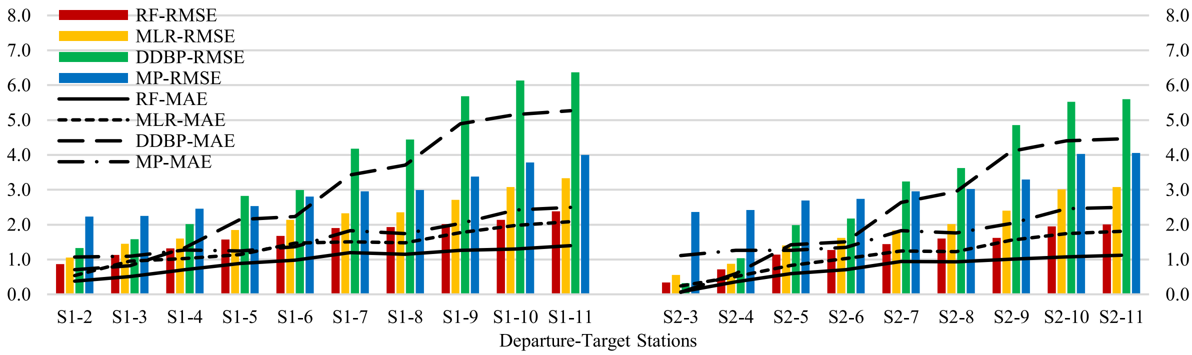

Fig. 5. The Performance of RF (NoW-NoT), MLR (NoW-NoT), and baseline models for departure stations S1 and S2 (in minutes).

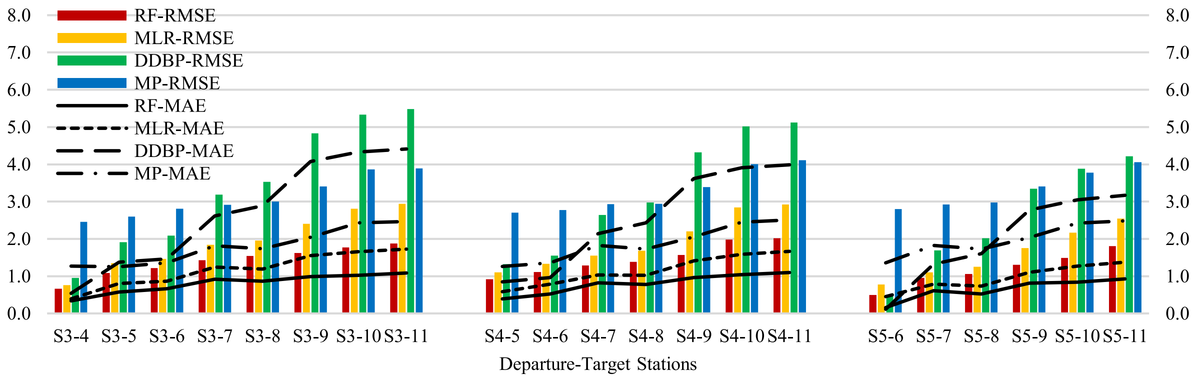

Fig. 6. The Performance of RF (NoW-NoT), MLR (NoW-NoT), and baseline models for departure stations S3, S4 and S5 (in minutes).

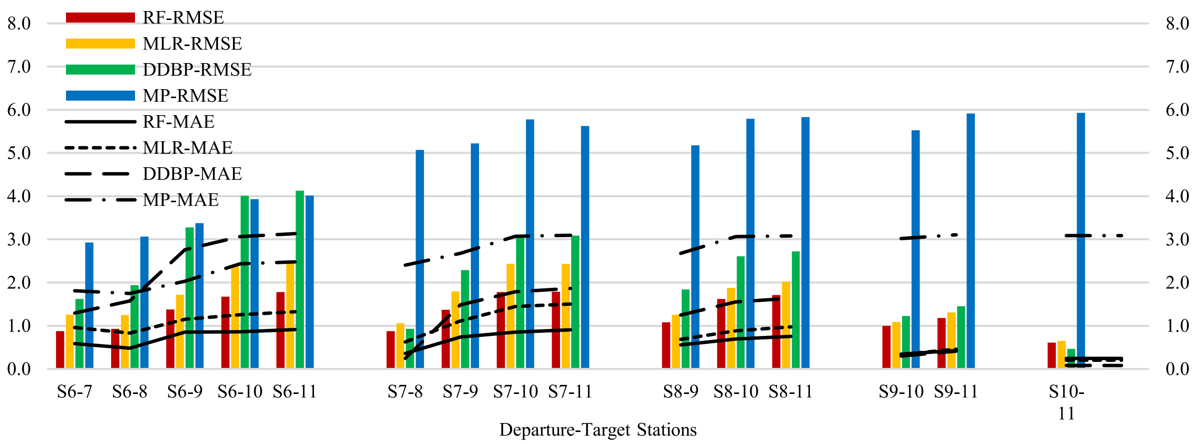

Fig. 7. The Performance of RF (NoW-NoT), MLR (NoW-NoT), and baseline models for departure stations S6, S7, S8, S9 and S10 (in minutes).

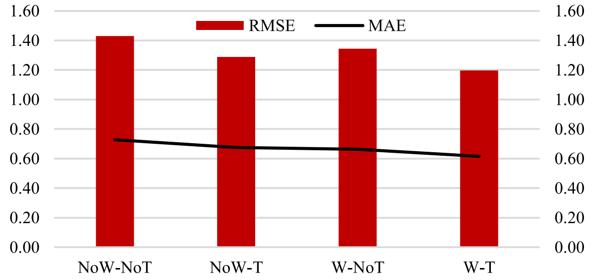

Fig. 8. The MAE and RMSE values of four RF models (in minutes).

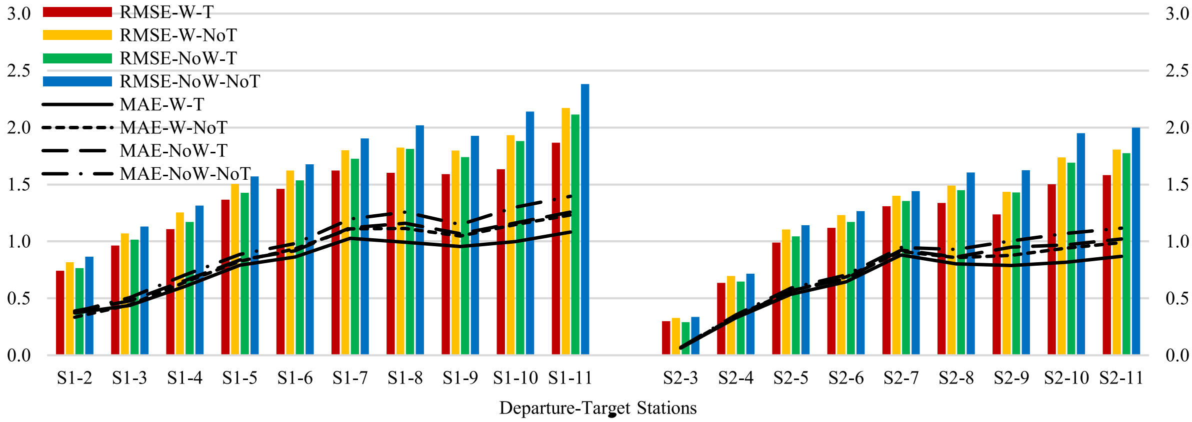

Fig. 9. The Performance of four RF models for departure stations S1 and S2 (in minutes).

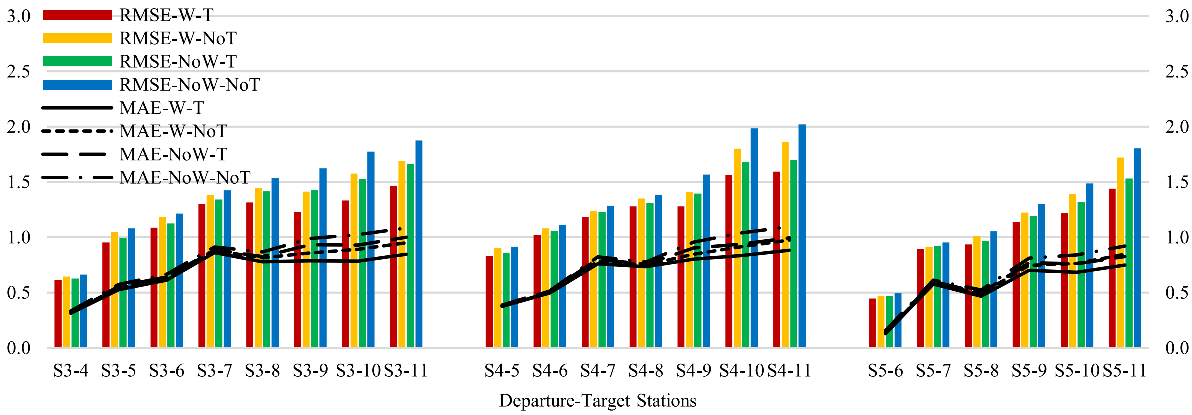

Fig. 10. The Performance of four RF models for departure stations S3, S4 and S5 (in minutes).

3. Results

Table 8 compares the MAE and RMSE values of the RF model (NoW-NoT) for the training and testing datasets. The table demonstrates that the MAE and RMSE values of the training dataset are very similar to those of the testing dataset. Moreover, there is no notable difference between the MAE and RMSE values of the training and testing datasets for the other three types of RF models and the pairs of departure and target stations (see Appendixes 1 and 2). Hence, we can conclude that the models do not suffer from overfitting or underfitting problems.

Figs. 5–7 display a comparison of the prediction performances of four models (RF model (NoW-NoT), MLR model (NoW-NoT),

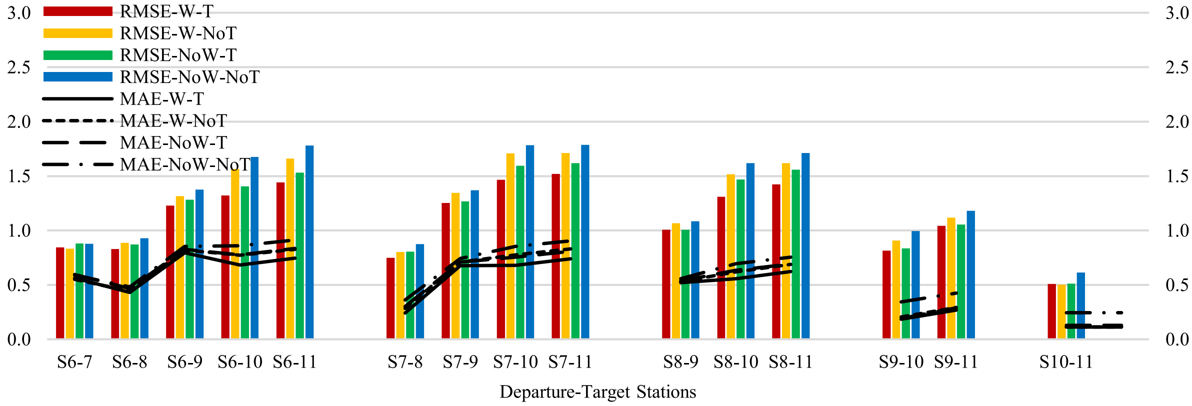

Fig. 11. The Performance of four RF models for departure stations S6, S7, S8, S9, and S10 (in minutes).

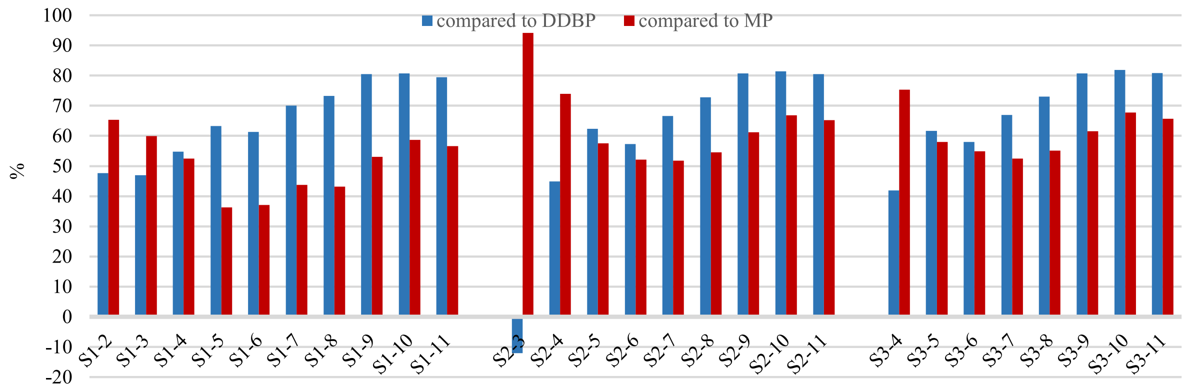

Fig. 12. The comparison of the RF (W–T) model with baseline models (DDBP and MP) for departure stations S1, S2, and S3 (%, in terms of MAE).

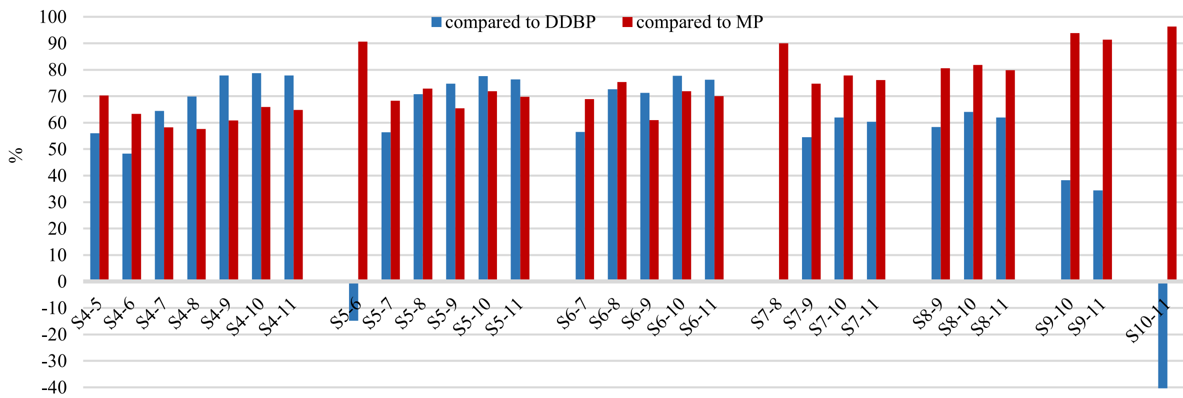

Fig. 13. The comparison of the RF (W–T) model with baseline models (DDBP and MP) for departure stations S4, S5, S6, S7, S8, S9, and S10 (%, in terms of MAE).

DDBP, and MP) across the pairs of departure and target stations. The results show that the RF model performs better than all other models in terms of MAE and RMSE. As the prediction horizon increases, indicating a greater distance between the departure and target stations, the prediction performance of all models declines. Nevertheless, the RF model outperforms the other models even for wide

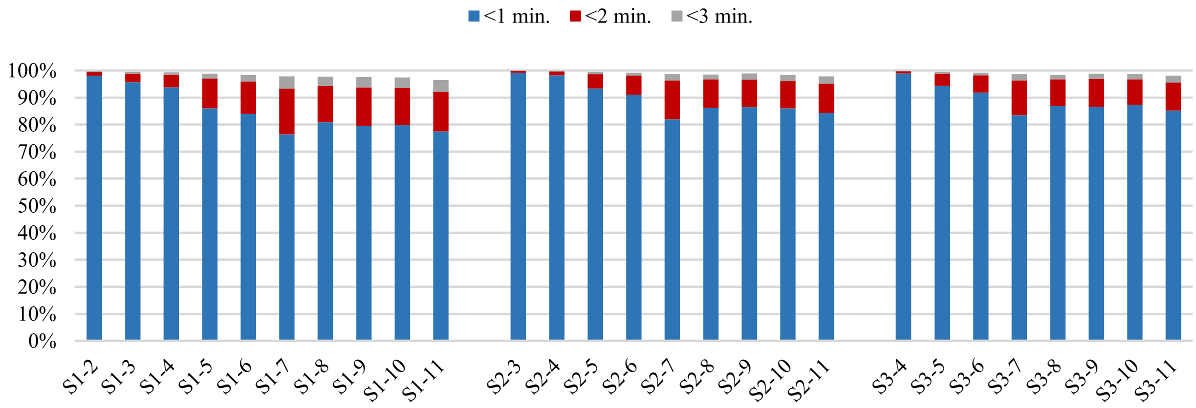

Fig. 14. Absolute prediction errors of the RF (W–T) model for departure stations S1, S2 and S3.

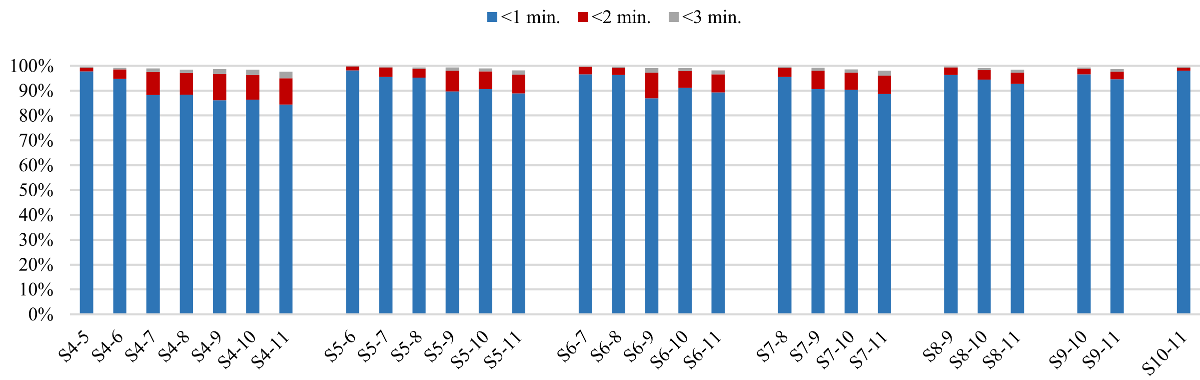

Fig. 15. Absolute prediction errors of the RF (W–T) model for departure stations S4, S5, S6, S7, S8, S9 and S10.

prediction horizons.

Fig. 8 displays the MAE and RMSE values of four types of RF models. The figue indicates that the prediction models, including the variables related to weather conditions and technical problems related to train control systems, namely NoW-T, W-NoT, and W-T, outperform the model without these variables (NoW-NoT). This suggests that including these variables enhances the model's prediction performance. When comparing the W-NoT and NoW-T models, we observe that the W-NoT model exhibits slightly better performance in terms of MAE, while the NoW-T model performs better in terms of RMSE. However, the W-T model, which incorporates the variables related to weather conditions and technical problems related to train control systems, yields the best performance among all the models.

Figs. 9–11 provide a better understanding of how weather conditions and technical problems related to train control systems contribute to predicting arrival delays. These figures demonstrate the accuracy of four RF models across pairs of departure-target stations, indicating the range of prediction horizons. We observe that the W-T model performs better than the other three models across all pairs of departure and target stations. In fact, its performance becomes superior to the others, particularly when the prediction horizon widens.

Figs. 12 and 13 demonstrate the extent to which the RF (W-T) model improves arrival delay predictions compared to two baseline models (DDBP and MP) for all pairs of departure and target stations. We observe that the RF (W-T) model produces between 40% and 90% more accurate predictions than the baseline models for all departure-target station pairs, except for three pairs (S_2–S_3, S_5–S_6, and S_{10}–S_{11}). In these three pairs, the accuracy of the DDBP model's predictions is better than those of the RF (W-T) model. However, it is important to note that the RF (W-T) and the DDBP models have MAEs of less than 0.2 min for these pairs of departure and target stations. Thus, the difference between the MAEs of the two models is insignificant in real life. Generally, the superiority of the RF (W-T) model's predictions increases with longer prediction horizons.

Figs. 14 and 15 offer some additional insights regarding the predictive capabilities of the RF (W-T) model. The figures display the accuracy rates of the model across a range of absolute prediction errors, which are the absolute differences between the predicted and actual arrival delays. When assuming a maximum acceptable prediction error of 1 min, the accuracy rates vary from 76% to 99% across

Table 9
The Standard deviations of predictions per train and target station.

Target Station	Standard Deviation		
	≤1 min.	2 min.	≥3 min.
S3	99.0%	0.5%	0.6%
S4	98.9%	0.7%	0.4%
S5	98.6%	1.0%	0.4%
S6	97.8%	1.5%	0.7%
S7	97.6%	1.8%	0.7%
S8	97.1%	2.0%	0.9%
S9	97.7%	1.6%	0.7%
S10	97.5%	1.7%	0.7%
S11	96.7%	2.2%	1.1%

Table 10
Negative actual prediction errors for target stations S11, S10 and S9.[a]

	RF (W-T)			DDBP		
Depart.-Target	(<-2 min.)	(-2 min)	(-1 min.)	(<-2 min.)	(-2 min)	(-1 min.)
S1-S11	3.1%	8.8%	25.4%	0.0%	0.0%	0.1%
S2-S11	1.6%	6.6%	24.7%	0.0%	0.2%	0.5%
S3-S11	1.2%	5.1%	22.2%	0.0%	0.1%	0.4%
S4-S11	1.5%	5.2%	23.2%	0.0%	0.1%	0.7%
S5-S11	1.1%	4.3%	24.1%	0.1%	0.4%	1.0%
S6-S11	0.8%	3.7%	22.3%	0.0%	0.1%	0.9%
S7-S11	1.1%	4.0%	25.1%	0.1%	0.3%	4.4%
S8-S11	0.6%	2.3%	21.4%	0.1%	0.4%	4.5%
S9-S11	0.2%	1.3%	5.0%	0.1%	0.7%	1.4%
S10-S11	0.1%	0.5%	4.6%	0.0%	0.3%	1.1%
S1-S10	2.7%	8.4%	26.2%	0.0%	0.0%	0.0%
S2-S10	1.5%	6.2%	25.6%	0.0%	0.1%	0.4%
S3-S10	0.8%	4.9%	23.2%	0.0%	0.1%	0.4%
S4-S10	1.0%	5.0%	23.9%	0.0%	0.1%	0.6%
S5-S10	0.8%	4.3%	24.1%	0.1%	0.2%	1.2%
S6-S10	0.6%	3.6%	22.1%	0.0%	0.0%	0.6%
S7-S10	0.8%	3.7%	24.8%	0.1%	0.1%	4.2%
S8-S10	0.4%	2.1%	21.3%	0.0%	0.1%	4.6%
S9-S10	0.2%	1.1%	4.2%	0.1%	0.4%	0.5%
S1-S9	3.2%	8.7%	26.0%	0.0%	0.0%	0.0%
S2-S9	1.6%	6.3%	25.4%	0.0%	0.2%	0.4%
S3-S9	1.0%	5.5%	23.2%	0.0%	0.1%	0.3%
S4-S9	1.0%	5.3%	22.5%	0.0%	0.1%	0.5%
S5-S9	0.7%	4.7%	23.1%	0.0%	0.2%	1.1%
S6-S9	0.8%	5.1%	21.5%	0.0%	0.0%	0.7%
S7-S9	0.4%	3.5%	25.9%	0.0%	0.0%	4.0%
S8-S9	0.1%	1.7%	20.7%	0.0%	0.0%	4.4%

[a] The results of MP model, the other baseline model, are not included in the table as its results regarding actual negative prediction errors are much higher than those of RF (W-T) and DDBP models.

pairs of departure and target stations. If we increase the acceptable error to 2 min, the accuracy rates improve significantly, ranging from 92% to 99%.

It is essential to consider certain criteria while measuring the arrival delay prediction performance from a passenger's perspective. As described in Section 1, one such criterion is the stability of arrival delay predictions for a specific train at a particular target station. In order to measure this, we calculate the standard deviation of all arrival delay predictions for each train at each target station. For example, if a train travels from S1 to S11, the W-T model updates the arrival delay at S11 ten times during the journey whenever it departs from S1, S2, S3, and so on until S10. The standard deviation of these ten predictions is considered a measure of the stability of the predictions for this train at S11. The lower the standard deviation of arrival delay predictions for a train at a particular station, the more stable the predictions are for that train at that station. Table 9 shows the percentage of trains based on the standard deviation of arrival delay predictions. The S11 has the lowest ratio (96.7%) for the standard deviation of less than or equal to 1 min, meaning that most trains (96.7%) have stable arrival delay predictions for this station. Assuming that a volatility of 2 min is acceptable, even this lowest percentage for S11 increases to 98.9%.

Table 11
Negative actual prediction errors for target stations S8, S7, S6, S5, S4, S3 and S2.[a]

	RF (W-T)			DDBP		
Depart.-Target	(<-2 min.)	(-2 min)	(-1 min.)	(<-2 min.)	(-2 min)	(-1 min.)
S1-S8	2.0%	8.0%	26.1%	0.0%	0.0%	0.1%
S2-S8	1.1%	6.5%	27.3%	0.0%	0.1%	0.5%
S3-S8	0.9%	5.0%	23.4%	0.0%	0.1%	0.5%
S4-S8	0.6%	4.8%	24.5%	0.0%	0.0%	0.4%
S5-S8	0.1%	2.0%	21.7%	0.0%	0.2%	0.8%
S6-S8	0.1%	1.5%	21.6%	0.0%	0.0%	0.0%
S7-S8	0.1%	0.3%	6.6%	0.0%	0.0%	0.1%
S1-S7	3.1%	10.9%	23.6%	0.0%	0.0%	0.0%
S2-S7	1.1%	8.9%	22.9%	0.0%	0.1%	0.8%
S3-S7	0.6%	7.0%	21.7%	0.0%	0.1%	0.7%
S4-S7	0.5%	4.5%	22.8%	0.0%	0.1%	0.7%
S5-S7	0.2%	1.8%	20.6%	0.0%	0.2%	1.2%
S6-S7	0.0%	1.0%	20.1%	0.0%	0.0%	0.0%
S1-S6	1.2%	7.5%	27.0%	0.0%	0.1%	0.2%
S2-S6	0.4%	4.0%	26.0%	0.0%	0.6%	2.3%
S3-S6	0.2%	3.2%	23.8%	0.0%	0.4%	2.1%
S4-S6	0.2%	1.8%	21.3%	0.0%	0.2%	2.5%
S5-S6	0.1%	0.9%	6.4%	0.0%	0.6%	4.2%
S1-S5	0.6%	6.2%	22.4%	0.0%	0.0%	0.1%
S2-S5	0.2%	2.7%	20.8%	0.0%	0.3%	1.8%
S3-S5	0.1%	2.0%	20.0%	0.0%	0.1%	1.5%
S4-S5	0.1%	0.4%	18.2%	0.0%	0.0%	0.6%
S1-S4	0.3%	1.9%	23.7%	0.0%	0.1%	0.3%
S2-S4	0.0%	0.7%	17.3%	0.0%	0.3%	2.1%
S3-S4	0.0%	0.2%	17.2%	0.0%	0.0%	1.8%
S1-S3	0.2%	0.9%	13.5%	0.0%	0.1%	0.3%
S2-S3	0.1%	0.3%	2.7%	0.0%	0.0%	1.1%
S1-S2	0.2%	0.6%	9.9%	0.0%	0.1%	0.3%

[a] The results of MP model, the other baseline model, are not included in the table as its results regarding actual negative prediction errors are much higher than those of RF (W-T) and DDBP models.

Another crucial factor to consider when evaluating arrival delay predictions from a passenger's perspective is the proportion of late predictions. Tables 10 and 11 show the rates of the negative actual prediction errors at target stations. Negative actual prediction errors denote late predictions, which means the predicted delay is greater than the actual arrival delay. The tables indicate that the RF (W-T) model produces more late predictions than the DDBP model across all target stations. While the ratio of 1-min late prediction for the RF (W-T) model mostly ranges from around 20%–25%, the percentage of 2-min (or greater) late prediction is between 0.5% and 14%. However, the rate of late predictions, especially for 2-min (or greater) late predictions, decreases as the prediction horizon diminishes. The results show that the highest rate of 2-min (or greater) late predictions for one station ahead predictions (e.g., S10–S11 and S9–S10) is 1.8% for S8–S9. For stop stations S2, S5, and S11, where passengers board and alight, the rates of 2-min (or greater) late predictions are 0.8%, 0.5%, and 0.6%, respectively.

4. Conclusions

The study developed a novel arrival delay prediction model using random forest regression based on the train operation data from the Ankara - Eskişehir high-speed train line in Turkey from July 2021 to October 2023. The model can simultaneously predict arrival delays at all downstream stations on this line and continuously update these predictions as new information about train movements becomes available. The accuracy rates of this model vary from 76% to 99% across pairs of departure and target stations if we assume that the prediction error of 1 min is acceptable. If we increase the acceptable error to 2 min, the accuracy rates improve significantly, ranging from 92% to 99%. The study has also explored the effect of weather conditions and technical problems related to train control systems on arrival delay prediction accuracy. The results show that incorporating these variables into the model improves prediction performance. Their contribution to the model performance increases as the prediction horizon widens. These results suggest that the model predictions may assist network managers in making better decisions about train operations.

To evaluate the model's performance from passengers' point of view, the study has proposed two methods, namely the stability of forecasts and the proportion of late predictions, in addition to other commonly used accuracy metrics. According to the stability of forecasts, the model performs quite well for all target stations. Most trains (between 96.7% and 99%) have stable arrival delay predictions at target stations. On the other hand, the proportion of late predictions, where the predicted delay exceeds the actual arrival delay, is relatively high. While the ratio of 1-min late predictions mostly ranges around 20%–25%, the percentage of 2-min (or greater) late predictions is between 0.5% and 14%. As the prediction horizon decreases, the ratios of late predictions decrease, too. For one-station ahead predictions, the highest rate of 2-min (or greater) late predictions is 1.8%. However, from passengers' perspective, while a one or 2-min late prediction might be tolerable for long prediction horizons (e.g., 40–50 min ahead), a prediction of more than 1-min late might have negative consequences (e.g., wrongly encouraging passengers to leave stations) as the train approaches the target station (e.g., 10–15 min ahead). Therefore, one might consider using the predictions of the DDBP model for one or two-station ahead horizons as its proportion of 2-min (or greater) late prediction is almost zero for such horizons. The RF (W-T) model might be preferred for longer prediction horizons. Nevertheless, it is necessary to be cautious while using the model's predictions to inform passengers, even for longer prediction horizons, due to the possibility of late predictions.

Further research should expand the models to cover further distances (e.g., more stations) on the same high-speed line and other high-speed lines with different technical and operational features. It might be interesting to see whether we observe the same patterns in the predictive powers of these models. Furthermore, the model might be further enhanced to lower the proportion of late predictions to use the model predictions safely to inform passengers.

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

The author thanks the two anonymous reviewers for their useful comments on the draft version of this article.

Appendix 1. MAE and RMSE for training and testing samples

RF-Model	MAE		RMSE	
	Training	Testing	Training	Testing
NoW-NoT	0.73	0.84	1.43	1.66
NoW-T	0.68	0.78	1.29	1.52
W-NoT	0.66	0.77	1.34	1.58
W-T	0.62	0.73	1.20	1.43

Appendix 2. MAE and RMSE for training and testing samples across the pairs of departure and target stations

Depart. Station	Target Station	NoW-NoT Model				NoW-T Model				W-NoT Model				W-T Model			
		MAE		RMSE		MAE		RMSE		MAE		RMSE		MAE		RMSE	
		Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.
S1	S2	0.38	0.47	0.87	1.21	0.39	0.44	0.77	1.06	0.33	0.45	0.82	1.18	0.37	0.43	0.74	1.05
S1	S3	0.51	0.60	1.13	1.42	0.48	0.55	1.02	1.28	0.44	0.58	1.07	1.39	0.44	0.53	0.96	1.25
S1	S4	0.70	0.80	1.32	1.45	0.66	0.74	1.17	1.36	0.64	0.76	1.25	1.43	0.60	0.71	1.11	1.33
S1	S5	0.89	1.02	1.57	1.77	0.84	0.95	1.43	1.64	0.83	0.98	1.50	1.73	0.79	0.91	1.37	1.57
S1	S6	0.98	1.09	1.68	1.92	0.92	1.03	1.54	1.80	0.94	1.06	1.62	1.87	0.86	0.99	1.46	1.74
S1	S7	1.20	1.39	1.90	2.05	1.12	1.32	1.73	1.91	1.11	1.29	1.80	1.93	1.03	1.23	1.62	1.80
S1	S8	1.15	1.33	1.93	2.13	1.07	1.23	1.74	1.94	1.05	1.23	1.80	2.00	0.95	1.12	1.59	1.77
S1	S9	1.26	1.52	2.02	2.41	1.16	1.40	1.81	2.20	1.11	1.32	1.82	2.09	0.99	1.22	1.60	1.88
S1	S10	1.30	1.65	2.14	3.37	1.16	1.51	1.88	3.01	1.15	1.47	1.93	3.14	1.00	1.32	1.63	2.77
S1	S11	1.40	1.58	2.38	2.89	1.26	1.42	2.11	2.61	1.23	1.40	2.17	2.71	1.08	1.25	1.87	2.39
S2	S3	0.07	0.08	0.34	0.39	0.06	0.08	0.29	0.37	0.06	0.08	0.33	0.39	0.07	0.08	0.30	0.36
S2	S4	0.36	0.38	0.72	0.82	0.34	0.37	0.65	0.76	0.34	0.37	0.70	0.81	0.33	0.36	0.64	0.75
S2	S5	0.59	0.67	1.14	1.18	0.57	0.64	1.04	1.12	0.56	0.64	1.10	1.14	0.54	0.62	0.99	1.08
S2	S6	0.71	0.83	1.27	1.40	0.67	0.78	1.17	1.29	0.69	0.79	1.23	1.35	0.64	0.76	1.12	1.25
S2	S7	0.95	1.08	1.44	1.87	0.92	1.05	1.36	1.71	0.91	1.07	1.40	1.85	0.88	1.03	1.31	1.66
S2	S8	0.93	0.99	1.60	1.56	0.86	0.98	1.45	1.53	0.86	0.95	1.49	1.49	0.80	0.93	1.34	1.42
S2	S9	1.01	1.18	1.62	1.92	0.95	1.12	1.43	1.87	0.88	1.05	1.43	1.66	0.79	0.97	1.24	1.56
S2	S10	1.07	1.25	1.95	2.00	0.97	1.15	1.69	1.83	0.94	1.07	1.74	1.79	0.82	0.96	1.50	1.61
S2	S11	1.12	1.31	2.00	2.20	1.02	1.21	1.77	1.97	0.99	1.16	1.81	2.01	0.87	1.05	1.58	1.77
S3	S4	0.33	0.33	0.66	0.68	0.32	0.33	0.63	0.65	0.32	0.33	0.65	0.67	0.31	0.33	0.62	0.64
S3	S5	0.57	0.64	1.08	1.22	0.56	0.63	0.99	1.15	0.54	0.62	1.05	1.18	0.53	0.60	0.95	1.11

(*continued on next page*)

(*continued*)

Depart. Station	Target Station	NoW-NoT Model				NoW-T Model				W-NoT Model				W-T Model			
		MAE		RMSE		MAE		RMSE		MAE		RMSE		MAE		RMSE	
		Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.	Train.	Test.
S3	S6	0.67	0.78	1.21	1.41	0.64	0.76	1.12	1.33	0.64	0.76	1.18	1.37	0.61	0.74	1.09	1.27
S3	S7	0.91	1.06	1.42	1.79	0.89	1.05	1.34	1.68	0.88	1.03	1.38	1.76	0.86	1.00	1.30	1.60
S3	S8	0.87	1.06	1.54	1.78	0.83	1.00	1.41	1.61	0.81	0.99	1.45	1.69	0.78	0.95	1.31	1.56
S3	S9	0.99	1.17	1.62	1.87	0.93	1.10	1.43	1.73	0.86	1.01	1.41	1.67	0.79	0.94	1.23	1.52
S3	S10	1.02	1.28	1.77	2.62	0.93	1.15	1.53	2.16	0.89	1.12	1.58	2.47	0.78	0.98	1.33	2.12
S3	S11	1.08	1.33	1.87	2.64	1.00	1.18	1.66	2.14	0.95	1.17	1.69	2.48	0.85	1.03	1.47	2.13
S4	S5	0.39	0.42	0.91	0.81	0.38	0.41	0.85	0.81	0.38	0.40	0.90	0.79	0.37	0.41	0.83	0.80
S4	S6	0.52	0.56	1.11	1.06	0.52	0.56	1.06	1.04	0.50	0.55	1.08	1.04	0.50	0.54	1.02	1.01
S4	S7	0.82	0.94	1.29	1.68	0.80	0.92	1.23	1.57	0.78	0.93	1.24	1.66	0.76	0.89	1.18	1.50
S4	S8	0.78	0.86	1.38	1.70	0.75	0.84	1.31	1.63	0.75	0.84	1.35	1.68	0.73	0.82	1.28	1.58
S4	S9	0.96	1.13	1.57	1.81	0.90	1.08	1.40	1.64	0.85	1.02	1.41	1.68	0.80	0.95	1.28	1.48
S4	S10	1.04	1.17	1.98	1.89	0.94	1.07	1.68	1.68	0.92	1.05	1.80	1.76	0.83	0.95	1.56	1.53
S4	S11	1.10	1.24	2.02	2.13	0.99	1.15	1.70	2.05	0.98	1.12	1.86	2.00	0.88	1.05	1.59	1.90
S5	S6	0.16	0.17	0.49	0.57	0.14	0.15	0.47	0.56	0.14	0.15	0.47	0.57	0.13	0.14	0.45	0.53
S5	S7	0.61	0.68	0.95	1.03	0.60	0.67	0.92	1.01	0.58	0.67	0.91	1.01	0.58	0.67	0.90	0.99
S5	S8	0.52	0.57	1.05	1.03	0.50	0.56	0.97	0.99	0.49	0.55	1.01	0.99	0.47	0.54	0.94	0.94
S5	S9	0.81	0.98	1.30	1.55	0.77	0.95	1.19	1.45	0.74	0.92	1.22	1.47	0.70	0.88	1.14	1.37
S5	S10	0.84	1.05	1.49	2.43	0.76	0.98	1.32	1.96	0.76	0.97	1.39	2.35	0.68	0.86	1.22	1.83
S5	S11	0.92	1.08	1.80	1.81	0.83	1.03	1.53	1.76	0.85	0.99	1.72	1.70	0.75	0.92	1.44	1.59
S6	S7	0.59	0.64	0.88	0.97	0.60	0.64	0.88	0.95	0.55	0.63	0.83	0.95	0.56	0.63	0.84	0.94
S6	S8	0.48	0.55	0.93	1.13	0.47	0.54	0.87	1.06	0.44	0.52	0.89	1.10	0.43	0.52	0.83	1.04
S6	S9	0.85	1.00	1.38	1.83	0.83	0.99	1.28	1.75	0.82	0.97	1.32	1.79	0.80	0.96	1.23	1.71
S6	S10	0.86	1.01	1.68	1.68	0.77	0.93	1.41	1.50	0.77	0.90	1.56	1.57	0.68	0.83	1.32	1.41
S6	S11	0.91	1.02	1.78	1.65	0.84	0.94	1.53	1.46	0.83	0.95	1.66	1.59	0.74	0.87	1.44	1.41
S7	S8	0.36	0.37	0.88	0.81	0.30	0.30	0.81	0.74	0.28	0.31	0.80	0.78	0.24	0.26	0.75	0.70
S7	S9	0.74	0.82	1.37	1.31	0.71	0.79	1.27	1.25	0.71	0.79	1.34	1.27	0.68	0.76	1.25	1.21
S7	S10	0.85	0.93	1.78	1.75	0.75	0.84	1.59	1.53	0.78	0.87	1.71	1.71	0.68	0.80	1.46	1.54
S7	S11	0.90	1.03	1.79	2.16	0.81	0.96	1.62	1.95	0.83	0.98	1.71	2.13	0.74	0.89	1.52	1.86
S8	S9	0.56	0.61	1.08	1.04	0.54	0.59	1.01	1.00	0.53	0.60	1.07	1.05	0.52	0.58	1.01	1.02
S8	S10	0.70	0.78	1.62	1.87	0.64	0.72	1.47	1.72	0.62	0.72	1.52	1.80	0.56	0.66	1.31	1.60
S8	S11	0.76	0.87	1.71	1.99	0.69	0.79	1.56	1.83	0.69	0.79	1.62	1.85	0.62	0.72	1.42	1.63
S9	S10	0.34	0.39	1.00	1.13	0.19	0.24	0.84	1.03	0.20	0.27	0.91	1.15	0.18	0.25	0.81	1.07
S9	S11	0.43	0.45	1.18	1.15	0.29	0.31	1.05	0.98	0.29	0.33	1.12	1.10	0.27	0.29	1.04	0.99
S10	S11	0.24	0.25	0.61	0.66	0.13	0.14	0.51	0.57	0.11	0.13	0.50	0.58	0.11	0.13	0.51	0.57

Appendix 3. MAE, RMSE and OBB values of RF (W-T) model for training and testing samples

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
500	8	30	1.37	0.62	1.20	0.73	1.42
400	8	30	1.37	0.62	1.20	0.73	1.43
200	8	30	1.38	0.62	1.20	0.73	1.43
300	8	30	1.38	0.62	1.20	0.73	1.43
400	6	30	1.38	0.62	1.21	0.73	1.43
500	6	30	1.38	0.62	1.21	0.73	1.44
400	8	25	1.38	0.63	1.21	0.73	1.44
300	6	30	1.38	0.62	1.21	0.73	1.44
300	8	25	1.38	0.63	1.21	0.73	1.44
500	8	25	1.38	0.63	1.22	0.73	1.43
500	6	25	1.38	0.64	1.22	0.74	1.44
200	8	25	1.38	0.63	1.22	0.73	1.44
500	10	30	1.39	0.63	1.21	0.74	1.44
400	10	30	1.39	0.63	1.21	0.74	1.44
100	8	30	1.39	0.62	1.20	0.73	1.43
300	10	30	1.39	0.62	1.21	0.73	1.44
200	6	30	1.39	0.63	1.21	0.73	1.45
400	6	25	1.39	0.64	1.22	0.74	1.44
300	6	25	1.39	0.64	1.22	0.74	1.44
200	10	30	1.39	0.62	1.21	0.73	1.44
200	6	25	1.39	0.64	1.22	0.74	1.45
200	10	25	1.39	0.64	1.22	0.74	1.45
500	10	25	1.39	0.64	1.23	0.75	1.45

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
400	10	25	1.39	0.64	1.23	0.75	1.45
100	8	25	1.39	0.64	1.22	0.74	1.44
100	10	30	1.39	0.62	1.21	0.73	1.44
300	10	25	1.40	0.63	1.23	0.74	1.45
100	6	25	1.40	0.64	1.22	0.74	1.44
100	6	30	1.40	0.63	1.21	0.73	1.45
100	10	25	1.40	0.63	1.22	0.74	1.44
500	8	20	1.41	0.69	1.28	0.76	1.46
400	8	20	1.41	0.69	1.28	0.76	1.46
200	8	20	1.41	0.69	1.28	0.76	1.47
300	8	20	1.41	0.69	1.28	0.76	1.46
200	4	30	1.41	0.64	1.24	0.75	1.47
500	4	30	1.41	0.64	1.24	0.74	1.48
400	6	20	1.42	0.69	1.29	0.76	1.47
500	6	20	1.42	0.69	1.29	0.76	1.47
300	6	20	1.42	0.69	1.29	0.76	1.48
300	4	25	1.42	0.66	1.25	0.75	1.48
100	8	20	1.42	0.69	1.28	0.76	1.47
500	4	25	1.42	0.66	1.26	0.75	1.49
300	4	30	1.42	0.64	1.25	0.75	1.49
200	6	20	1.42	0.69	1.29	0.76	1.48
200	4	25	1.42	0.66	1.26	0.75	1.50
500	10	20	1.42	0.70	1.29	0.77	1.48
300	10	20	1.42	0.69	1.29	0.76	1.47
100	4	30	1.43	0.65	1.24	0.75	1.47
400	10	20	1.43	0.70	1.30	0.77	1.48
200	10	20	1.43	0.70	1.29	0.77	1.48
100	6	20	1.43	0.70	1.29	0.77	1.49
400	4	25	1.43	0.66	1.26	0.75	1.49
400	4	30	1.43	0.64	1.26	0.75	1.50
100	10	20	1.44	0.70	1.30	0.78	1.48
100	4	25	1.45	0.66	1.27	0.76	1.50
500	4	20	1.45	0.71	1.33	0.78	1.52
300	4	20	1.45	0.71	1.33	0.78	1.52
200	4	20	1.46	0.71	1.33	0.78	1.51
400	4	20	1.46	0.71	1.33	0.78	1.53
100	4	20	1.48	0.72	1.34	0.78	1.54
500	8	15	1.49	0.79	1.43	0.82	1.54
300	8	15	1.50	0.79	1.43	0.82	1.55
400	8	15	1.50	0.79	1.43	0.82	1.55
200	8	15	1.50	0.79	1.43	0.82	1.55
500	6	15	1.50	0.79	1.44	0.82	1.56
200	6	15	1.50	0.79	1.44	0.82	1.55
400	6	15	1.50	0.79	1.44	0.82	1.56
300	6	15	1.50	0.79	1.44	0.82	1.55
100	8	15	1.50	0.79	1.43	0.82	1.55
400	10	15	1.51	0.79	1.43	0.82	1.55
500	10	15	1.51	0.80	1.44	0.83	1.55
300	10	15	1.51	0.79	1.43	0.82	1.55
200	10	15	1.51	0.79	1.44	0.82	1.55
100	10	15	1.51	0.80	1.44	0.83	1.56
100	6	15	1.51	0.80	1.44	0.83	1.57
400	4	15	1.54	0.81	1.47	0.83	1.60
200	4	15	1.54	0.81	1.47	0.84	1.60
500	4	15	1.54	0.81	1.47	0.83	1.60
300	4	15	1.54	0.81	1.47	0.83	1.60
100	4	15	1.55	0.81	1.47	0.84	1.60
100	2	30	1.64	0.80	1.48	0.88	1.72
200	2	30	1.65	0.80	1.50	0.88	1.73
400	2	30	1.65	0.81	1.51	0.88	1.74
300	8	10	1.65	0.92	1.64	0.92	1.70
500	2	30	1.65	0.82	1.51	0.89	1.75
300	2	30	1.65	0.82	1.51	0.89	1.74
300	2	25	1.65	0.82	1.51	0.89	1.74
500	2	25	1.65	0.82	1.52	0.89	1.75
500	8	10	1.65	0.92	1.64	0.92	1.70
400	2	25	1.65	0.82	1.52	0.89	1.75
400	8	10	1.65	0.91	1.64	0.92	1.70

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
200	8	10	1.65	0.92	1.64	0.92	1.70
500	6	10	1.65	0.92	1.64	0.92	1.70
300	6	10	1.65	0.92	1.64	0.92	1.70
200	2	25	1.65	0.82	1.51	0.88	1.75
200	6	10	1.66	0.92	1.64	0.92	1.71
400	6	10	1.66	0.92	1.64	0.92	1.71
100	8	10	1.66	0.92	1.64	0.92	1.70
100	6	10	1.67	0.92	1.65	0.92	1.70
400	10	10	1.67	0.93	1.65	0.94	1.71
400	2	20	1.67	0.86	1.56	0.91	1.77
500	10	10	1.67	0.93	1.66	0.94	1.72
300	10	10	1.67	0.93	1.66	0.94	1.71
200	10	10	1.67	0.93	1.65	0.94	1.71
100	10	10	1.68	0.93	1.66	0.94	1.72
300	2	20	1.68	0.87	1.57	0.92	1.76
100	2	25	1.68	0.84	1.53	0.91	1.74
200	2	20	1.68	0.87	1.57	0.92	1.78
500	2	20	1.69	0.87	1.58	0.92	1.79
400	4	10	1.70	0.95	1.68	0.96	1.75
500	4	10	1.70	0.95	1.68	0.96	1.76
300	4	10	1.70	0.95	1.68	0.96	1.76
200	4	10	1.70	0.95	1.68	0.96	1.76
100	4	10	1.71	0.96	1.69	0.97	1.76
100	2	20	1.71	0.86	1.57	0.91	1.79
100	2	15	1.77	0.97	1.69	1.00	1.84
400	2	15	1.77	0.98	1.71	1.01	1.87
500	2	15	1.78	0.98	1.71	1.01	1.88
300	2	15	1.78	0.99	1.71	1.01	1.87
200	2	15	1.78	0.99	1.71	1.01	1.87
500	8	5	1.98	1.19	1.98	1.19	2.00
300	8	5	1.98	1.19	1.98	1.19	2.00
400	8	5	1.98	1.19	1.97	1.19	1.99
200	8	5	1.98	1.19	1.98	1.19	2.00
100	8	5	1.98	1.18	1.97	1.18	1.99
500	6	5	1.99	1.19	1.98	1.19	2.02
400	6	5	2.00	1.20	1.99	1.20	2.02
300	6	5	2.00	1.19	1.99	1.19	2.02
200	6	5	2.00	1.20	1.99	1.20	2.03
500	2	10	2.02	1.20	1.99	1.22	2.10
300	2	10	2.03	1.21	1.99	1.22	2.11
300	10	5	2.03	1.24	2.04	1.24	2.07
100	10	5	2.03	1.25	2.04	1.25	2.07
400	2	10	2.03	1.21	2.00	1.22	2.11
500	10	5	2.03	1.24	2.04	1.24	2.07
400	10	5	2.03	1.24	2.04	1.24	2.07
200	10	5	2.03	1.24	2.04	1.24	2.07
100	6	5	2.03	1.21	2.02	1.22	2.06
200	2	10	2.04	1.21	2.00	1.22	2.11
100	2	10	2.05	1.22	2.01	1.23	2.13
500	4	5	2.16	1.34	2.17	1.35	2.24
300	4	5	2.17	1.35	2.17	1.35	2.24
400	4	5	2.17	1.35	2.18	1.36	2.25
200	4	5	2.19	1.39	2.21	1.40	2.28
100	4	5	2.24	1.40	2.24	1.41	2.32
400	2	5	2.70	1.66	2.69	1.67	2.80
300	2	5	2.70	1.66	2.69	1.67	2.80
500	2	5	2.71	1.67	2.71	1.68	2.82
200	2	5	2.71	1.67	2.70	1.68	2.81
100	2	5	2.72	1.67	2.71	1.68	2.83

Appendix 4. MAE, RMSE and OBB values of RF (W-NoT) model for training and testing samples

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
400	7	30	1.53	0.66	1.34	0.77	1.58
500	7	30	1.53	0.67	1.35	0.78	1.59
500	7	25	1.53	0.68	1.36	0.78	1.59
300	7	30	1.53	0.67	1.35	0.78	1.58
400	7	25	1.54	0.68	1.36	0.78	1.59
500	9	30	1.54	0.67	1.35	0.79	1.59
200	7	30	1.54	0.66	1.35	0.78	1.59
400	9	30	1.54	0.67	1.35	0.79	1.59
300	9	30	1.54	0.67	1.34	0.78	1.59
300	7	25	1.54	0.68	1.36	0.78	1.59
500	9	25	1.54	0.69	1.36	0.79	1.59
500	5	30	1.54	0.67	1.35	0.78	1.60
400	9	25	1.54	0.68	1.36	0.78	1.59
300	5	30	1.54	0.67	1.36	0.78	1.60
200	9	30	1.54	0.67	1.35	0.79	1.59
400	5	25	1.54	0.68	1.36	0.79	1.60
200	5	30	1.54	0.67	1.35	0.78	1.60
200	7	25	1.54	0.68	1.36	0.78	1.59
100	9	30	1.54	0.66	1.35	0.78	1.58
400	5	30	1.55	0.67	1.36	0.78	1.60
300	9	25	1.55	0.68	1.36	0.79	1.59
500	5	25	1.55	0.68	1.37	0.79	1.60
100	7	30	1.55	0.67	1.36	0.78	1.59
200	9	25	1.55	0.69	1.37	0.79	1.59
100	7	25	1.55	0.68	1.36	0.79	1.59
300	5	25	1.55	0.69	1.37	0.79	1.61
200	5	25	1.55	0.68	1.37	0.79	1.61
100	5	30	1.56	0.68	1.36	0.79	1.61
100	9	25	1.56	0.69	1.37	0.80	1.60
100	5	25	1.56	0.69	1.38	0.79	1.62
400	7	20	1.57	0.73	1.42	0.80	1.61
500	7	20	1.57	0.73	1.42	0.80	1.62
300	9	20	1.57	0.73	1.42	0.81	1.62
300	7	20	1.57	0.73	1.42	0.81	1.62
200	7	20	1.57	0.73	1.43	0.81	1.62
500	9	20	1.57	0.74	1.43	0.82	1.62
400	5	20	1.57	0.74	1.43	0.81	1.63
400	9	20	1.57	0.74	1.43	0.82	1.62
100	7	20	1.57	0.73	1.42	0.81	1.62
200	9	20	1.57	0.74	1.43	0.82	1.62
500	5	20	1.58	0.74	1.43	0.81	1.63
300	5	20	1.58	0.74	1.44	0.81	1.63
200	5	20	1.58	0.74	1.44	0.81	1.64
100	9	20	1.58	0.75	1.43	0.82	1.63
100	5	20	1.59	0.74	1.44	0.81	1.64
400	3	30	1.60	0.71	1.42	0.81	1.68
200	3	30	1.60	0.70	1.41	0.81	1.66
500	3	30	1.60	0.70	1.42	0.81	1.68
500	3	25	1.60	0.72	1.43	0.81	1.68
400	3	25	1.60	0.72	1.43	0.81	1.67
300	3	25	1.60	0.72	1.43	0.81	1.67
300	3	30	1.61	0.71	1.42	0.81	1.68
200	3	25	1.61	0.72	1.44	0.82	1.69
100	3	25	1.61	0.72	1.43	0.82	1.68
500	3	20	1.63	0.77	1.49	0.83	1.70
100	3	30	1.63	0.71	1.42	0.81	1.69
300	3	20	1.63	0.77	1.50	0.83	1.70
400	3	20	1.63	0.77	1.50	0.83	1.69
200	3	20	1.63	0.77	1.49	0.84	1.70
500	7	15	1.64	0.83	1.58	0.86	1.69
100	3	20	1.64	0.77	1.50	0.84	1.71
400	7	15	1.65	0.83	1.58	0.86	1.69
300	7	15	1.65	0.83	1.58	0.86	1.69
200	7	15	1.65	0.83	1.58	0.86	1.70
400	9	15	1.65	0.84	1.58	0.87	1.70
500	9	15	1.65	0.84	1.59	0.87	1.70

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
300	9	15	1.65	0.84	1.58	0.87	1.70
100	7	15	1.65	0.83	1.58	0.86	1.70
200	9	15	1.65	0.84	1.58	0.87	1.70
300	5	15	1.65	0.83	1.59	0.86	1.70
500	5	15	1.65	0.83	1.59	0.86	1.71
400	5	15	1.66	0.83	1.59	0.86	1.71
200	5	15	1.66	0.83	1.59	0.86	1.71
100	9	15	1.66	0.84	1.59	0.87	1.70
100	5	15	1.67	0.83	1.60	0.86	1.72
400	3	15	1.71	0.86	1.65	0.89	1.78
300	3	15	1.72	0.87	1.65	0.89	1.78
500	3	15	1.72	0.87	1.66	0.89	1.79
200	3	15	1.72	0.87	1.66	0.89	1.80
100	3	15	1.74	0.87	1.66	0.90	1.80
500	7	10	1.77	0.94	1.76	0.95	1.81
400	7	10	1.77	0.94	1.76	0.95	1.81
200	7	10	1.77	0.94	1.76	0.94	1.81
300	7	10	1.77	0.94	1.76	0.94	1.81
100	7	10	1.77	0.94	1.76	0.95	1.81
300	9	10	1.77	0.94	1.76	0.95	1.81
500	9	10	1.77	0.95	1.77	0.96	1.82
400	9	10	1.77	0.95	1.77	0.96	1.82
200	9	10	1.78	0.94	1.77	0.95	1.82
500	5	10	1.78	0.95	1.77	0.95	1.83
200	5	10	1.78	0.95	1.77	0.95	1.82
100	9	10	1.78	0.95	1.77	0.96	1.82
400	5	10	1.78	0.95	1.77	0.95	1.83
300	5	10	1.78	0.94	1.77	0.95	1.83
100	5	10	1.79	0.95	1.78	0.95	1.83
500	3	10	1.86	1.03	1.85	1.04	1.93
300	3	10	1.87	1.03	1.85	1.04	1.93
200	3	10	1.87	1.04	1.85	1.05	1.93
400	3	10	1.87	1.03	1.86	1.04	1.93
100	3	10	1.91	1.06	1.88	1.07	1.96
400	7	5	2.02	1.19	2.03	1.19	2.05
500	7	5	2.03	1.20	2.04	1.19	2.05
300	7	5	2.03	1.20	2.04	1.20	2.06
200	7	5	2.03	1.20	2.04	1.20	2.06
100	7	5	2.03	1.19	2.03	1.19	2.06
400	9	5	2.06	1.22	2.08	1.22	2.10
500	9	5	2.06	1.25	2.09	1.25	2.11
200	9	5	2.06	1.23	2.08	1.23	2.10
100	9	5	2.06	1.25	2.08	1.25	2.09
300	9	5	2.06	1.24	2.08	1.24	2.10
500	5	5	2.08	1.22	2.08	1.23	2.12
400	5	5	2.08	1.22	2.08	1.23	2.12
200	5	5	2.08	1.22	2.08	1.22	2.11
300	5	5	2.08	1.23	2.08	1.23	2.12
100	5	5	2.09	1.22	2.08	1.22	2.12
300	3	5	2.36	1.45	2.37	1.45	2.46
500	3	5	2.37	1.46	2.38	1.46	2.46
400	3	5	2.37	1.45	2.38	1.46	2.47
200	3	5	2.38	1.45	2.38	1.46	2.46
300	1	30	2.39	1.34	2.32	1.38	2.49
200	1	25	2.42	1.37	2.34	1.40	2.52
300	1	25	2.42	1.36	2.35	1.40	2.53
300	1	20	2.43	1.38	2.37	1.41	2.53
100	3	5	2.43	1.48	2.42	1.49	2.51
400	1	25	2.45	1.38	2.38	1.41	2.56
500	1	30	2.45	1.38	2.38	1.41	2.56
200	1	20	2.45	1.38	2.37	1.41	2.54
400	1	20	2.45	1.38	2.38	1.41	2.56
400	1	30	2.45	1.38	2.38	1.42	2.56
300	1	15	2.45	1.41	2.40	1.43	2.56
200	1	30	2.46	1.37	2.37	1.40	2.56
100	1	20	2.46	1.38	2.37	1.41	2.53
500	1	25	2.47	1.39	2.40	1.43	2.58
500	1	20	2.47	1.41	2.41	1.44	2.58

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
100	1	25	2.48	1.39	2.38	1.42	2.56
100	1	30	2.51	1.40	2.42	1.44	2.60
200	1	15	2.51	1.44	2.46	1.47	2.61
400	1	15	2.53	1.47	2.48	1.49	2.64
500	1	15	2.54	1.47	2.49	1.49	2.64
100	1	15	2.55	1.47	2.49	1.49	2.65
300	1	10	2.69	1.60	2.66	1.62	2.79
400	1	10	2.70	1.61	2.68	1.63	2.80
500	1	10	2.71	1.62	2.69	1.64	2.82
200	1	10	2.72	1.60	2.69	1.62	2.81
100	1	10	2.77	1.64	2.73	1.66	2.86
300	1	5	3.17	1.92	3.17	1.93	3.30
200	1	5	3.18	1.93	3.18	1.94	3.30
500	1	5	3.20	1.93	3.20	1.94	3.32
400	1	5	3.20	1.93	3.21	1.94	3.33
100	1	5	3.26	1.95	3.25	1.96	3.38

Appendix 5. MAE, RMSE and OBB values of RF (NoW-T) model for training and testing samples

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
400	7	30	1.47	0.68	1.29	0.78	1.51
500	7	30	1.47	0.68	1.30	0.79	1.52
400	7	25	1.47	0.69	1.30	0.79	1.52
200	7	30	1.47	0.68	1.29	0.79	1.52
500	7	25	1.47	0.69	1.31	0.79	1.52
500	5	30	1.47	0.69	1.31	0.79	1.52
300	5	30	1.47	0.69	1.31	0.79	1.52
400	5	25	1.47	0.70	1.31	0.79	1.52
300	7	25	1.47	0.69	1.30	0.79	1.52
300	7	30	1.47	0.69	1.30	0.79	1.52
400	5	30	1.47	0.69	1.31	0.79	1.52
500	5	25	1.47	0.70	1.32	0.79	1.52
200	5	30	1.47	0.69	1.31	0.79	1.52
300	5	25	1.47	0.70	1.31	0.79	1.52
200	5	25	1.47	0.70	1.31	0.79	1.52
200	7	25	1.48	0.70	1.31	0.79	1.53
100	7	30	1.48	0.68	1.29	0.78	1.52
100	5	30	1.48	0.69	1.31	0.79	1.53
100	7	25	1.48	0.70	1.31	0.80	1.53
100	5	25	1.48	0.70	1.32	0.79	1.52
400	7	20	1.48	0.74	1.35	0.80	1.53
200	7	20	1.48	0.73	1.35	0.80	1.53
300	7	20	1.48	0.74	1.35	0.81	1.54
500	7	20	1.48	0.74	1.36	0.81	1.54
500	5	20	1.49	0.73	1.36	0.80	1.53
400	5	20	1.49	0.73	1.36	0.80	1.54
300	5	20	1.49	0.73	1.36	0.80	1.53
200	5	20	1.49	0.73	1.36	0.80	1.54
100	7	20	1.49	0.74	1.36	0.81	1.54
100	5	20	1.49	0.73	1.36	0.80	1.54
200	3	30	1.53	0.74	1.39	0.81	1.58
400	3	30	1.53	0.74	1.39	0.81	1.58
300	3	25	1.53	0.74	1.39	0.81	1.58
200	3	25	1.53	0.74	1.40	0.81	1.58
300	3	30	1.53	0.74	1.39	0.81	1.58
500	3	30	1.53	0.74	1.40	0.81	1.58
500	3	25	1.53	0.74	1.40	0.81	1.58
400	3	25	1.53	0.74	1.40	0.81	1.58
100	3	30	1.53	0.74	1.39	0.81	1.58
300	3	20	1.54	0.76	1.43	0.81	1.59
200	3	20	1.54	0.76	1.43	0.82	1.59
100	3	25	1.54	0.74	1.40	0.81	1.58
400	7	15	1.54	0.81	1.47	0.84	1.59

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
500	5	15	1.54	0.81	1.48	0.84	1.59
500	7	15	1.54	0.82	1.48	0.85	1.60
200	7	15	1.54	0.81	1.47	0.84	1.59
500	3	20	1.54	0.76	1.43	0.82	1.59
300	7	15	1.54	0.81	1.48	0.84	1.59
300	5	15	1.54	0.81	1.49	0.84	1.60
400	3	20	1.55	0.76	1.44	0.82	1.60
200	5	15	1.55	0.81	1.49	0.84	1.59
400	5	15	1.55	0.81	1.49	0.84	1.59
100	7	15	1.55	0.82	1.48	0.85	1.59
100	5	15	1.55	0.81	1.49	0.84	1.60
100	3	20	1.56	0.77	1.44	0.82	1.60
400	3	15	1.59	0.83	1.53	0.85	1.64
500	3	15	1.59	0.83	1.54	0.85	1.65
300	3	15	1.60	0.83	1.54	0.85	1.64
200	3	15	1.60	0.83	1.54	0.85	1.65
100	3	15	1.61	0.83	1.54	0.85	1.65
500	5	10	1.67	0.92	1.67	0.93	1.72
400	5	10	1.67	0.92	1.67	0.93	1.72
200	5	10	1.67	0.92	1.67	0.93	1.72
300	5	10	1.67	0.92	1.67	0.93	1.72
100	5	10	1.68	0.92	1.67	0.93	1.73
500	7	10	1.68	0.93	1.67	0.94	1.73
400	7	10	1.68	0.93	1.68	0.94	1.74
300	7	10	1.68	0.93	1.68	0.94	1.74
200	7	10	1.68	0.92	1.67	0.93	1.73
100	7	10	1.69	0.92	1.67	0.93	1.73
100	3	10	1.72	0.96	1.70	0.96	1.76
200	3	10	1.72	0.95	1.71	0.95	1.76
500	3	10	1.72	0.95	1.71	0.96	1.77
400	3	10	1.72	0.95	1.71	0.96	1.77
300	3	10	1.72	0.95	1.71	0.96	1.77
400	5	5	1.98	1.19	1.98	1.19	2.00
200	5	5	1.98	1.18	1.98	1.18	1.99
300	5	5	1.98	1.18	1.98	1.18	1.99
500	5	5	1.98	1.19	1.98	1.19	2.00
100	5	5	1.98	1.19	1.98	1.19	2.00
200	7	5	2.03	1.25	2.04	1.25	2.07
500	7	5	2.03	1.24	2.04	1.24	2.07
400	7	5	2.03	1.24	2.04	1.24	2.06
300	7	5	2.03	1.24	2.05	1.24	2.07
100	7	5	2.03	1.24	2.04	1.24	2.07
200	3	5	2.15	1.32	2.15	1.32	2.20
300	3	5	2.16	1.32	2.16	1.33	2.22
100	3	5	2.17	1.31	2.16	1.32	2.21
500	3	5	2.17	1.33	2.18	1.34	2.24
400	3	5	2.17	1.33	2.18	1.34	2.23
300	1	30	2.40	1.39	2.36	1.40	2.49
300	1	15	2.42	1.41	2.38	1.43	2.51
300	1	20	2.42	1.39	2.38	1.41	2.51
200	1	15	2.43	1.41	2.39	1.43	2.52
400	1	20	2.43	1.41	2.40	1.43	2.53
400	1	30	2.44	1.41	2.41	1.43	2.54
500	1	30	2.44	1.41	2.41	1.43	2.54
200	1	25	2.44	1.41	2.40	1.43	2.53
500	1	25	2.44	1.41	2.41	1.43	2.55
400	1	15	2.45	1.43	2.42	1.44	2.54
500	1	15	2.46	1.43	2.43	1.44	2.55
400	1	25	2.46	1.41	2.42	1.43	2.56
200	1	20	2.46	1.43	2.43	1.45	2.57
300	1	25	2.47	1.41	2.43	1.42	2.57
500	1	20	2.47	1.42	2.44	1.44	2.57
100	1	20	2.49	1.43	2.44	1.45	2.57
100	1	30	2.49	1.45	2.45	1.47	2.59
100	1	25	2.49	1.45	2.44	1.46	2.58
100	1	15	2.50	1.46	2.46	1.47	2.59
200	1	30	2.50	1.44	2.46	1.45	2.59
200	1	10	2.54	1.50	2.51	1.51	2.62

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
300	1	10	2.57	1.50	2.55	1.51	2.66
500	1	10	2.60	1.52	2.58	1.53	2.70
100	1	10	2.60	1.54	2.58	1.56	2.70
400	1	10	2.61	1.52	2.58	1.53	2.70
200	1	5	2.95	1.77	2.95	1.78	3.07
300	1	5	2.97	1.77	2.97	1.78	3.09
400	1	5	3.00	1.79	3.00	1.80	3.12
500	1	5	3.01	1.79	3.01	1.80	3.14
100	1	5	3.02	1.81	3.01	1.82	3.12

Appendix 6. MAE, RMSE and OBB values of RF (NoW-NoT) model for training and testing samples

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
400	6	30	1.62	0.73	1.43	0.84	1.66
500	6	25	1.62	0.74	1.44	0.84	1.67
200	6	30	1.62	0.72	1.43	0.83	1.66
500	6	30	1.62	0.73	1.43	0.84	1.67
300	6	30	1.62	0.73	1.43	0.84	1.66
400	6	25	1.62	0.74	1.44	0.84	1.67
300	6	25	1.62	0.74	1.44	0.84	1.67
200	6	25	1.62	0.74	1.45	0.84	1.67
100	6	30	1.62	0.72	1.43	0.83	1.66
100	6	25	1.63	0.74	1.44	0.84	1.66
500	6	20	1.63	0.77	1.49	0.85	1.67
400	6	20	1.63	0.77	1.49	0.85	1.68
500	4	30	1.63	0.74	1.46	0.83	1.67
300	4	30	1.63	0.74	1.46	0.84	1.67
500	4	25	1.63	0.74	1.46	0.84	1.67
400	4	30	1.63	0.74	1.46	0.84	1.67
400	4	25	1.63	0.74	1.47	0.84	1.67
300	6	20	1.63	0.78	1.49	0.85	1.68
200	6	20	1.63	0.78	1.49	0.85	1.68
300	4	25	1.63	0.75	1.47	0.84	1.67
200	4	25	1.63	0.75	1.47	0.84	1.67
200	4	30	1.64	0.74	1.46	0.84	1.68
400	4	20	1.64	0.77	1.51	0.84	1.68
300	4	20	1.64	0.77	1.51	0.84	1.68
500	4	20	1.64	0.77	1.51	0.84	1.68
100	4	25	1.64	0.75	1.47	0.84	1.68
100	6	20	1.64	0.78	1.50	0.85	1.68
200	4	20	1.64	0.78	1.51	0.84	1.68
100	4	30	1.65	0.74	1.47	0.84	1.68
100	4	20	1.65	0.78	1.52	0.85	1.69
500	6	15	1.68	0.85	1.62	0.88	1.73
300	6	15	1.68	0.85	1.62	0.88	1.73
400	6	15	1.68	0.85	1.62	0.88	1.73
100	6	15	1.68	0.85	1.62	0.88	1.73
200	6	15	1.69	0.85	1.63	0.88	1.73
400	4	15	1.69	0.85	1.63	0.87	1.73
500	4	15	1.69	0.85	1.64	0.87	1.73
300	4	15	1.69	0.85	1.63	0.87	1.73
100	4	15	1.69	0.85	1.63	0.87	1.73
200	4	15	1.69	0.85	1.64	0.87	1.73
500	2	30	1.71	0.82	1.60	0.88	1.77
300	2	30	1.72	0.82	1.60	0.88	1.77
400	2	25	1.72	0.82	1.60	0.88	1.77
400	2	30	1.72	0.82	1.60	0.88	1.78
200	2	25	1.72	0.82	1.60	0.88	1.77
500	2	25	1.72	0.82	1.60	0.88	1.78
200	2	30	1.72	0.82	1.60	0.88	1.78
300	2	25	1.72	0.82	1.60	0.88	1.78
500	2	20	1.73	0.84	1.63	0.88	1.78
400	2	20	1.73	0.84	1.63	0.89	1.78

(*continued on next page*)

(*continued*)

Tree num	Try num	Max depth	OBB	Training		Testing	
				MAE	RMSE	MAE	RMSE
300	2	20	1.73	0.84	1.63	0.88	1.78
100	2	30	1.73	0.82	1.60	0.88	1.77
200	2	20	1.73	0.84	1.63	0.89	1.78
100	2	25	1.73	0.82	1.60	0.88	1.77
100	2	20	1.74	0.84	1.63	0.89	1.78
500	2	15	1.76	0.89	1.72	0.92	1.82
400	2	15	1.77	0.89	1.72	0.92	1.82
200	2	15	1.77	0.89	1.72	0.92	1.82
300	2	15	1.77	0.90	1.72	0.92	1.83
100	2	15	1.78	0.89	1.72	0.92	1.82
300	4	10	1.79	0.95	1.78	0.95	1.83
500	4	10	1.79	0.95	1.78	0.95	1.83
500	6	10	1.79	0.96	1.79	0.96	1.83
400	6	10	1.79	0.96	1.79	0.96	1.83
400	4	10	1.79	0.95	1.79	0.95	1.83
300	6	10	1.79	0.96	1.79	0.96	1.83
200	6	10	1.79	0.96	1.79	0.96	1.84
100	6	10	1.79	0.96	1.79	0.96	1.84
200	4	10	1.79	0.95	1.79	0.95	1.83
100	4	10	1.80	0.95	1.79	0.95	1.83
500	2	10	1.89	1.04	1.88	1.05	1.95
300	2	10	1.90	1.04	1.89	1.05	1.95
400	2	10	1.90	1.04	1.89	1.05	1.95
100	2	10	1.90	1.04	1.88	1.05	1.95
200	2	10	1.91	1.05	1.89	1.06	1.96
500	4	5	2.04	1.20	2.05	1.20	2.07
400	4	5	2.04	1.20	2.05	1.20	2.07
300	4	5	2.04	1.20	2.05	1.20	2.07
200	4	5	2.05	1.20	2.05	1.20	2.07
100	4	5	2.05	1.20	2.05	1.20	2.08
200	6	5	2.06	1.24	2.08	1.24	2.10
400	6	5	2.07	1.24	2.09	1.24	2.10
300	6	5	2.07	1.24	2.09	1.24	2.11
100	6	5	2.07	1.25	2.09	1.25	2.10
500	6	5	2.07	1.24	2.09	1.24	2.11
100	2	5	2.38	1.43	2.36	1.44	2.44
500	2	5	2.39	1.46	2.40	1.46	2.47
300	2	5	2.40	1.46	2.40	1.47	2.47
400	2	5	2.40	1.46	2.40	1.47	2.48
200	2	5	2.41	1.47	2.41	1.48	2.48

References

Hauck, F., Kliewer, N., 2019. Data analytics in railway operations: using machine learning to predict train delays. In: Operations Research Proceedings 2019: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR). Springer, Dresden, Germany, pp. 741–747.

Huang, P., Wen, C., Fu, L., Lessan, J., Jiang, C., Peng, Q., Xu, X., 2020. Modeling train operation as sequences: a study of delay prediction with operation and weather data. Transport. Res. E Logist. Transport. Rev. 141, 102022.

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.

Ji, Y., Zheng, W., Dong, H., Gao, P., 2020. Train delays prediction based on feature selection and random forest. In: IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 1–6.

Kecman, P., Goverde, R., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transp 7 (3), 295–319.

Laifa, H., Ghezalaa, H.H.B., 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 Journal of Intelligent Transportation Systems 3, 539–550.

Li, Z., Huang, P., Wen, C., Jiang, X., Rodrigues, F., 2022. Prediction of train arrival delays considering route conflicts at multi-line stations. Transport. Res. C Emerg. Technol. 138, 103606.

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: IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 371–380.

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. Transport. Res. Rec. 2676, 353–365.

Nair, R., Hoang, T.L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., Walter, T., 2019. An ensemble prediction model for train delays. Transport. Res. C Emerg. Technol. 104, 196–209.

Neunhoeffer, M., Sternberg, S., 2019. How cross-validation can go wrong and what to do about it. Polit. Anal. 27, 101–106.

Økland, A., Olsson, N.O., 2021. Punctuality development and delay explanation factors on Norwegian railways in the period 2005–2014. Public Transport 13, 127–161.

Olsson, N.O., Haugland, H., 2004. Influencing factors on train punctuality—results from some Norwegian studies. Transport Pol. 11, 387–397.

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: 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 Transactions on Systems, Man, and Cybernetics: Systems 47, 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 research 11, 54–64.

Palmer, P.B., O'Connell, D.G., 2009. Regression analysis for prediction: understanding the process. Cardiopulm. Phys. Ther. J. 20, 23.

Peters, J., Emig, B., Jung, M., Schmidt, S., 2005. Prediction of delays in public transportation using neural networks. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). 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: 11th International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, pp. 307–312.

Spanninger, T., Trivella, A., Büchel, B., Corman, F., 2022. A review of train delay prediction approaches. Journal of Rail Transport Planning & Management 22, 100312.

Taleongpong, P., Hu, S., Jiang, Z., Wu, C., Popo-Ola, S., Han, K., 2022. Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network. Journal of Intelligent Transportation Systems 26, 311–329.

Tiong, K., Ma, Z., Palmqvist, C.-W., 2022. Real-time train arrival time prediction at multiple stations and arbitrary times. In: IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, pp. 793–798.

Tiong, K.Y., Ma, Z., Palmqvist, C.-W., 2023a. A review of data-driven approaches to predict train delays. Transport. Res. C Emerg. Technol. 148, 104027.

Tiong, K.Y., Ma, Z., Palmqvist, C.-W., 2023b. Analyzing Factors Contributing to Real-Time Train Arrival Delays Using Seemingly Unrelated Regression Models.

Xia, Y., Van Ommeren, J.N., Rietveld, P., Verhagen, W., 2013. Railway infrastructure disturbances and train operator performance: the role of weather. Transport. Res. Transport Environ. 18, 97–102.