International Journal of Computational Science and Engineering.
ISSN 2249-4251 Volume 13, Number 1 (2023), pp. 35-44

https://dx.doi.org/10.37622/IJCSE/13.1.2023.35-44

Train Delay Predicti On Using Machine Learning and Deep Learning Techniques

Dr. O. Obulesu*, A. Shivani Reddy, D. Ashritha Reddy, Dondeti Pavani and Shreya Reddy S

Department of CSE,
G. Narayanamma Institute of Technology & Science, Hyderabad, Telangana, India
**email: obulesh194@gmail.com*

Abstract:

Transport systems are critical pieces of infrastructure and they have substantially increased in size in many countries worldwide. This includes rail transport systems that have evolved significantly, including to provide long-distance travel services. Passenger train delay significantly influences riders' decision to choose rail transport as their mode choice. Poor on-time performance can impact passenger trust and their satisfaction, and it may result in a shift to other modes of transport, especially private vehicles and air transport. Service disruption is a root cause of lower rail punctuality and customer satisfaction. Major service disruptions result from various conditions or factors such as accidents, problems in train operation, malfunctioning or damaged equipment, routine maintenance, construction, passenger boarding or alighting, and even extreme weather conditions. Train delay can also negatively affect connecting trains and passengers' journeys or activities. Thus, delay estimates or predictions can help train operators develop better plans to manage, reschedule, or adjust the timetable of the current and consecutive trains more effectively, as well as to inform passengers in advance so they themselves can adjust their travel plans in time. In light of these problems, the main objective is to model passenger train delay prediction based on three Machine Learning and Deep Learning techniques.

Keywords: CNN, Train Delay Prediction, Random Forest, Logistic Regression, KNN, Naive Bayes

Introduction

Our aim is to build a web application on Train delay prediction to predict the influence degree of train operation interference and delay propagation, which is helpful to

realizing real-time risk analysis and early warning of dispatching.Transport systems are critical pieces of infrastructure and they have substantially increased in size in many countries worldwide. This includes rail transport systems that have evolved significantly, including to provide long-distance travel services. Passenger train delay significantly influences riders' decision to choose rail transport as their mode choice. Poor on-time performance can impact passenger trust and their satisfaction, and it may result in a shift to other modes of transport, especially private vehicles and air transport. Service disruption is a root cause of lower rail punctuality and customer satisfaction. Major service disruptions result from various conditions or factors such as accidents, problems in train operation, malfunctioning. It has happened so many times that you have been waiting at the railway station for someone to arrive and you don't have any exact information about train timing and other stuff. So here we present to you a project on Railway Tracking and Arrival Time Prediction. Using this system users can get the information about train timing, and is it on time or not, and other information.

In this, the system will track the train timing at what time the train departed from a particular station and pass these timing details to another station's system where it will display the timing according to the train departed from the previous station. If the system finds any delay in the train due to a signal it will automatically update the train timing in the next station and will be displayed to viewers. In this system there is an admin module, who enters the details about trains and its timing and these details will be passed through an internet server and is fetched by the system on other stations, and there is another system that shows train information to the viewers on platform.

Second system will get all the information of all trains but will automatically select the data that refers to a particular station and show that information on screen. Station masters on every station have a login wherein they may update train arrival time at their station when it arrives. This second System is installed on various locations on stations for viewers to view the information. Admin will add information like train departed from station, expected arrival at destination, delay in the train schedule, etc. 2 This project publishes real-time train schedule events to subscribing multiple client applications. In the present world, the major components of any transportation system include passenger Railway, cargo Railway, and air traffic control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the Railway transportation system. This has brought drastic change in the Railway operations. Train delays occasionally cause inconvenience to the modern passengers. Every year approximately 20% of Railway Trains are canceled or delayed, costing passengers more than 20 billion dollars in money and their time.

Related Works

In the area of Intelligent Transportation System (ITS) many pieces of research have been developed in the past. Following Literature review shows a few researches on train delays prediction systems that have been performed. Pipatphon Lapapinyo et al., [proposed a passenger train delay prediction model using random forest (RF), gradient boosting machine (GBM), and multi-layer perceptron (MLP). In this article, the impact on the PTPD models using Real-time based Data-frame Structure (RTDFS) and Real-

time with Historical based Data-frame Structure (RWH-DFS) is investigated. The results show that PTDP models using MLP with RWH-DFS outperformed all other models. The influence of the external variables such as historical delay profiles at the destination (HDPD), ridership, population, day of the week, geography, and weather information on the real-time PTPD models are also further analyzed and discussed.

A. Hansen et al., [4] proposed a model in which two things are taken into consideration, dispatching decisions & conflicts of train path. The model is built for predicting running times as well as arrival times for delayed and on-time, both kinds of trains and to evaluate the effectiveness of dispatching decisions.

Masoud Yaghini et al., [6] presented an ANN model with high accuracy to predict the train delays. He proposed a model that uses different strategies to define the input. To evaluate the quality of the results, they took advantage of Multinomial logistic regression models & decision trees. The model accuracy and training time may be improved over met heuristic methods such as genetic algorithms.

S. Pongnumkul et al., [3] proposed two algorithms using the average of historical travel times and average travel time of the k-nearest neighbors of the last known arrival time respectively. In this paper, a review found that both the algorithms bring in similar percent improvement in the prediction errors. There is some work to do on Thai railways for predicting train delays. They also discovered that one of the major factors for train delay is the number of stops between stations. In this report, they used 9 parameters & six months of train data to predict the inline delay.

Jia Hu et al, [5] proposed a prediction model that is built on the basis of Artificial Neural Network (ANN). To overcome the endogenous drawback of ANN, further Genetic Algorithm is implemented to boost the performance of ANN. A thorough survey of the following papers was done to understand thoroughly the concept of Train delay prediction systems, datasets, its attributes and methodology used: and for related research work for flight delay prediction systems. Delay in a train means that the train has not arrived at its pre scheduled time. The train delay does not include unexpected stopping time near to the station or in between the station due to poor signal or unavailability of the platform.

Jianqing Wu et al, [1] has proposed a hybrid deep learning solution by integrating long short-term memory (LSTM) and Critical Point Search (CPS). LSTM deals with long term prediction tasks of trains' running time and dwell time, while CPS uses predicted values with a nominal timetable to identify primary and secondary delays based on the delay causes, run-time delay, and dwell time delay. To validate the model and analyze its performance, we compare the standard LSTM with the proposed hybrid model. The results demonstrate that new variants outperform the standard LSTM, based on predicting time steps of dwell time feature. The experiment results also showed many irregularities of historical trends, which draws attention for further research.

Sybil Derrible et al., [2] proposed two algorithms using the average of historical travel times and average travel time using the support vector machines of the last known arrival time respectively. In this paper, a review found that both the algorithms bring in similar percent improvement in the prediction errors. There is some work to do on China railways for predicting train delays. They also discovered that one of the major factors for train delay is the number of stops between stations.

There are some important causes that leads to train delays like delay at the origin, engine breakdown, other train's engine breakdown, waiting time at overtaking point, Climate/weather condition (temperature, wind speed, heavy rain, snowfall) and other factors (railways assets condition, festivals, strikes, national level exams, etc.). In this paper, we included most of these factors in order to predict better results. Accuracy of the model can be improved through meta-heuristic methods like genetic algorithm, hybrid methods or ensemble learning.

Proposed Methodology

The main aim of the project is to build a website to predict the expected delay in the arrival of the train. Convolutional Neural Networks have been proven to be so effective for prediction problems. A convolutional neural network, a class of artificial neural networks, is a network architecture for deep learning that learns directly from the data. Neural networks are made up of layers where each layer is connected to other layers, thus forming a network-like structure.

A typical network model consists of multiple layers, each comprising several nodes. The layers in the neural networks can be classified as an input layer, one or more hidden layers and an output layer(fully-connected layer). Convolutional neural networks consist of three main types of layers, namely convolutional layers, pooling layers and fully-connected layers. The complexity of a convolutional neural network is directly proportional to the number of layers. Every node in the network is associated with a threshold value and a weight. A node is said to be activated if the output of the node is greater than the specified threshold, and forwards the input to the next node.

Architecture of Convolutional Neural Network

The general architecture of a CNN consists of the following components:

- Input layer: Input layer represents the first layer of the CNN which brings the information into the network for further processing of the data. It is the first convolutional layer responsible for taking the inputs.
- Convolutional layers: Convolutional layers are responsible to perform the core computation tasks, thus are the building blocks of CNN. These layers extract features from the input data by applying a set of filters to the data.
- Activation function: the output of the convolution operation from the convolutional layer is given as input to an activation function. This activation function gives a non-linear expression to the input. The output of the activation function is used to decide the firing i.e activation of the neuron(node).
- Pooling layers: These layers are known as downsampling layers, in other words they downsample the output of the convolutional layers, by conducting dimensionality reduction and reducing the input parameters.

This helps to reduce the number of parameters in the model, reducing overfitting and improving computational skills. Pooling layers computation is similar to convolutional layers, but the neurons in these layers are not associated with weights.

- Fully connected layers: The function of fully connected layers is to ensure that every input vector has an influence on the output factor; they have weights connected to all outputs of the previous layer. Fully connected layers are applied to the output from pooling layers along with a set of weights to make the predictions.
- Output layer: It is the final layer of the CNN responsible for producing the final output. It represents the desired predictions based on the input data.

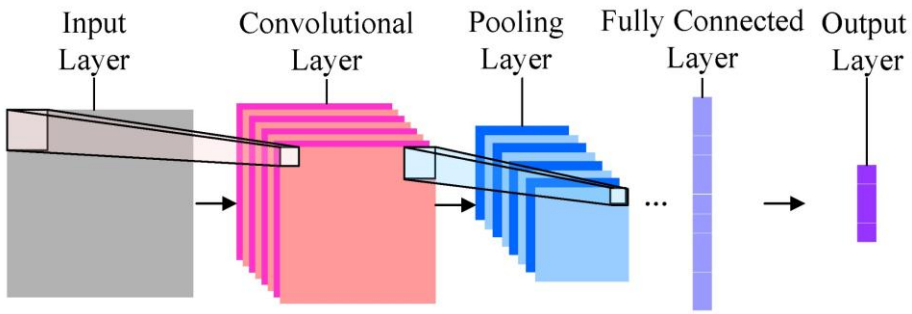
The use of convolutional layers, pooling layers and fully-connected layers enable Convolutional Neural Networks to learn increasingly complex features of the input data, allowing it to make accurate predictions.

Working of CNN

Convolutional neural network (CNN) for train delay prediction, primarily started with the collection of historical data that includes train schedules, routes, stations, weather, and other factors that affect train delays. As the next step, we preprocess the data by collecting, formatting and cleaning it, this also includes removing duplicate data, handling missing values.

The pre-processed data must be trained for the model to predict the expected delay of the train. The data must be split before training into a separate set for training, validation, and testing. The training set will be used to train the CNN model, while the validation set will be used to tune the hyperparameters of the model, and the testing set will be used to evaluate the performance of the trained model. In general, 80 percent of the data is used for testing, 10 percent data each is used for validating and testing.

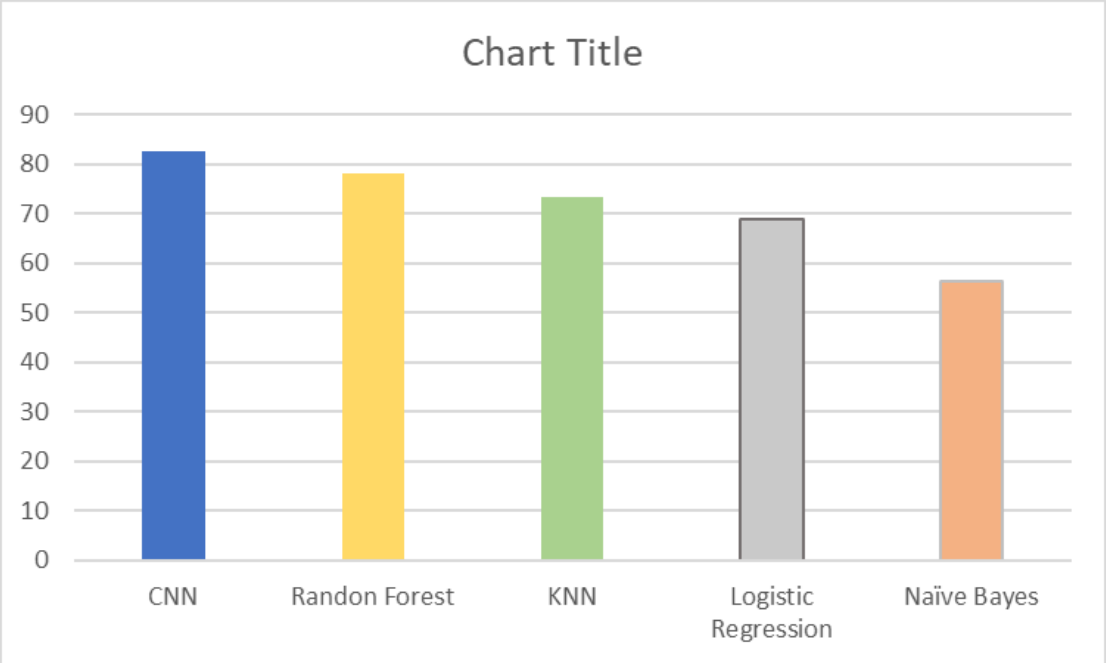
The CNN architecture consists of several layers of convolutional and pooling layers, followed by one or more fully connected layers. The convolutional layers will help the model to learn and extract the features of the input data, while the pooling layers will help reduce the dimensionality of the data, and handle overfitting. Next, we compile the model by specifying the loss function, optimizer, and metrics that will be used during training, here three convolutional layers are used with ReLU activation function. ReLU is Rectified Linear Unit, it introduces a non-linearity to the model improving the computational performance. While training the model for train delay prediction, accuracy is used as a metric along with loss functions like mean squared error (MSE), root mean squared error (RMSE) to evaluate the performance of the model.



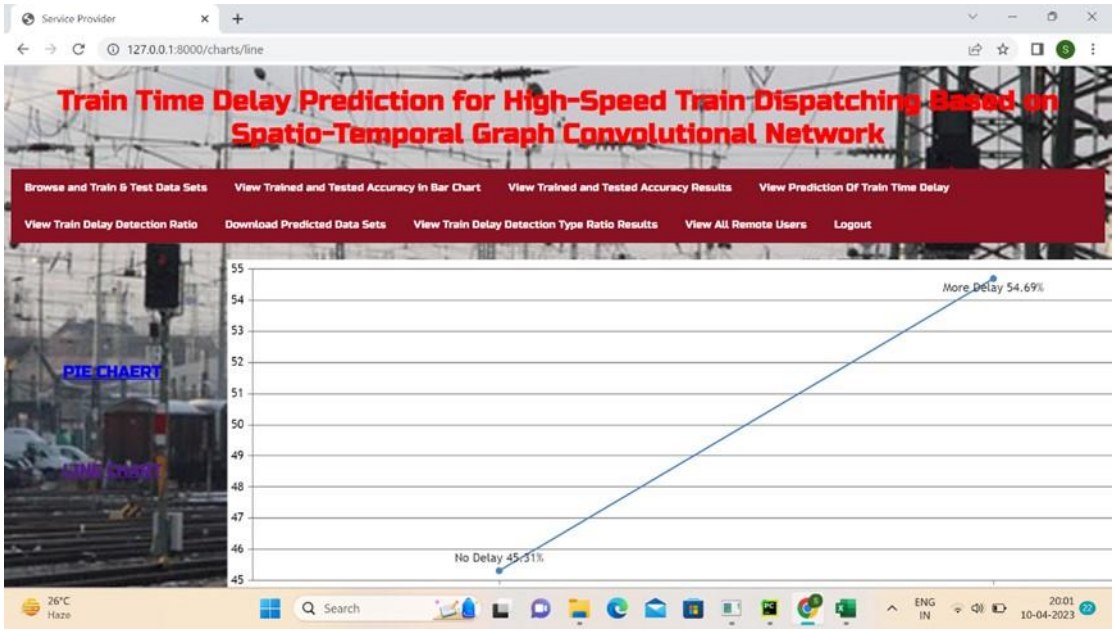
The model is trained using the training data, using a suitable batch size and number of epochs. The performance of the model is monitored on the validation set during training to prevent overfitting. The performance of the trained model is evaluated on the testing set, using metrics such as MAE or RMSE. The model is fine-tuned by adjusting the hyperparameters, adding or removing layers, or using a different architecture.
Overall, the process of implementing a CNN for train delay prediction involves data collection and preprocessing, model architecture design, model training and evaluation, and deployment.

Results and Discussion

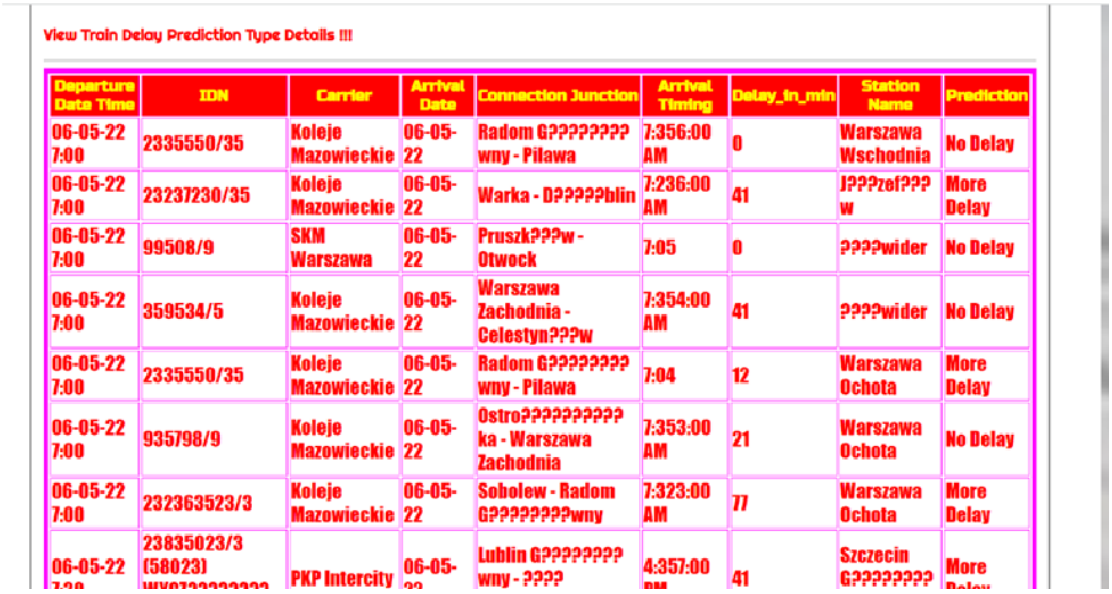
This section discusses the results given by Convolutional Neural Network in comparison with Random Forest, Linear Regression, K-Nearest Neighbour, and Naive Bayes algorithms. This research project aimed to develop a website to predict the expected delay in the arrival of trains.



During the experimental comparative analysis, it has been observed the highest accuracy is observed for CNN, which is followed by Random forest with an accuracy of 71 percent. KNN algorithm with 68 percent accuracy has performed better than Logistic Regression and Naive Bayes.



The figure represents the analysis of results queried by the users. It gives a percentage overview of the result in terms of predicted delay of the train. From the figure it is observed that 54.6 percent of the queries resulted in a delay in the train arrival while 45.4 percent of the results mentioned that there wouldn't be any delay in the arrival of the train.



View Train Delay Prediction Type Details !!!

Departure Date Time	IDN	Carrier	Arrival Date	Connection Junction	Arrival Timing	Delay_in_min	Station Name	Prediction
06-05-22 7:00	2335550/35	Koleje Mazowieckie	06-05-22	Radom G???????? wny - Pilawa	7:356:00 AM	0	Warszawa Wschodnia	No Delay
06-05-22 7:00	23237230/35	Koleje Mazowieckie	06-05-22	Warka - D?????blin	7:236:00 AM	41	J???zef??? w	More Delay
06-05-22 7:00	99508/9	SKM Warszawa	06-05-22	Pruszk???w - Otwock	7:05	0	????wider	No Delay
06-05-22 7:00	359534/5	Koleje Mazowieckie	06-05-22	Warszawa Zachodnia - Celestyn???w	7:354:00 AM	41	????wider	No Delay
06-05-22 7:00	2335550/35	Koleje Mazowieckie	06-05-22	Radom G???????? wny - Pilawa	7:04	12	Warszawa Ochota	More Delay
06-05-22 7:00	935798/9	Koleje Mazowieckie	06-05-22	Ostro?????????? ka - Warszawa Zachodnia	7:353:00 AM	21	Warszawa Ochota	No Delay
06-05-22 7:00	232363523/3	Koleje Mazowieckie	06-05-22	Sobolew - Radom G????????wny	7:323:00 AM	77	Warszawa Ochota	More Delay
06-05-22	23835023/3 (58023)	PKP Intercity	06-05-	Lublin G???????? wny - ????	4:357:00	41	Szczecin G????????	More

The above figure represents the statistics of the predicted delays which were queried by the user. An additional option to download the details is also provided via the UI.

Conclusion

Based on the research conducted on train delay prediction using convolution neural networks, it can be concluded that this approach can be effective in predicting train delays with a high degree of accuracy. The use of convolutional neural networks allows for the detection of patterns and trends in large sets of data, making it possible to identify factors that contribute to train delays. The use of features such as demand data, passenger count, and historical train performance can also enhance the accuracy of the model.

However, it is important to note that the accuracy of the model may be influenced by the quality of the data used for training and testing, as well as the availability of relevant features. In the experimental stage, we compare our CNN with RF, Naive Bayes, Logistic Regression and KNN. The experimental results show that CNN is clearly better for the train delay cumulative effect prediction for train dispatching. Overall, convolution neural networks can be a useful tool for predicting train delays and improving the efficiency and reliability of train operations. Further research can be conducted to explore the potential of this approach in other transportation systems and to identify ways to optimize the model's performance.

Future Scope

Some of the areas which can result in improvement of the model are

- **Advanced analytics:** Advanced analytics tools such as predictive modeling and data visualization can help railway authorities to identify the factors that contribute to train delays and take appropriate actions to prevent them in the future.
- **Internet of Things (IoT):** IoT devices such as sensors and cameras can be used to collect data on train movements and track conditions, which can be used to predict delays more accurately.
- **Artificial Intelligence (AI):** The use of AI technologies such as natural language processing and image recognition can help improve the accuracy of train delay predictions by analyzing unstructured data such as social media posts and news articles.

Overall, the future of train delay prediction is promising, and with the right technology and tools, railway authorities can provide a better and more reliable service.

References

[1] Jianqing Wu et al., “A GTFS data acquisition and processing framework and its application to train delay prediction”, doi:10.1016/j.ijtst.2022.01.005, March 2022.

[2] Sybil Derrible et al., “Real-Time Passenger Train Delay Prediction Using Machine Learning: A Case Study With Amtrak Passenger Train Routes”, doi:https://doi.org/10.1109/OJITS.2022.319 4879, Vol. 3 pp. 539 – 550, Jan 2022

[3] S. Pongumkal et al., “ A review of train delay prediction approaches.”, 2022.

[4] A. Hansen et al., "Prediction of Train Delay in Indian Railways through Machine Learning Techniques.", Department of CS & IT, Maulana Azad National Urdu University, Hyderabad, 2019.
[5] Jia Hu et al., "Transit signal priority accommodating conflicting requests under Connected Vehicles technology", doi: 10.1016/j.trc.2016.06.001, Aug 2016.
[6] Masoud Yaghini et al., "Railway passenger train delay prediction via neural network model.", Masoud Yaghini, Mohammad M.Koshraftar and Masoud Seyedabadi, 2012.
[7] L. Oneto et al., "Train delay prediction systems: A big data analytics perspective, " Big Data Res., vol. 11, pp. 54–64, Mar. 2018, doi: 10.1016/j.bdr.2017.05.002.
[8] Z. Alwadood, A. Shuib, and N. A. Hamid, "Rail passenger service delays: An overview, " in Proc. IEEE Bus. Eng. Ind. Appl. Colloquium (BEIAC), Apr. 2012, pp. 449–454, doi: 10.1109/BEIAC.2012.6226102.
[9] S. Milinkovic, M. Markovic, S. Veskovi ´ c, M. Ivi ´ c, and ´ N. Pavlovic, "A fuzzy Petri net model to estimate train delays, " Simulat. Model. Pract. Theory, vol. 33, pp. 144–157, Apr. 2013.
[10] B. W. Schlake, C. P. L. Barkan, and J. R. Edwards, "Train delay and economic impact of in-service failures of railroad rolling stock, " Transport. Res. Rec., vol. 2261, no. 1, pp. 124– 133, Jan. 2011, doi: 10.3141/2261-14.
[11] R. Wang and D. B. Work, "Data driven approaches for passenger train delay estimation, " in Proc. IEEE 18th Int. Conf. Intell. Transport. Syst., Sep. 2015, pp. 535–540, doi: 10.1109/ITSC.2015.94.
[12] L. Oneto et al., "Advanced analytics for train delay prediction systems by including exogenous weather data, " in Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2016, pp. 458–467, doi: 10.1109/DSAA.2016.57.
[13] A. Gal, A. Mandelbaum, F. Schnitzler, A. Senderovich, and M. Weidlich, "Traveling time prediction in scheduled transportation with journey segments, " Inf. Syst., vol. 64, pp. 266–280, Mar. 2017, doi: 10.1016/j.is.2015.12.001.
[14] A. Estes, M. O. Ball, and D. Lovell, "Predicting performance of ground delay programs, " presented at the 12th USA/Europe air traffic management R&D seminar, Seattle, WA, USA, 2017.
[15] R. Gaurav and B. Srivastava, "Estimating train delays in a large rail network using a zero shot Markov model, " Jun. 2018, arXiv:1806.02825.
[16] R. Nair et al., "An ensemble prediction model for train delays, " Transport. Res. C, Emerg. Technol., vol. 104, pp. 196–209, Jul. 2019, doi: 10.1016/j.trc.2019.04.026.
[17] P. Taleongpong, S. Hu, Z. Jiang, C. Wu, S. Popo-Ola, and K. Han, "Machine learning techniques to predict reactionary delays and other associated key performance indicators on the British railway network, " J. Intell. Transport. Syst., vol. 26, no. 3, pp. 311–329, Dec.2020, doi:10.1080/15472450.2020.1858 822.2018.
[18] S. Jing, "Research on delay prediction of high speed railway train based on data Analysis, " Ph.D dissertation, Southwest Jiaotong UnivChengdu, China, 2019.
[19] R. Nilsson and K. Henning. Predictions of Train Delays Using Machine Learning. ht. kb.se/reslve?urn=urn:nbn:se:kth:diva-2347,

[20] Amtrak Five Year Service Line Plans FY20-24, Amtrak, Washington, DC, USA, 2019.
[21] N. O. E. Olsson and H. Haugland, "Influencing factors on train punctuality — Results from some Norwegian studies, " Transp. Policy, vol. 11, no. 4, pp. 387–397, Oct. 2004, doi: 10.1016/j.tranpol.2004.07.001.
[22] W. Peetawan and K. Suthiwart Narueput, "Identifying factors affecting the success of rail infrastructure development projects contributing to a logistics platform: A Thailand case study, " Kasetsart J. Social Sci., vol. 39, no. 2, pp. 320–327, 2018, doi: 10.1016/j.kjss.2018.05.002.
[23] P. Wang and Q. Zhang, "Train delay analysis and prediction based on big data fusion."