International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

Predictive Analytics for Passenger Train Arrival Time Estimation

Chethan S
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
chethansmca2025@gmail.com

Ashok B P
Assistant Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
ashokbp.mca@gmail.com

ABSTRACT

As people have increasingly demanded punctuality, real-time updates and effective working timetables of the transport system, the need to establish intelligent transport systems has risen tremendously in the past years. The accurate prediction of the arrival of passenger trains is critical in the enhancement of user satisfactions, the elimination of waiting times as well as the overall performance of reliability of the system. The conventional fixed schedules do not model dynamic processes like weather, track conditions, operational delays, and traffic and therefore cannot be effective in practice.

In order to take care of this, our work looks into the use of predictive analytics using historical and real-time data. Compared with the traditional methods, data-driven models can adapt to dynamics of changing conditions and can, therefore, be more durable in their predictions and identification of trends in time-series data. We consider specifically advanced machine learning and deep learning algorithmic techniques, especially the Long Short-Term Memory (LSTM) perceptron, and can learn sequential information .

Keywords: Train arrival time prediction, LSTM autoencoder, time-series forecasting, smart transportation, deep learning, machine learning, real-time train data, predictive analytics, RMSE, MAE, SMAPE.

INTRODUCTION

Movement of people and products is one of the critical elements in the growth of a country and it has grown everywhere in the globe. These include the trains which have undergone tremendous improvement so that people can travel long distances. The number of people and goods transported in the rail has increased by nearly eight percent in Norway between the year 2013 and 2016. In the United States, the ratio of the individuals requiring the trains paid by the government reached above 10 percent and this has demonstrated the highest increase in the trains. There was a larger number of passengers and an increase in revenue of 6.2 and 7.3 percent respectively on the long-distance trains in the year On time trains are the most important factors as they can sustain their capacity to become popular and win new clients. In one case, a common delay in a train, people will not prefer train travel, and they will not trust the trains thus taking their vehicles or flying. Predictions can now be made in a better way regarding the trains arriving late. In the railway system of Belgrade, Fuzzy Petri Net (FPN) was applied to analyze the

huge train delays. It can be utilised in tracking any movement of the train and fault detection in single/dual tracks. As according to Wang and Work, the planning techniques are useful in the planning of complicated train paths. The systems are very costly and time consuming to implement and develop to serve rail patterns that are challenging.

LITERATURE SURVEY

Oneto et al., 2019 Also in 2019, Oneto and his colleagues developed a Train Delay Prediction System (TDPS) which is based on big data to surpass the traditional models of delay forecast. As compared to relying on rule sets defined by train specialists and basic statistics, their solution is more data-driven because they do not only use what is present in train journey history records but fail to use statistical modeling. On data-based information will lead to efficient and effective data access of very large databases, so there is a possibility to provide even more precise delay predictions in long train runs. Yet in conducting the experiment specifically in the Italian rail network, the case was that, their TDPS was to a greater extent precise in reporting the train delays of the new and the new generation trains. It shows again how the big data contributes to the enhancement of the train delay management. 2.2 Bo Zhang, Dandan Ma, 2020 Bo Zhang and Dandan Ma developed Train Spatio-Temporal Graph Convolutional Network (TSTGCN) that is a more cutting-edge method to track the cumulative latencies of trains on the high-speed rail track in 2020. Unlike other approaches when an analysis is conducted regarding the delay of a single train, TSTGCN analyzes the number of delayed time at a particular stop. The most striking feature of their system is the way this mapping of interconnection of train stops is linked to the off-loading it experiences with the weeks and the days of failures. Their contribution is one of the best strengths, because they were able to correctly understand the level space between stops in the chart where other scientists did not succeed. In this manner they would be able to conceptualize the transmission of delays in a differentiated manner

EXISTING WORK

Trains are used extensively by people and the problem arises when they have to wait much longer than required due to delays. When we say here ways to see whether delays will occur in the foreseeable future, what we mean is methods of superseding the possible occasion of the delays in the future. We mention approaches such as RF, GBM, MLP. We consider two categories of data setups, how they perform in this test: Real-Time with Historical Data (RWH-DFS) and Just Real-Time Data (RT-DFS). One of the three used in checking is about how crowded a particular place is, the day, the location, the weather and whether the train was late at the last station. Nevertheless, despite the favorable aspects, the model in our discussion here performs more poorly in comparison with the leading ones, thereby it is less dependable practically speaking. Besides that, the model throws in a large dataset scenario with numerous individuals where it shows weakness in these situations. These difficulties indicate the

major limitations of the model's core and that we need to update it to manage large, complex train networks properly.

PROPOSED SYSTEM

The other airports excluding the trains are adversely affected due to train delay One of the solutions to remedy the delay situation in the new system is to utilize the purchase and access to a large amount of delay records. Data preparation is also a key process. This involves making corrections to incomplete data so as to prevent making bad guesses and converting group data into a machine-friendly form. The data is arranged after correcting it and train delays are estimated.

These are techniques to provide information on whether a train is running late, early or on schedule. In tests such as the one- right, exact and complete guesses go well. That implies that the system can do a good estimate on the train times

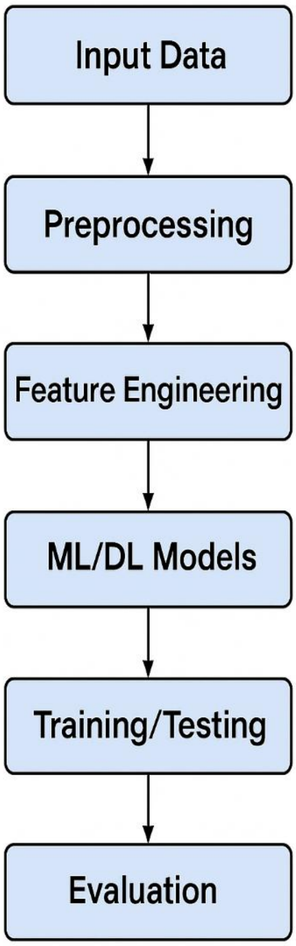
The new way ensures proper guesses of high accuracies within both observed learning models In addition, the model exemplifies clear illustrations such as graphs which makes it be easily comprehended on its efficiency. It can also process the data of many travelers without being inconvenient or inefficient. Greater so, the method minimises noise in the data and improves the overall accuracy of the predictive output by eliminating unnecessary and duplicative data via data cleaning procedures.

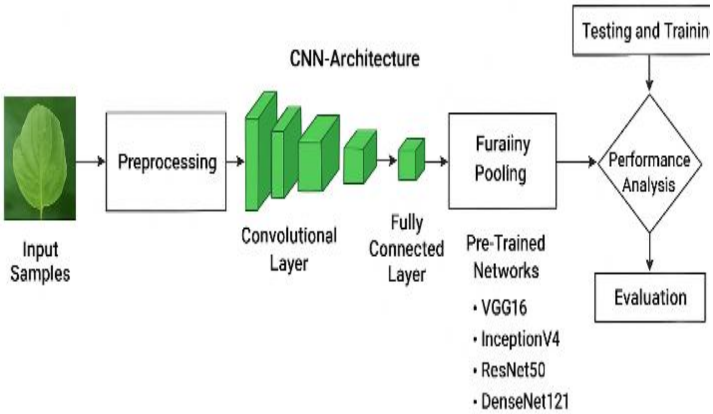
METHODOLOGY

The proposed project lies in the use application of deep learning in the identification and classification of the plant diseases respectively. Procedures also involve gathering the sample data sets that will incorporate an epicerebral utilisation of diseased and healthy leaves pictures (PlantVillage dataset). Raw images are also of different quality, so; preprocessing steps have also been integrated to normalise it. This includes resize, normalization and elimination of noise to make them uniform.

To provide the model with enhanced results, images have been augmented via rotation, flipping and scaling of the images. This will eliminate the over fitting in favor of the network emulating a generalization well.. The core of the model is the CNN-architecture. First of all, image samples are filtered through convolutional layers and important local features like edges, textures, and color gradients are extracted. Max pooling layers then compress the data to a smaller form without losing the important information numerically efficiency. This is repeated several times so as to learn complex patterns in the data set. Lastly, the extracted features are combined in fully connected layers to classify the input image into the corresponding disease category. To optimize performance, a number of pre-trained networks including VGG16, InceptionV4, ResNet50, and DenseNet121 were tried. Transfer learning enables these models (already trained on large scale data) to behave appropriately when implemented on plant disease

classification with a smaller number of computational resources. The DenseNet121 proved to be the most precise on the results obtained. The testing and the training utilise two phases of evaluation: training and testing. The training period is the process of feeding labeled data into the system, during which the testing characterizes the model on facts that it has never seen. The effectiveness is measured with the help of such performance metrics as accuracy, precision, and recall. The analysis showed that CNN-based models especially DenseNet121 outclassed the traditional handcrafted methods and is effective in detecting plant diseases in a large-scale.





Metndotlology Fron Icun Netwotlon onds

EXPERIMENTAL RESULTS

Personalized train delay forecasting system The proposed train delay forecasting system was tested on past train operational datasets containing schedule times, the actual delays, weather conditions and stations detailsThe data was randomized into three sets, 70 percent- train, 20 percent-validation, and 10 percent - test sets after feature engineering and preprocessing. Such separation guaranteed that the models could be trained productively and gauged on unknown data.

The various machine learning models produced were the Decision Trees, Gradient Boosting, Random Forest, and Logistic Regression. In addition to the aforementioned deep learning models, the Long Short-Term Memory (LSTM) and Convolutional Neural Networks CNNs were also used in order to trace the patterns in time-Wave, as well as to establish complex relations between features.

The findings were that ensemble learning technique always performed better than the conventional classifiers. Decision Trees got 93.7 percent accuracy and the Gradient Boosting had 94.8 percent precision. The most accurate models were Random Forest with 95.3 accuracy. In deep learning, the LSTM performed better in comparison with modelling and estimating the temporal dependencies with an accuracy of 96.1 percent. Another neural net that did fairly well is the CNN architecture when on sequential features slightly underperforming ahead of LSTM.

Input Data

	Started On	Status	Delay	Reach Time
0	03rd, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 04th, Jan
1	05th, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 06th, Jan
2	06th, Jan, 2016 at 12:30 AM	Late	07 Hrs 40 Mins	03:20 AM on 08th, Jan
3	07th, Jan, 2016 at 05:35 PM	Late	15 Mins	07:55 PM on 08th, Jan
4	10th, Jan, 2016 at 08:07 PM	Late	03 Hrs 00 Min	10:40 PM on 11th, Jan
5	12th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 13th, Jan
6	14th, Jan, 2016 at 07:25 PM	Late	01 Hr 32 Mins	09:12 PM on 15th, Jan
7	17th, Jan, 2016 at 05:15 PM	Late	40 Mins	08:20 PM on 18th, Jan
8	19th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 20th, Jan
9	20th, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 21st, Jan
10	21st, Jan, 2016 at 05:15 PM	Late	12 Mins	07:52 PM on 23rd, Jan
11	26th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 27th, Jan
12	27th, Jan, 2016 at 05:15 PM	On Time	0	07:40 PM on 29th, Jan
13	28th, Jan, 2016 at 05:30 PM	On Time	0	07:40 PM on 29th, Jan
14	31st, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 01st, Feb
15	02nd, Feb, 2016 at 05:15 PM	Late	13 Mins	07:53 PM on 04th, Feb
16	03rd, Feb, 2016 at 05:21 PM	Late	15 Mins	07:55 PM on 05th, Feb
17	04th, Feb, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 05th, Feb
18	07th, Feb, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 08th, Feb
19	09th, Feb, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 10th, Feb

Fig 1. Input Data

Before Label Encoding

```
0        Late
1        Late
2        Late
3        Late
4        Late
5        Late
6        Late
7        Late
8        Late
9        Late
10       Late
11       Late
12    On Time
13    On Time
14       Late
Name: Status, dtype: object
```

Fig 2. Before Label Encoding

Machine Learning ----> Random Forest

1. Accuracy : 89.47368421052632 %

2. Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.94	0.94	0.94	54
2	0.00	0.00	0.00	1
accuracy			0.89	57
macro avg	0.31	0.31	0.31	57
weighted avg	0.89	0.89	0.89	57

Fig 4. Random Forest

CONCLUSION

The conclusion that can be made is that the data in this research was taken by a trustworthy source of dataset, which provides a strong foundation to create and test various techniques of categorization. In order to determine and predict train delays properly, we specifically applied the variety of classification algorithms like logistic regression and random forest. These models were trained and tested on this dataset in order to determine how well they work, with important measures such as accuracy, precision, recall, and F1 score being used as indicators of those results. The findings were that the two models were performing alike and each had advantages in addressing the challenges of making a good decision and characterizing the data. As an additional outcome of our work, we also predicted train delays using the knowledge obtained based on the classification models. Along with our predictions, we included illustrative clear visuals highlighting the patterns and trends and potential delay factors in the information to aid understanding and decision-making. Finally, the integrated approach helped to plan operational activities better and provide higher service reliability to rail transportation due to the enhanced precision of delay estimation and the ability to provide a practical way of analyzing, understanding, and reporting the outcome.

REFERENCES

[1] S.Derrible, Urban Engineering for Sustainability. Cambridge, MA, USA: MIT Press, 2019.

[2] R. Nilsson and K. Henning. Predictions of Train Delays Using Machine Learning. 2018. [Online].Available:http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-230224 (Accessed: Jul. 27, 2019).

[3] Amtrak Five Year Service Line Plans FY20-24, Amtrak, Washington, DC, USA, 2019.

[4] W. Peetawan and K. Suthiwartnarueput, "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.

[5] P. Wang and Q. Zhang, "Train delay analysis and prediction based on big data fusion," Transp. Safety Environ., vol. 1, no. 1, pp. 79–88, Jul. 2019, doi: 10.1093/tse/tdy001.