# Prediction of Train Delay in Indian Railways through Machine Learning Techniques

**Article** *in* INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING · February 2019

DOI: 10.26438/ijcse/v7i2.405411

Muqeem Ahmed
MANUU(Central University) ,Hyderabad Telangana India
**27** PUBLICATIONS **72** CITATIONS

IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

# Prediction of Train Delay in Indian Railways through Machine Learning Techniques

**Mohd Arshad**[1*], **Muqeem Ahmed**[2]

[1,2]Department of CS & IT, Maulana Azad National Urdu University, Hyderabad, India

[*]*Corresponding Author: azmi.arshad9044@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i2.405411 | Available online at: www.ijcseonline.org

Accepted: 15/Feb/2019, Published: 28/Feb/2019

***Abstract*** - Train delay is one of the foremost problems in the railway systems across the world. According to the TOI newspaper, In India there are about 25.3 million people were used to travel by train in 2006 and this drastically increased year by year. In 2018, every day at least 80 million people in India prefer to travel by trains[1]. Categorically in India, most of the trains unable to run on their scheduled time due to poor signaling and less number of railway tracks. This implies that travellers might get delayed to reach their respective destinations. The aim of this paper is to present the prediction of Train delay in Indian Railways through machine learning techniques to achieve higher accuracy. In the proposed model, we used 3 different machine learning methods (Multivariate regression, Neural Network, and Random Forest) which have been compared with different settings to find the most accurate method. To compare different methods, we consider training time and accuracy of the method over the test data set. Trains in India get delayed frequently, and if we can predict this in advance - it would be a great help for the passengers to plan their journey according to their works.

***Keywords***: *Train delay, Multivariate Regression, Neural Network, Random Forest.*

## I. INTRODUCTION

Indian railways is 4th largest railway networks in the world that spread over an area of 67,368 km length and revenue of Indian railways is 1.87 lakh crores [1]. Total number of trains running in India has nearly 20 thousands and operated by Indian Railways [2]. In India, most of the trains unable to run on scheduled time due to high congestion, poor signaling and more number of trains.

In a country like India, a vast majority of population depends on the Railways. Approximately 70 million passengers i.e. nearly 5.2% of Indian population travel everyday by train [2]. Trains in India got delayed frequently by which passengers have to face lot of inconvenience, and if we can predict in advance then it would be a great help for the passengers to plan their journey according to their works. Sometimes people do not able to get the reservation of train from source to destination directly, so the people generally prefer break journey. The major drawback of break journey is - if they found their first train late then probably they will miss the second one. To overcome from this problem, prediction of train delay system plays vital role. Prediction of train delay in advance gives much more flexibility to re-schedule your journey.

The main objective of this paper is to provide a way for the development of basic understanding of train delay prediction system and can act as a foundation for further studies within the field. We discuss various tools and machine learning algorithm that may suitable for prediction of train delay prediction system in a better way.

## II. LITERATURE SURVEY

The following tables (Table1 and Table 2) shows summaries of train delay prediction system. In this review paper, we walk around different related research work of the researchers on train delay prediction system and show it in tabular form (Table 1). Out of many research works we conclude 9 papers in tabular form (Table 1) in which year of publication, reference, methodologies or tools, attributes used, conclusion, and research gap/future work are shown and two papers related to flight and bus delay are shown in tabular form (Table 2).

Table 1: Related research work in train delay prediction

| **Year of Publication** | **Reference/ Author's** | **Methodology/Tools** | **Attributes Used / conclusion** | **Research Gap / Future work** |
|---|---|---|---|---|
| 2010 | [8]<br>A. Hansen et al. | Offline Statistical analysis | **Dataset:**<br>Netherlands historical train | 1. Not suitable for large data set.<br>2. Piecewise linear function may |

| | | | | |
|---|---|---|---|---|
| | | | data**,** Dutch railway corridor Rotterdam | maximize the accuracy of prediction in case of large data set. |
| 2012 | [9] Masoud Yaghini et al. | Artifical Neural Network | **Dataset:** historical train data of Iranian Railways<br>**Conclusion**: to evaluate the quality of result decision tree and multinomial logistic regression are used. Approximately 90% accuracy achieved. | Can be Improve through meta heuristic methods such as genetic algorithm or hybrid algorithm. |
| 2013 | [18]<br>Kecman, Pavle et al. | Timed event graph with dynamic arc weights, using Recursive depth first search algorithm | **Dataset:**<br>Netherlands Train data<br>**Conclusion:** produced accurate estimates for train traffic and route conflicts within 30 min | adding some more factors like railways assets conditions, weather data may increase the efficiency |
| 2014 | [10] Suporn Pongnum Kul at al. | *k-Nearest Neighbors-based Algorithm* | **Dataset:** Historical data of Thailand local passenger train.<br>**Conclusion:** Algorithm based on k-NN improves the prediction error by 23% on average (when k=16). | Missing most important factor i.e. weather info. |
| 2014 | [11] Jia Hu, Bernd Noche et al. | Multi-layered perceptron, GA-BPNN model | **Conclusion:**<br>Implementing GA in BPNN improves prediction performance | Small data set, less parameter and The run time of the GA-BPNN is more (usually 3 hours to 10 hours) |
| 2015 | [16] Yaghini, Masoud et al. | Decision Tree and Multinomial logistic Regression | **Dataset:**<br>Iranian Railways | Accuracy of model may be improved through meta heuristic methods like genetic algorithm, hybrid methods etc. |
| 2016 | [12] Robert Nilsson et al. | Decision tree, decision tree with Ada boosting , neural Network | **Dataset:**<br>local traffic of Stockholm using Trafikverket's API<br>**Conclusion:** average error reported 3 minutes in case of neural network | Other methodologies may give better prediction |
| 2016 | [13]<br>L. Oneto et al. | Kernel method, Extreme learning machines, Ensemble Methods | **Dataset:** The Italian Case, Italian train historical data & weather data<br>**Conclusion:** addition of weather info increases the accuracy by 5-10% | railways assets conditions |
| 2018 | [14]<br>R. Gaurav et al. | N-Markov Model, N-OMLMPF algorithm, Random forest Regressor (RFR) model | **Dataset:** Indian railways historical data<br>**Conclusion:** Used to predict late minutes at an inline station. | produce good results but adding some factors like railways assets conditions increase the efficiency |
| 2018 | [15]<br>Jasti Satyakrishna et al. | *Extreme Learning Machines, Shallow & Deep Extreme Learning Machines* | **Dataset:** Train delay historical data | Widening the scope of region/station. Missing most important factor i.e. weather info |

Table 2: Related research work in flight and bus delay prediction

| **Year of Publication** | **Reference/ Author's** | **Methodology/Tools** | **Attributes Used / conclusion** | **Research Gap / Future work** |
|---|---|---|---|---|
| 2013 | [17] Zorkany, M & Zaki, et al. | ANN & Linear Quadratic Estimation | **Dataset:**<br>real-time location data from GPS receivers, historical travel speed as well as temporal and spatial variations of traffic conditions data . | We may increase accuracy of prediction delay by including Weather data. |
| 2017 | [18] Sternberg, Alice, Soares at al. | **Statistical analysis:** regression model, Multivariate analysis (MVA)<br>**Probabilistic analysis:** Genetic algorithm | **Dataset:**<br>Datasets from United States Department of Transportation, National Oceanic and Atmospheric Administration and Weather Company | |

## III. FINDING OF LITERATURE SURVEY

Delay is period of time by which something is late or postponed. Train delay means train has not arrived at its pre-scheduled time. Actually, 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. There are some important causes that leads to the train delays are as follows:

***Delay at the origin:*** (actual departure time - schedule departure time)

***Engine Breakdown:*** Train engine not working properly during the journey.

***Other's train Engine Breakdown:*** other's train engine running on the same track, not working properly during the journey that may leads to the delay.

***Waiting time at overtaking point:*** Passing the others train that are running on the same track according to its priority.

***Climate/weather condition:*** *temperature,* wind speed, visibility.

***Other factors:*** *railways assets condition,* festivals, strikes, national level exams.

## IV. TOOLS USED IN MACHINE LEARNING:

There are lots of tools available to perform Machine Learning task, some of them are listed below in table (shown in Table 3).

Table 3: List of Tools that are used to perform Machine Learning task

| **Ref.** | **Tool** | **Language** | **Type** | **Description** | **Advantages / Disadvantages** | **Use** | **Source?** | **System Requirement** | **Interface** |
|---|---|---|---|---|---|---|---|---|---|
| [21] | LIBLINEAR | C++ & Java | Library | A library for large data sets Classification & Regression | **Advantage:**<br>Easy to use, more suitable for text classification problems. | Support Vector machine and Logistic Regression | Open Source | Average | CLI |
| [22] | R | R | Environment /Language | Functional Language and environment for Statistic. | **Advantage:**<br>It is extensible & offers rich functionality.<br>**Disadvantage:**<br>Memory management, speed, and Security are probably the | Numeric Analysis & Machine Learning | Open Source, GNU license | Average | Both CLI & GUI |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | biggest challenges in R | | | | |
| [23] | MATLAB | MATLAB | Environment /Language | High Level computing Language, Best for computational Analysis. | **Advantages:** computation, visualization and programming is easy. **Disadvantages:** It is very costly. | Algorithm development, data visualization, numerical & data analysis | Closed Source, Proprietary commercial software | High | GUI |
| [24] | Weka 3 | Java | Library | Collection of Machine Learning for data mining tasks. | **Advantages:** platform independent, easy to use. **Disadvantages**: performance issue, relatively slow (Java ) | For developing new ML Algorithm, data mining tasks. | Open Source, GNU license | Average | GUI |
| [25] | Pandas | Python | Library | High Performance, Easy to use data structures and data analysis tools | **Advantages**: handle larger datasets, faster, flexible and powerful. | Data Analysis | Open Source, BSD licensed library. | Average | CLI |
| | NeuroLab | Python | Library | Basic neural networks algorithms with flexible network configurations and learning algorithms for Python. | **Advantages**: Supports large number of Neural Networks. | Neural Network, deep learning | Open Source | High | CLI |
| [26] | Scikit-learn | Python | Library | Scikit-learn support both supervised and unsupervised learning. | **Advantages**: Platform independent, easy to use, robust and supports large data sets. | Clustering, Cross validation, Datasets, Dimensionality Reduction, Ensemble Methods, Feature extraction, Feature selection etc. | Open Source, BSD License Library. | Average | CLI |

By doing comprehensive survey on various machine learning tools, we can conclude that some of these tools are user friendly, easy to install and require less programming knowledge; while some of them might be difficult to use. Due to user Friendly GUI, WEKA can be the easiest tool for the beginners. For users with beginner or intermediate knowledge of Python language, Scikit-learn is one of the best tool to perform machine learning tasks because it contains numerous amount of machine learning libraries and can be combined with python libraries like numpy, scipy, etc. For users with good programming background, R and Matlab can also be one of the options but Matlab & R requires more memory space & heavy processor.

## V. PROPOSED METHODOLOGY

Machine learning is a data analytics technique where machine can learns from its own with prior experiences [3]. There are the following broad phases for prediction of train delay using machine learning:
1. Data collection
2. Exploratory data Analysis
3. Understanding data and feature engineering
4. Training of model
5. Evaluation of result

It is important to find that which variables will affect train delay and how we can use them to predict it. After careful analysis of the problem, we found that there is a close relationship between past delays in particular region and climate of that region, (you can clearly see in the figure). So, we can use both past delay and weather information together of that region to predict future train delay. Data filtration/clustering techniques are useful when we have too much data to process [27].

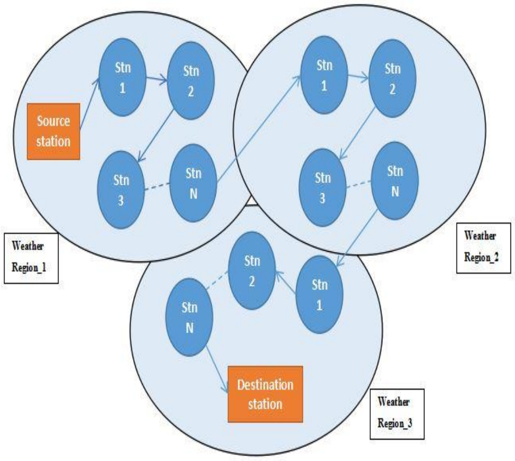

Figure 1: Prediction of Train delay shows the stations in a particular regions and weather information.

There are too many prominent methodologies and tools are used to for the prediction and experiment of train delay.

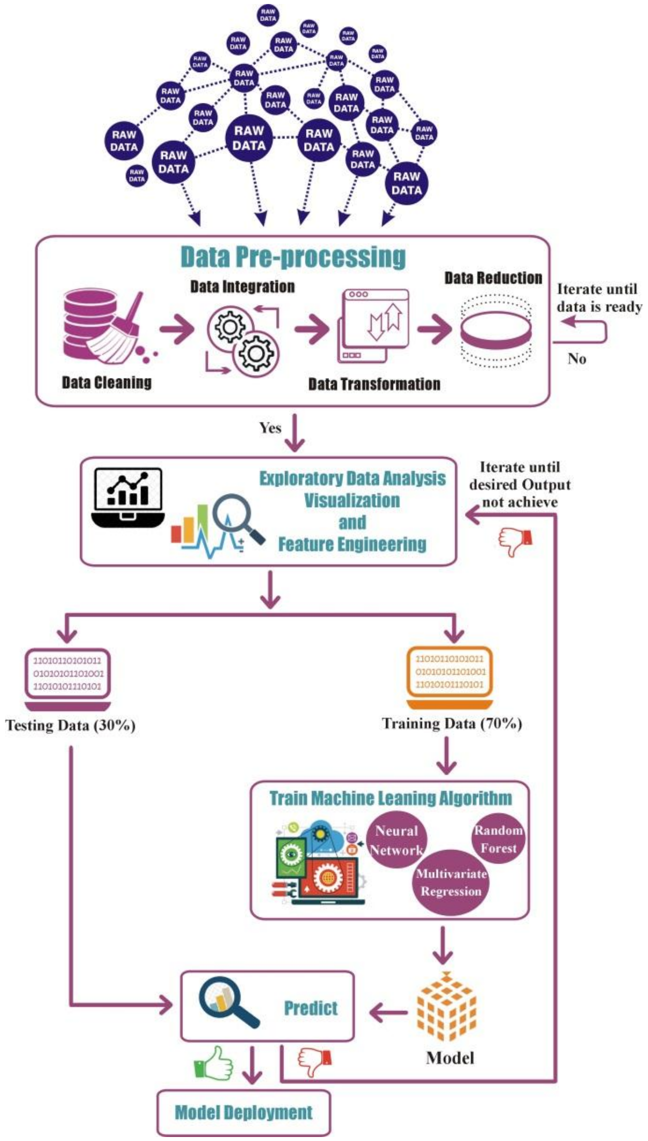

Figure 2: Train delay prediction System using Machine learning.

Train details have been collected from Indian railway's API [19] using python 'requests' library[6], Train delays information have been collected from Indian railways website [20] and weather data have been collected from OpenWheatherMap API. Train delay data in a particular region will be combining with the weather data and the combinations will be saved in a JSON file. The combined data will be split into two parts: 70% is used as training data and the rest as testing data.

Exploratory data analysis will be done in order to remove the null values and erroneous data. To fill null values we use mean value. Feature Engineering will be done in order to understand the data and selection of appropriate algorithm. Python library such as 'matplotlib' is used for data visualization for better understanding of the data.

Machine learning task where the dataset inputs and related outputs are already given comes under supervised learning. A supervised learning contains entire data sets that are further divided into training and testing purpose. The algorithm examines the training dataset and produces an inferred function, which is then used for mapping new examples.

We compare several classes of methods for solving the regression problems. For the first method, Multivariate Regression is used. Second method, Artificial Neural Networks (ANN), we use standard network architectures for the regression problems and third method is Random Forest (RF). Finally, we compare the performance of these three methods.

### A. Multivariate Regression

In Multivariate regression there is more than one input variables used to estimate the target. A model with two input variables can be expressed as:

$$\mathbf{y = K_0 + K_1.x_1 + K_2.x_2}$$

A generalized equation for the multivariate regression model with ‘n’ input variables can be expressed as:

$$\mathbf{y = K_0 + K_1.x_1 + \ldots\ldots\ldots + K_n.x_n}$$

Where K = regression coefficient & n = number of predictors.

### B. Neural Network (NN)

Neural Network (NN) [4] is worked by assembling number of neurons together in layers to produce a desire output. First layer is known as input layer and the last is the output layer. All the layers between first and last is called hidden layers. Every neuron has an activation function (Identity, binary step, Sigmoid, tanh, etc). The parameters of the network are the weights and biases of each layer. The goal of the neural network is to learn the network parameters such that the predicted outcome is the same as the ground truth. We use Back Propagation (BPN) feed forward network for prediction of train delay because it perform good prediction with least error.

### C. Random Forest

Random forest algorithm can use both for classification and the regression kind of problems. RF is an ensemble of Decision Tree (DT) generally trained via bagging method. We can use random forest classifier, which is more convenient and optimized for DT). The RF algorithm includes extra randomness by searching for best feature when splitting the tree. This causes high variance and low bias generally resulting a better model.

## VI. CCONCLUSION & FUTURE WORK

In this paper we are attempting to find the factors that are affected in train delays. It is worthy to mention that this study only examines the latest literature available from 2010 to 2018 that is very much related to train delay prediction system. A total of 11 papers studied to explore the affecting factors of train delay. This research may provide a way for the development of basic understanding of train delay prediction system and can act as a foundation for further studies within the field.

Weather data play an important factor in prediction of train delay in India because weather in India is differing from one state to another state, unlike from other countries. If we consider Real world data from Indian Railways shows that the proposal of this paper will be able to remarkably improve the accuracy of train delay prediction systems and would achieve less error (difference between actual delay and predicted delay). It would also assist the Indian railways companies by giving a possibility to find frequent delays during certain times of the week. The companies could thereafter implement further delay preventions during these particular times of the week in order to maintain a good on-time arrival rate.

## REFERENCES

[1] Ministry of Railways, "Indian Railways year book 2015-16," in ministry of Railways (Railway Board) , 2015.

[2] Ministry of Railway. "Indian Railways Statistical Publications 2016-17: PassengerBusiness" p. 23. Archived (PDF) from the original on 3 March 2018. Retrieved: 2 March 2018

[3] Mathworks, "What is machine learning?", https://se.mathworks. com/discovery/ machine-learning.html Retrieved: 20 October 2018

[4] K. Jain, Anil & Mao, Jianchang & Moidin Mohiuddin, "Artificial Neural Networks: A Tutorial,” 29. 31 - 44. 10.1109/2.485891. (1996).

[5] James G, Witten D, Tibshirani T H R. "An introduction to Statistical Learning with Applications in R" .New York. Springer; 2013

[6] Lutz, Mark “Learning Python (5th ed.)”. O’’Reilly Media. ISBN 978-0-596-15806-4. (2013).

[8] Hansen, I.A. & Goverde, Rob & J. van der Meer, Dirk. “Online train delay recognition and running time prediction.” IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC. 1783 - 1788. 10.1109/ITSC.2010.5625081. (2010).

[9] Y. Masoud et al. “ Railway passenger train delay prediction via neural network model” JOURNAL OF ADVANCED TRANSPORTATION, 47:355–368, (2013)

[10] S. Pongnumkul et al. "Improving arrival time prediction of Thailand's passenger trains using historical travel times," 11th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 307-312, 2014

[11] Jia Hu, Bernd Noche. “Application of Artificial Neuron Network in Analysis of Railway Delays”, Open Journal of Social Sciences, 4, 59-68, 2016

[12] R. Nilsson and K. Henning, “Predictions of train delays using machine learning,” Dissertation DIVA, 2018.

[13] L. Oneto et al. "Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data," IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, 2016, pp. 458-467. 2016

[14] R. Gaurav et al., "Estimating Train Delays in a Large Rail Network Using a Zero Shot Markov Model", arXiv:1806.028251v1 [stat.AP], 7 June 2018

[15] Jasti Satyakrishna et al., " Train Delay Prediction Systems Using Big Data Analytics" , International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 3, March 2018

[16] Yaghini, Masoud & Setayesh Sanai, Maryam & Amin Sadrabady, Hossein., "Passenger Train Delay Classification," International Journal of Applied Metaheuristic Computing 4. 21-31. 10.4018/jamc.2013010102. (2015)

[17] Zorkany, M & Zaki, Mohamed & Ashour, I & Hisham, Basma. "Online bus arrival time prediction using hybrid neural network and Kalman filter techniques." International Journal of Modern Engineering Research (IJMER). V3. (2013)

[18] Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. "A Review on Flight Delay Prediction." (2017)

[19] RailApi, "Indian railway apis," https://railwayapi.com, 2018

[20] https://runningstatus.in, 2018

[21] Fan, Rong-En & Chang, Kai-Wei & Hsieh, Cho-Jui & Wang, Xiang-Rui & Lin, Chih-Jen. (2008). "LIBLINEAR: a library for large linear classification." Journal of Machine Learning Research. 9. 1871-1874. 10.1145/1390681.1442794.

[22] T. R. Foundation. R: What is r? [Online]. Available: https://www.rproject.org/about.html

[23] M. Paluszek and S. Thomas, MATLAB Machine Learning. Apress, 2016.

[24] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[25] pandas: Python data analysis library. Online, 2012, Available: http://pandas.pydata.org/

[26] Scikit-learn. Scikit-learn: machine learning in python. [Online]. Available: http://scikit-learn.org/

[27] Jitendra Oza et al. "Public Transport Tracking and its Issues," International Journal of Computer Sciences and Engineering, Vol.5(11), Nov 2017, E-ISSN: 2347-2693

## Authors Profile

***Mohd Arshad*** pursed Bachelor of Technology in Computer Science from Maulana Azad National Urdu University, Hyderabad, India in 2017 and he is currently Pursuing Master of Technology in Computer Science. He has attended three workshops and one paper present in National Conference conducted by MANUU in 2018. His main research work focuses on Data Mining, Machine Learning and IoT.

***Dr. Muqeem Ahmed*** working as an Assistant Professor at the Department of Computer Science and Information Technology Hyderabad (India).He received his doctoral degree in Computer Science from Jamia Millia Islamia New Delhi India. His professional experience spans over more than 10 years of teaching, research, and project supervision. He has supervised various students for interdisciplinary research and industrial projects. Over the years, he has published many research papers with national and international journals of repute. In addition to these, he is also in the Editorial Boards and Reviewers' Panels of various journals. His primary area of research focuses on semantic web applications, Distributed Database Machine learning and Big data Analytics.