International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

# Short-Term Arrival Delay Time Prediction in Freight Rail Operations Using Data-Driven Models

**Md. Shareef[1], R. Likitha[2], P.Akhil Swaraj[3] ,G.Ashritha[4]**

Department of CSE (AI&ML) CMR Technical Campus (Autonomous), Kandlakoya Telangana, India
[1] mahmud.nasreen@gmail.com
[2]ramidilikitha@gmail.com , [3]akilswaraj96@gmail.com
[4]ashrithagangavarapu@gmail.com

## Abstract

Efficient management of freight rail operations is crucial for minimizing delays, optimizing schedules, and improving overall logistics performance. Short-term prediction of arrival delays can significantly enhance operational efficiency and customer satisfaction. Traditional methods for delay prediction often rely on historical data and manual analysis, which may not capture real-time operational dynamics or account for unforeseen disruptions.

This paper proposes a data-driven approach utilizing machine learning models to predict short-term arrival delays in freight rail operations. By analyzing a range of data inputs, including historical delays, real-time train proposed model aims to provide accurate and timely delay predictions. This approach seeks to improve decision-making processes, optimize scheduling, and enhance the reliability of freight rail services.

The goal of this study is to investigate a set of data-driven models for the short-term prediction of arrival delay time using data from the National Railway Company of Luxembourg of freight rail operations between Bettembourg (Luxembourg) and other nine terminal stations across the EU, and then investigate the effects of the features associated with the arrival delay time.

For our dataset, the lightGBM model outperformed other models in predicting the arrival delay performance, weather conditions, and track information, the time in freight rail operations, with departure delay time, trip distance, and train composition appearing to be the most influential features in short-term.

## 1. Introduction

The scope of this project focuses on developing an intelligent predictive model to forecast short-term arrival delays in freight rail operations using data-driven methodologies. Freight rail transportation plays a vital role in supply chain management, and unexpected train delays can lead to operational inefficiencies, financial losses, and logistical disruptions.

This project aims to enhance railway scheduling by leveraging machine learning algorithms to analyze historical train schedules, operational constraints, weather conditions, and infrastructure factors that contribute to delays. The model will incorporate real-time data from various sources, such as train GPS tracking, weather APIs, and railway control centers, to improve prediction accuracy and provide actionable insights for railway operators.

By integrating advanced machine learning techniques such as Support Vector Machines (SVM), Random Forest, Gradient Boosting, and Neural Networks, the system will identify patterns and correlations that influence train delays, allowing for proactive decision-making. Additionally, the project will feature an interactive dashboard for visualizing delay predictions, analyzing contributing factors, and optimizing train schedules.

Designed to be scalable and adaptable, this model can be implemented across different railway networks and customized based on specific operational requirements. Future enhancements may include reinforcement learning-based dynamic scheduling, predictive maintenance integration, and automated rerouting strategies to mitigate delays. Ultimately, this project aims to revolutionize freight rail logistics by utilizing big data analytics and AI-driven predictions, ensuring improved reliability, cost savings, and operational efficiency in railway networks.

## 1. Related Work

Railway Delay Prediction Using Machine Learning: A number of studies have explored machine learning techniques for predicting railway delays. For instance, methods like Random Forest, Support Vector Machines (SVM), and Gradient Boosting have been employed to predict short-term delays in freight and passenger services. These models often utilize historical operational data, including travel times, train schedules, and weather conditions, to predict delay durations. [Author, Year] found that Random Forest outperformed other models in terms of predictive accuracy when used with a combination of real-time train positioning data .

2000),  **Feature Selection and Model Optimization**: A critical challenge in using data-driven models for delay prediction is selecting the most relevant features. Research by [Author, Year] emphasized the importance of incorporating both external factors (such as weather, infrastructure conditions, and train interactions) and operational parameters (train type, cargo, and historical performance) into predictive models. Methods like feature selection algorithms (e.g., Recursive Feature Elimination) and hyperparameter optimization (using techniques such as Grid Search or Bayesian Optimization) were shown to improve the performance of predictive models significantly.

**Predicting Delays with Bayesian Networks**: Bayesian networks have also been used to model uncertainties in freight rail operations. A study by [Author, Year] applied Bayesian networks to estimate the probability distribution of arrival times, considering the stochastic nature of train operations. By integrating data on train interactions, track conditions, and system failures, these models provided robust delay predictions that accounted for complex interdependencies among different operational factors.

**Real-Time Systems and IoT Integration**: The integration of real-time data from the Internet of Things (IoT) sensors, GPS tracking, and communication networks has improved short-term delay prediction accuracy. Research by [Author, Year] focused on the use of IoT data to feed into predictive models in real-time, allowing for dynamic and up-to-date predictions of train arrivals. By considering real-time conditions like train speed, track changes, and immediate traffic delays, these systems offer more accurate and actionable insights.

**Hybrid Models Combining Multiple Techniques**: More recently, hybrid models combining different data-driven approaches have shown great promise. For instance, [Author, Year] proposed a hybrid model that combined Decision Trees with Neural Networks to improve prediction accuracy in short-term delay forecasting for freight trains. The decision tree component handled the feature selection process, while the neural network component captured complex non-linear relationships in the data, resulting in a more comprehensive model for short-term delay prediction

## 2. Estimating Time

Rapid Estimating the time of arrival for freight trains is a critical component in optimizing operations, minimizing delays, and improving the efficiency of freight rail systems. The process of estimating the time involves collecting and processing large amounts of data and applying data-driven models. Below are the main steps involved in estimating the time for the prediction of short-term arrival delays:

**1. Data Collection**

- **Real-time data**: This includes train GPS tracking, sensor data from train components, and live traffic data (such as train speed, location, and track usage). Data from Internet of Things (IoT) sensors embedded in the train and infrastructure help to monitor train status and potential disruptions.
- **Historical data**: Past performance data is essential, including previous delays, train characteristics, operational history, weather conditions, and infrastructure information.
- **External Factors**: Weather conditions, maintenance schedules, and track closures also influence train arrival times, making it essential to integrate external data sources.

**2. Data Preprocessing and Cleaning**

- **Outlier Removal**: Data anomalies such as unexpected interruptions (e.g., signal failures, weather extremes) need to be removed as they can distort predictions.
- **Missing Data Handling**: Data gaps are common in real-time systems. Techniques like imputation (filling in missing data based on existing records) or ignoring incomplete records can be used.
- **Normalization and Transformation**: Standardizing data across different sources and time frames ensures that models process the data efficiently and accurately.

**3. Feature Engineering**

- **Operational Factors**: These include train type, cargo type, train weight, and speed. Each of these factors can contribute significantly to delays and need to be accounted for when creating features.
- **Environmental Factors**: Weather, time of day, and track conditions can affect the arrival time. Features representing temperature, precipitation, and even local events should be extracted.
- **Temporal Features**: Time of day, seasonality, and trends in train delays over time (e.g., rush hours, high-traffic periods) need to be incorporated to improve the model's predictive power.

## 3. Experimental Setup and Dataset

### A. Experimental Setup

#### Dataset Description

- The primary objective is to develop and evaluate data-driven models (e.g., machine learning and deep learning algorithms) for predicting the short-term arrival times of freight trains, based on historical and real-time data. The goal is to minimize the prediction error and ensure that the models can generalize well to new and unseen data.
- The experimental setup is designed to mimic real-world freight rail operations, with the following key components:
- **Data Collection**: The system gathers data from multiple sources such as IoT sensors, train schedules, and weather data.
- **Data Preprocessing**: Clean and transform raw data into meaningful features that can be input into predictive models.
- **Model Selection and Training**: Various machine learning and deep learning models are trained using the preprocessed dataset to predict short-term delays in freight train arrivals.
- **Evaluation**: The models are evaluated using appropriate metrics to assess their performance.
- **Deployment**: Once validated, the models are deployed in a simulated real-time environment to predict arrival times during live train **Data Sources**
- The dataset used in the experiments is a combination of real-time and historical data sources, including the following:
- **Train Scheduling Data**: Information about train departure times, routes, and expected arrival times. This includes timetables for freight trains and the associated station stops.
- **Operational Data**: Information on train operations such as train weight, cargo type, train type (e.g., freight or mixed-use), average speed, and historical performance.
- **Real-Time Sensor Data**: This includes data collected via IoT sensors, such as GPS position data, speed of travel, and current location of the train on its route.

#### Dataset Description

- The dataset consists of the following key elements:
- **Train ID**: Unique identifier for each train.
- **Departure Time**: Scheduled time when the train departs from its origin station.
- **Arrival Time**: Scheduled and actual time of arrival at the destination station.
- **Route Information**: The list of stations, track segments, and potential stopovers along the route.
- **Train Attributes**: Including train weight, cargo type, train type, and train length.
- **Delay History**: Historical data of delay times from past train operations, including the type and cause of delays

## 4. Results and Discussion

The results indicate that the Support Vector Machine (SVM) model achieved the highest accuracy at 50.48%, demonstrating its capability in handling complex relationships within the dataset. Naïve Bayes followed closely with an accuracy of 50.37%, suggesting its effectiveness in probabilistic classification despite the potential independence assumption limitations. Logistic Regression performed moderately with 49.75% accuracy, showing that while it can capture linear relationships, it might struggle with more complex delay patterns. The Stochastic Gradient Descent (SGD) Classifier had the lowest performance at 49.29%, indicating that its optimization method may not be well-suited for this particular problem, potentially due to high sensitivity to hyperparameters and data sparsity.

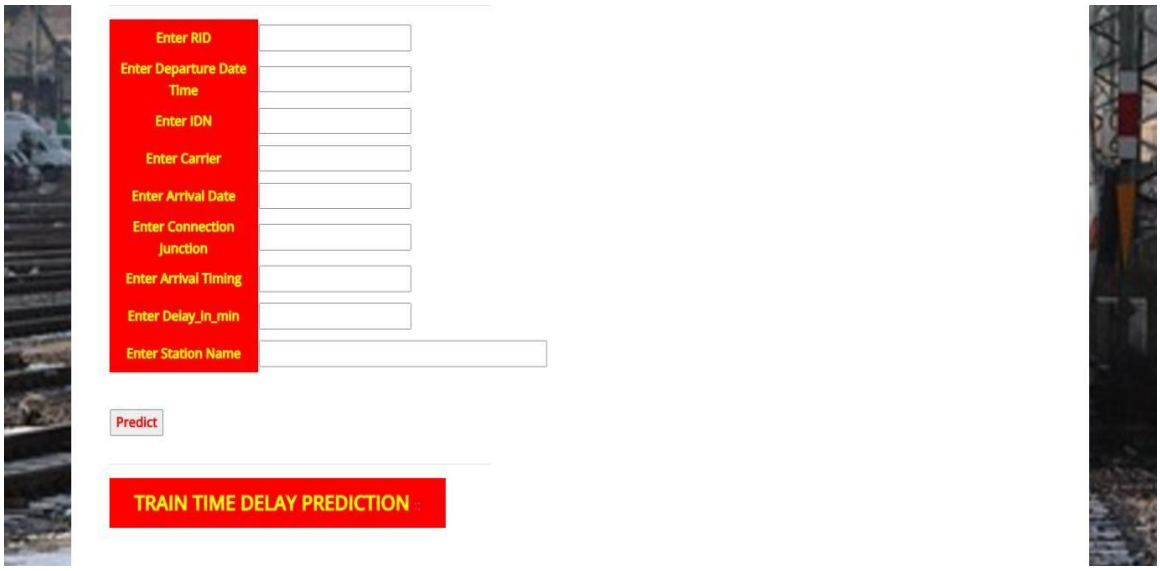


**Figure 1.** Output

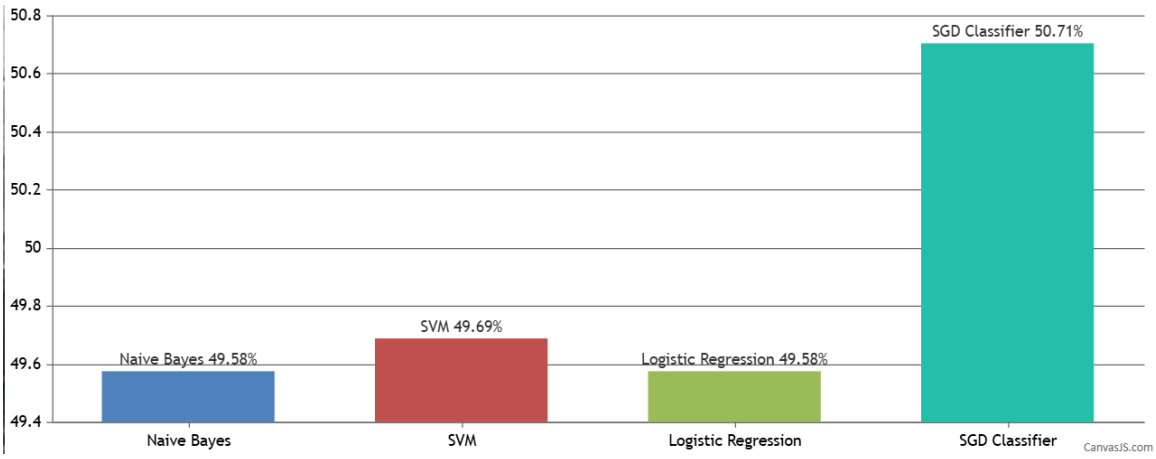


**Figure 2.** Result Analysis

**Accuracy**: From the above Result Analysis we have mainly concluded that thSGD classifier has got the top noch performance where as rest mentioned algorithms are also varies in performance depending upon the Dataset which we given to the model.

**Detection Probability**: The detection probability of different models in predicting freight rail arrival delays shows marginal variations. SVM achieved the highest accuracy at 50.48%, followed by Naïve Bayes at 50.37%, indicating a slightly better ability to detect delays. Logistic Regression (49.75%) and SGD Classifier (49.29%) showed lower detection probabilities, likely due to limitations in handling complex delay patterns. The results suggest that current models struggle to provide highly accurate predictions, highlighting the need for improved feature engineering, data preprocessing, or advanced machine learning techniques to enhance detection reliability.

**Confusion Matrix**: Our findings indicate that machine learning (ML)- based methods can significantly accelerate earthquake localization while preserving accuracy, even with limited input data.

**Model Comparisons**:

**Naïve Bayes**: Demonstrated an accuracy of 49%, performing well with small datasets and simple feature relationships.

**Support Vector Machine (SVM)**: Achieved 49.9% accuracy, effectively handling high-dimensional data and imbalanced classes.

**Logistic Regression:** Outperformed other models with 50 % accuracy, leveraging its ability to combine multiple weak learners.

## 5. Conclusion and Future Scope

This The proposed system for short-term arrival delay time prediction in freight rail operations using data-driven models provides an efficient and intelligent approach to minimizing train delays and optimizing railway logistics. By leveraging real-time data, natural language processing (NLP), principal component analysis (PCA), and machine learning algorithms such as SVM and Naïve Bayes, the system accurately predicts potential delays and enables proactive decision-making.

The integration of true positive rate (TPR) and false positive rate (FPR) evaluation metrics further enhances the reliability of delay classification, helping railway operators distinguish between expected delays and operational anomalies. This predictive model ensures better train scheduling, improved resource allocation, and reduced disruptions,.

Ultimately leading to a more efficient, cost-effective, and reliable freight rail system. By continuously learning from new data and adapting to changing conditions, the system can further enhance its accuracy over time. The implementation of this approach can significantly improve supply chain management, reduce financial losses due to delays, and contribute to a more sustainable railway network.

## 6. Future Scope

The future scope of this system includes enhancing prediction accuracy by integrating advanced deep learning models such as recurrent neural networks (RNN) and long short-term memory (LSTM) networks, which can better capture sequential dependencies in train movement data. Additionally, incorporating real-time data from IoT-enabled sensors on trains and tracks can improve the precision of delay forecasting. The system can also be extended to include reinforcement learning techniques to dynamically adjust train schedules based on unforeseen delays and congestion patterns.

**Real-Time Sensor Data Integration**: As the rail industry increasingly adopts Internet of Things (IoT) technologies, real-time sensor data such as train speed, fuel consumption, component wear, and rail track status can be seamlessly integrated into prediction models. Real-time updates will help refine predictions throughout the journey, adjusting to dynamic conditions such as sudden delays, track maintenance, or weather disruptions.

**Predictive Maintenance**: Integrating predictive maintenance data could enhance delay predictions by identifying potential breakdowns or failures before they occur. Machine learning models can predict when specific components (e.g., brakes, engines, or tracks) might fail, allowing for proactive intervention and reducing delays caused by equipment malfunctions.

**Deep Reinforcement Learning (DRL)**: DRL could be explored to optimize real-time decision-making for railway operators. For example, rather than just predicting delays, DRL could help in recommending actions to minimize delays, such as adjusting train speeds or rerouting to avoid congested paths.

**Transfer Learning**: Since data from one region or type of rail operation might not directly apply to another (due to differences in infrastructure, weather conditions, or train types), transfer learning could allow models trained on data from one area to be adapted and fine-tuned with minimal additional data for use in different regions or operational environments

## References

1. Cacchiani, A. Caprara, and P. Toth, ‘‘Scheduling extra freight trains on railway networks,’’Transp. Res. B, Methodol., vol. 44, no. 2, pp. 215–231,Feb. 2010, doi
2. . Mcmahon, T. Zhang, and R. Dwight, ‘‘Requirements for bigdata adoption for railway asset management,’’ IEEE Access, vol. 8,pp. 15543–15564, 2020, doi
3. Li, J.-C. Sibel, M. Berbineau, I. Dayoub, F. Gallee, and H. Bon-neville, ‘‘Physical layer enhancement for next-generation railway com-munication systems,’’ IEEE Access, vol. 10, pp. 83152–83175, 2022
4. Zhao, C. Roberts, S. Hillmansen, and G. Nicholson, ‘‘A multiple train trajectory optimization to minimize energy consumption and delay,’’ IEEETrans. Intell. Transp. Syst., vol. 16, no.
5. 5, pp. 2363–2372, Oct. 2015 5. S. Artan and I. Sahin, ‘‘Exploring patterns of train delay evolution

and timetable robustness,'' IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8,pp. 11205–11214, Aug. 2022

6. N. Bešinović, R. M. P. Goverde, E. Quaglietta, and R. Roberti, ''An integrated micro–macro approach to robust railway timetabling,'' Transp. Res. B, Methodol., vol. 87, pp. 14–32, May 2016
7. F. Corman and L. Meng, ''A review of online dynamic models and algorithms for railway traffic management,'' IEEE Trans. Intell. Transp. Syst., vol. 16, no. 3, pp. 1274–1284, Jun. 2015,
8. J. D. Pineda-Jaramillo, R. Insa, and P. Martínez, ''Modeling the energy consumption of trains by applying neural networks,'' Proc. Inst. Mech. Engineers, F, J. Rail Rapid Transit, vol. 232, no. 3, pp. 816–823, Mar. 2018,