Predicting Short-Term Arrival Delay in Freight Rail Operations Using XGBoost and Voting Classifiers

Anas Syed Shoebul Hasan[1], Sridhar Gummalla[2], Subramanian K.M[3]
[1] *PG Scholar, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086*
[2] *Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086*
[3] *Professor, Department of Computer Science and Engineering, Shadan College of Engineering and Technology, Hyderabad, Telangana, India-500086*

***Abstract*— Despite rail's growing popularity as a mode of freight transportation due to its role in intermodal transportation and numerous economic and environmental benefits, optimizing all aspects of rail infrastructure use remains a significant challenge. To address this issue, various methods for developing train disruption prediction models have been used. However, these models continue to struggle with accurately predicting short-term arrival delay times, as well as identifying the causes of delays and the expected impact on operations. The lack of information available to operators makes it difficult for them to effectively mitigate the effects of disruptions. 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 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 time in freight rail operations, with departure delay time, trip distance, and train composition appearing to be the most influential features in predicting the arrival delay time in the short-term. Knowing a train's arrival delay time allows you to estimate future operational time, providing more support to reduce disruptions and subsequent operational delays via a simple web service.**

***Index Terms*— Arrival delay prediction, Fright rail operations, LightGBM, Machine Learning, Real-Time Prediction, Transportation Analytics, Voting Classifier, XGBoost**

1. INTRODUCTION

The freight transportation industry is constantly changing, and rail transportation is becoming a more popular option due to its advantages in terms of operational costs, efficiency, reliability, emissions, and safety. This trend has resulted in the gradual integration of rail into intermodal transportation, with public agencies encouraging a shift away from other alternatives.

As rail intermodal operations become more important for the efficiency and dependability of the freight transport industry, optimizing all aspects of rail infrastructure is critical. This includes ensuring that the infrastructure is well-maintained, properly managed, and capable of meeting the current transportation system's demands. Furthermore, using technology and data analytics to optimize rail infrastructure and improve overall performance of the rail transport sector is critical.

Data-driven models offer promise in addressing this issue, but past studies struggled with accuracy. A new project focuses on freight rail operations and shows XGBoost as a top performer in predicting arrival delays, using factors like departure delay time, trip distance, and train composition. This predictive model can reduce disruptions and operational delays, enhancing rail transportation efficiency.

1.1 OBJECTIVE

- To develop a data-driven model for short-term arrival delay prediction in freight rail operations, focusing on accurate estimates at each station.
- To evaluate the performance of machine learning algorithms, including LIGHTGBM, XGBOOST, Random Forest, KNN, and Linear Regression, using metrics like RMSE, MAPE, MAE, and R2.

- As an extension to the project, we aim to experiment with the advanced XGBoost algorithm with 100 estimators and voting classifier to optimize feature selection and enhance R2 scores.

1.2 PROPOSED SYSTEM

The proposed system employs data-driven models, particularly lightGBM, to enhance freight rail operations. It prioritizes factors like departure delay, trip distance, and train composition for precise arrival delay predictions, offering decision support to estimate future operational times, minimize disruptions, and optimize rail infrastructure.
This comprehensive approach addresses the limitations of traditional methods, providing a forward-looking solution for more efficient and reliable railway operations.

ADVANTAGES OF PROPOSED SYSTEM

- Enhances short term arrival delay predictions, reducing disruptions.
- Empowers operators with estimates for proactive decision-making.
- Optimizes rail asset usage, improving efficiency.
- Helps understand delay reasons for operational improvements.

2. LITERATURE SURVEY

A. Scheduling extra freight trains on railway networks:
We study the problem of freight transportation in railway networks, where both passenger and freight trains run. While the passenger trains have a prescribed timetable that cannot be changed, freight train operators send the infrastructure manager requests to insert new freight trains. For each freight train, the associated train operator specifies a preferred ideal timetable, which can be modified by the infrastructure manager in order to respect safeness operational constraints. In particular, this modification may correspond to routing the train along a path which is different with respect to the one in the ideal timetable. Roughly speaking, the objective is to introduce as many new freight trains as possible by assigning them timetables that are as close as possible to the ideal ones. For this timetabling problem on a generic railway network, we present an integer linear programming formulation, that generalizes some formulations already presented for the case of a single railway line, and a Lagrangian heuristic based on this formulation. Computational results on real-world instances are reported.

B. Requirements for Big Data Adoption for Railway Asset Management:
Nowadays, huge amounts of data have been captured along with the day-to-day operation of assets including railway systems. Hence, we have come to the era of big data. The utilization of big data technologies for asset condition information management is becoming indispensable for improving asset management decision making. The vital information such as precursor information collected on failure modes and knowledge that may be available for analysis is hidden within the large extent of data. There are analysis tools incorporated with techniques such as multiple regression analysis and machine learning that are facilitated by the availability of big data. Therefore, the utilization of big data technologies for asset condition information management is becoming indispensable for improving asset management decision making. This paper provides a review of the requirements and challenges for big data analytics applications to railway asset management. The review focuses on railway asset data collection, data management, data applications with the implementation of Blockchain technology as well as big data analytics technologies. The need for, and the importance of big data analytics in railway asset management; and the requirement for the asset condition data collection in the railway industry are highlighted.

C. A multiple train trajectory optimization to minimize energy consumption and delay:
In railway operations, if the journey of a preceding train is disturbed, the service interval between it and the following trains may fall below the minimum line headway distance. If this occurs, train interactions will happen, which will result in extra energy usage, knock-on delays, and penalties for the operators. This paper describes a train trajectory (driving speed curve) optimization study to consider the tradeoff between reductions in train energy usage against increases in

delay penalty in a delay situation with a fixed block signaling system. The interactions between trains are considered by recalculating the behavior of the second and subsequent trains based on the performance of all trains in the network, apart from the leading train. A multitrain simulator was developed specifically for the study. Three searching methods, namely, enhanced brute force, ant colony optimization, and genetic algorithm, are implemented in order to find the optimal results quickly and efficiently. The result shows that, by using optimal train trajectories and driving styles, interactions between trains can be reduced, thereby improving performance and reducing the energy required. This also has the effect of improving safety and passenger comfort.

D. Exploring Patterns of Train Delay Evolution and Timetable Robustness:

Uncertainty affects the overall performance of railway systems. It can be captured tracking the successive departure and arrival events of trains at stations in the scheduled railways. Markov chains have shown to be an effective means to model variabilities in the departure and arrival delays of trains. The empirical data of actual operations are used to develop the Markov matrices. Transitions between delay states comprise the internal dynamics of running and dwelling processes in probabilistic terms, presenting particular patterns of deterioration, recovery and state keeping in the prevailing timetable setting. All these features conform to the practice of railway operations as well as train dispatchers and drivers. Hence, delay-based Markov chains becomes a simple and effective tool to predict delays in the following events of trains and distribution of delays over a longer period of time using simple performance measures. We have developed the homogeneous and non-homogeneous Markov models and tested their performances using the actual data collected in the Netherlands railway network. The analysis and evaluation show that the railway corridor considered in this study performs well and the corresponding Markov chain models can be used effectively to predict train delays and to evaluate timetable robustness. The outcomes of the homogeneous and non-homogeneous model are evaluated with respect to the observed data resulting that the non-homogeneous model performs better than the homogeneous model.

3. SYSTEM DESIGN

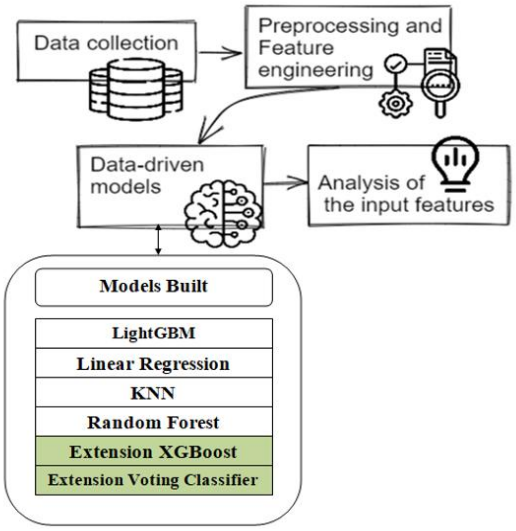

Fig 1. System Architecture

A. Data Collection:

This module gathers historical and real-time freight rail data, including train schedules, arrival/departure times, weather data, track conditions, and operational logs. The quality and diversity of this data are crucial for accurate prediction.

B. Data Preprocessing:

Raw data often contains missing values, noise, or inconsistent formats. This module cleans the data by:

- Filling or removing missing values
- Normalizing numerical values (e.g., delay time, speed)
- Creating new features like time of day or delay difference

C. Model Training and Tuning:

In this core module, machine learning models are trained on the preprocessed data. Two approaches are used:

- XGBoost: A gradient boosting algorithm known for high accuracy and speed.
- Voting Classifier: Combines multiple classifiers (e.g., XGBoost, Random Forest, Logistic Regression) to improve performance.

D. Prediction System:

Once the best model is selected, this module accepts new real-time input data and predicts whether a train

will be delayed or not (classification). It can be integrated with existing logistics or monitoring systems.

E. Visualization & Output:
The final results are presented using visual dashboards and the output is displayed on the screen.

4. RESULT

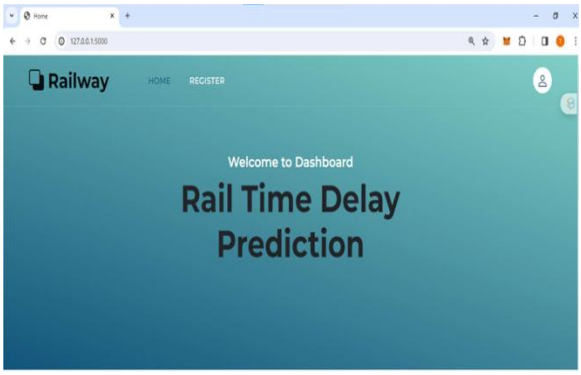

FIG 2. Homepage
Figure 2. Dashboard of the application

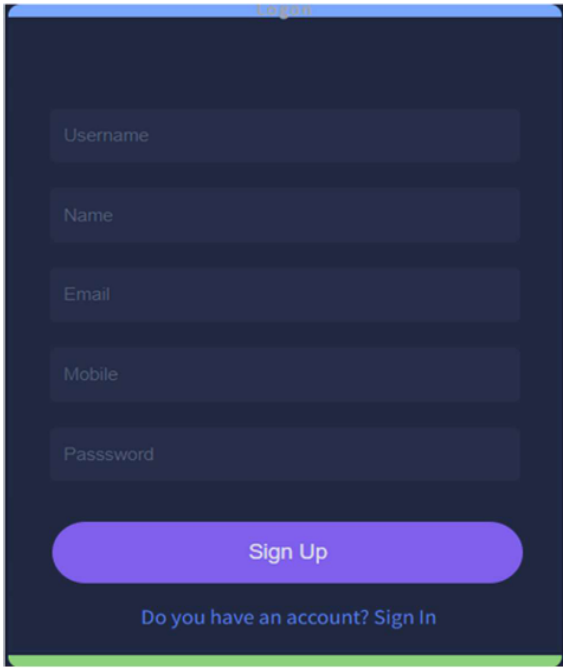

FIG 3. User Sign Up page
Figure 3. User registers his/her account

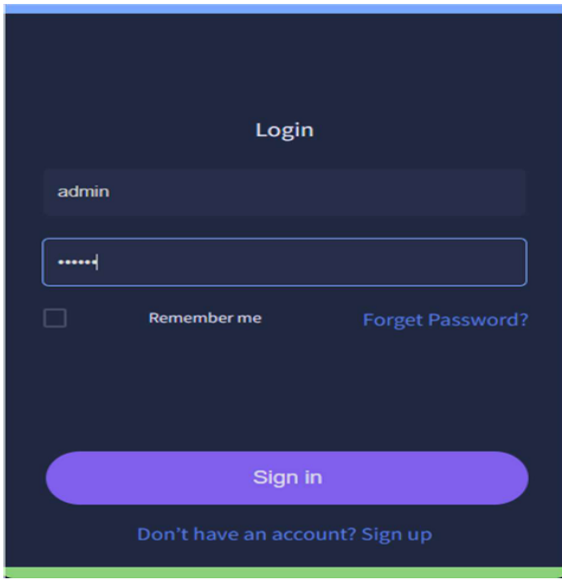

FIG 4. User Login page
Figure 4. User signs in into the application

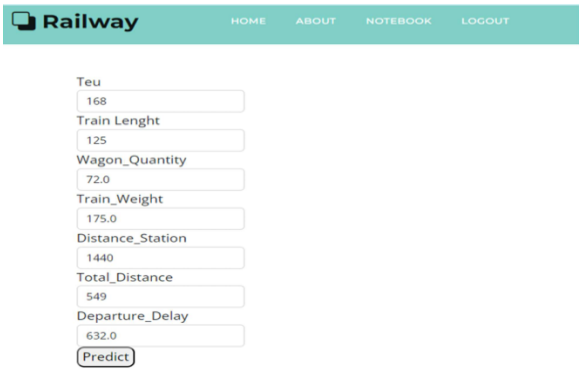

FIG 5. Enter input for prediction
Figure 5. User enters the real time values for prediction

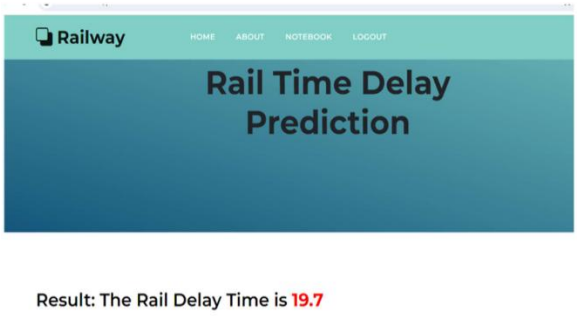

FIG 6. Prediction results displayed
Figure 6. The results are displayed on the screen

5. CONCLUSION

This study presents a comprehensive approach to predicting freight rail arrival delay times, as well as

investigating the underlying causes of delays and their expected impact on operations. The goal is to predict operational delays in real time and to create a STDSS that will assist decision-makers in future operational interventions to reduce disruptions and the resulting delays in freight operations. This will improve railway reliability in the freight transport sector in the long run.

Previous studies have developed models that predict the occurrence of disruptions or delay times in railway operations, but most of them have focused on passenger trains. Freight train research has primarily focused on examining the impact of network delays rather than train delays and has been unable to predict short-term delay times once the train has departed from the previous control station.

In this study, we used regression algorithms to train five data-driven models and analyzed predefined evaluation metrics (R2, RMSE, MAPE and MAE). For the examined dataset, which included railway operations carried out between Luxembourg and nine stations in Belgium, France, Germany, Poland, and Italy over a 17-month period, the lightGBM model stood out as the best data-driven model to predict arrival delay times in freight rail operations. The lightGBM model has demonstrated that departure delay time, trip distance, and train composition are variables with a significant impact on the prediction of arrival delay times in railway operations.

Our findings show that longer trains, longer distances, and heavier trains all have a direct relationship with arrival delay times in general. These findings may pave the way for future research into optimizing the routes of these freight trains' operations to reduce distances, resulting in not only shorter operating times, but also shorter arrival delay times.

Ultimately, accurate arrival time predictions benefit transportation, logistics, and manufacturing, enhancing resource allocation, reducing costs and efforts and ensuring reliable delivery schedules for improved operational efficiency.

6. FUTURE ENHANCEMENTS

- Future research can focus on optimizing freight train routes to reduce distances and operating times, ultimately leading to shorter arrival delay times.
- Automation of the entire workflow, from data extraction to model construction and implementation, can be explored to enhance efficiency.
- The short-term prediction model can be implemented as a user-friendly web service for real-time arrival delay predictions and proactive measures to reduce disruptions.
- Insights from analyzing delay characteristics and the use of the SHAP method can guide further research in improving freight rail operations and understanding delay causes and impacts.

REFERENCE

[1] V. 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: 10.1016/j.trb.2009.07.007.

[2] P. Mcmahon, T. Zhang, and R. Dwight, ‘‘Requirements for big data adoption for railway asset management,’’ IEEE Access, vol. 8, pp. 15543–15564, 2020, doi: 10.1109/ACCESS.2020.2967436.

[3] Q. Li, J.-C. Sibel, M. Berbineau, I. Dayoub, F. Gallee, and H. Bonneville, ‘‘Physical layer enhancement for next-generation railway communication systems,’’ IEEE Access, vol. 10, pp. 83152–83175, 2022, doi: 10.1109/ACCESS.2022.3192971.

[4] N. Zhao, C. Roberts, S. Hillmansen, and G. Nicholson, ‘‘A multiple train trajectory optimization to minimize energy consumption and delay,’’ IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2363–2372, Oct. 2015, doi: 10.1109/TITS.2014.2388356.

[5] M. 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, doi: 10.1109/TITS.2021.3101530.

[6] F. Bigi, T. Bosi, J. Pineda-Jaramillo, F. Viti, and A. D'Ariano. (2023). Addressing the Impact of Maintenance in Shunting Operations Through Shunt-In Policies for Freight Trains Operations.

[7] 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, doi: 10.1016/j.trb.2016.02.004.

[8] 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, doi: 10.1109/TITS.2014.2358392.

[9] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, and D. Dera, ‘‘Machine learning in transportation data analytics,’’ in Data Analytics for Intelligent Transportation Systems, M. Chowdhury, A. Apon, and K. Dey, Eds. Amsterdam, The Netherlands: Elsevier, 2017, pp. 283–307, doi: 10.1016/B978- 0-12-809715 1.00012-2.

[10] 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, doi: 10.1177/0954409717694522.

[11] S. Milinković, M. Marković, S. Vesković, M. Ivić, and N. Pavlović, ‘‘A fuzzy Petri net model to estimate train delays,’’ Simul. Model. Pract. Theory, vol. 33, pp. 144–157, Apr. 2013, doi: 10.1016/j.simpat.2012.12.005.

[12] V. De Martinis and F. Corman, ‘‘Data-driven perspectives for energy efficient operations in railway systems: Current practices and future opportunities,’’ Transp. Res. C, Emerg. Technol., vol. 95, pp. 679–697, Oct. 2018, doi: 10.1016/j.trc.2018.08.008.

[13] J. Pineda-Jaramillo, P. Martínez-Fernández, I. Villalba-Sanchis, P. Salvador Zuriaga, and R. Insa-Franco, ‘‘Predicting the traction power of metropolitan railway lines using different machine learning models,’’ Int. J. Rail Transp., vol. 9, no. 5, pp. 461–478, Sep. 2021, doi: 10.1080/23248378.2020.1829513.

[14] H. Alawad, S. Kaewunruen, and M. An, ‘‘A deep learning approach towards railway safety risk assessment,’’ IEEE Access, vol. 8, pp. 102811–102832, 2020, doi: 10.1109/ACCESS.2020.2997946.

[15] P. Kecman and R. M. P. Goverde, ‘‘Predictive modelling of running and dwell times in railway traffic,’’ Public Transp., vol. 7, no. 3, pp. 295–319, Dec. 2015, doi: 10.1007/s12469-015-0106-7.

[16] D. Li, W. Daamen, and R. M. P. Goverde, ‘‘Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station,’’ J. Adv. Transp., vol. 50, no. 5, pp. 877–896, Aug. 2016, doi: 10.1002/atr.1380.

[17] P. Huang, C. Wen, Q. Peng, C. Jiang, Y. Yang, and Z. Fu, ‘‘Modeling the influence of disturbances in high-speed railway systems,’’ J. Adv. Transp., vol. 2019, pp. 1–13, Mar. 2019, doi: 10.1155/2019/8639589.

[18] C. Wen, Z. Li, J. Lessan, L. Fu, P. Huang, and C. Jiang, ‘‘Statistical investigation on train primary delay based on real records: Evidence from Wuhan Guangzhou HSR,’’ Int. J. Rail Transp., vol. 5, no. 3, pp. 170–189, Jul. 2017, doi: 10.1080/23248378.2017.1307144.