Train Delay Prediction Using Machine Learning

Pushpalatha.A[1], Arish A[2], Chinrasu S[3], Manojkumar S[4], Muthuraman P[5]
[1]Assistant Professor, Department of Computer Science Engineering,
Maha Barathi Engineering College, Affiliated to Anna University, Chinnasalem (Tk), Kallakurichi (Dt)-606 201.
[2,3,4,5]UG Students, Department of Computer Science Engineering,
Maha Barathi Engineering College, Affiliated to Anna University, Chinnasalem (Tk), Kallakurichi (Dt)-606 201.

***Abstract*—Train transport systems are integral parts of urban mobility and national logistics; However, unexpected delay often disrupts operations, reduces efficiency, and affects passenger satisfaction. Traditional approaches to delay the management struggle for the dynamic and interaction nature of the railway network. This paper introduces a three-pronged prediction framework designed to enhance the reliability and accuracy of train delay forecasts. A hybrid sequence-regression model is developed, where Long Short-Term Memory Extreme Gradient Boosting (LSTM+ XGBoost) networks analyse the data to extract the output, and XGBoost refines predictions based on contextual framework such as weather conditions, train type, and station congestion. To find the spatial delay propagation, a Graph Neural Network (GNN) models the railway network as a graph, enabling the system to forecast delay cascades across interconnected stations. Additionally, a Transformer-Based Attention (TBA) mechanism is employed to improve the interpretability of the predictions by prioritizing significant temporal features and external influences. Simulated and real-world datasets from public railway systems are used for training and evaluation. Results achieve 91% classification accuracy and reduce the Mean Absolute Error (MAE) by 18% compared to conventional methods. The findings confirm that integrating temporal modelling, spatial reasoning, and attention-based mechanisms significantly improves delay prediction performance in complex rail environments.**

***Index Terms*—Train Delay Prediction, LSTM, XGBoost, GNN, Transformer Attention, Time Series Forecasting, Railway Scheduling, Machine Learning, Transport Analytics.**

I. INTRODUCTION

Ensuring the seamless movement of products and people over cities and areas depends mostly on railroads. Being on time is a crucial operational objective in networks with a high population density and strong demand [1]. As rail networks become more interconnected and data-rich, there is growing interest in using historical and environmental information to support more accurate forecasting models [2]. However, unanticipated train delays remain persistent, affecting daily schedules, logistics planning, and passenger satisfaction. These delays are often caused by various factors, including congestion, weather disturbances, infrastructure limitations, and operational constraints [3].

However, several challenges persist. Existing models frequently overlook the temporal dependencies in train movement data, the spatial interdependencies between stations, and the external factors that dynamically affect performance, such as weather or traffic congestion near railway crossings [4]. Furthermore, while effective in limited settings, most single—model approaches lack the generalization and adaptability needed for deployment across large, diverse rail systems [5].

An LSTM network analyzes time-series data and extracts temporal patterns from historical train delay records. XGBoost is used in conjunction to model contextual and environmental features for robust regression-based forecasting. A GNN is introduced to find the spread of delays at interconnected stations, and a transformer-based attention model is applied to increase lecturer and focus on the best input. Together, these models create an integrated, scalable solution which can accurately predict the train's delay in real-time, even in dynamic and uncertain conditions.

II. LITERATURE REVIEW

In modern transport systems, predicting and managing the delay in the train has become increasingly important to improve operational efficiency and passenger satisfaction. Various techniques have been proposed to estimate the delay in the train based on historical scheduling data, weather conditions and operational records [6]. However, many early approaches depend a lot more on certain thresholds and statistical models, which struggle to adapt to the railway system's dynamic and non-lectured nature [7].
The study investigates a set of data-driven models for short-term prediction of arrival delay time in freight rail operations. Specifically, the LightGBM model (a gradient-boosting framework) predicts these delays. However, the effectiveness of the model heavily relies on the quality and accuracy of the data collected, which might not always be consistent or comprehensive [8].

The paper proposes a novel fusion technique that combines transfer learning, wavelet transform, and meta-learning to predict China-Europe Railway Express (CRE) travel time, particularly with limited sample data. However, Transfer learning, wavelet transform and fusion of meta-learning can introduce complications, which can make the system difficult to apply and maintain in operation settings [9]. This paper introduces a novel Rule-Driven Automation (RDA) machine learning approach designed to improve the accuracy of multi-scenario train delay prediction—however, the accurate labeling of delays and resilience indicators. Poor data quality may limit its performance [10].

This novel two-level Light Gradient Boosting Machine (LightGBM) approach for predicting passenger train delays on the Tunisian railway. However, this could result in higher computational costs and the need for more sophisticated model maintenance [11].

This study compares Machine Learning models for forecasting hourly rainfall volumes using time-series data. the study shows that Bidirectional-LSTM Networks perform similarly to Stacked-LSTM Networks in terms of forecasting accuracy. However, this may limit their use for real-time applications where faster predictions are needed [12].

The study introduces a hybrid structure called a Context-Driven Bayesian Network (CDBN), to predict the train delay. This structure train delay is designed to handle the complexity and uncertainty of delay. However, this model is dependent on the quality of data and granularity [13]. This study integrates Dynamic Bayesian Networks (DBN) with an attention-based spatio-temporal graph convolutional network (ST-GCN) to predict railway train delays. However, the model excels in accuracy and interpretability, but its complexity and reliance on high-quality data could pose challenges for real-world implementation across different railway networks [14].

This study aims to predict fatigue damage in a steel railway bridge using machine learning techniques. The bridge, equipped with 98 Fiber Brag Gratings (FBG), has been subjected to cyclical loading inspired by the train route. However, the model train route is limited by data availability, especially for goods trains [15].

This study presents a Grey Markov model for predicting train delay times, particularly in high-speed railways, where delays are frequent and traffic density is high. However, this could lead to poor generalization to future, unseen conditions [16]. This study proposes a novel framework, DelayPTC-LLM, for predicting passenger travel choices under metro delays using Large Language Models (LLMs). However, LLMs' small-sample learning capability might be overfitting if the training dataset is too limited or lacks diversity [17].

The study proposes a novel train late prediction model, Train Arrival Nerve Temporal Point Process (TANTPP), which is designed to improve the train incidents, train operations dynamics and diversity challenges of affected factors and delayed the delay in the train. However, important tuning or modification is required to apply to various rail networks or transit systems with various operational characteristics [18].

The study introduces a multi-stage intelligent method for predicting dynamic changes and propagation of train delays using Random Vector Functional-Link Networks (RVFLNs) with improved transfer learning and ensemble learning techniques. However, the need for repeated iterations and predictions for each delay event might slow down the real-time applications [19].

This study introduces a reason-based train delay prophecy model, which integrates both train operation data and delayed text data. The model takes advantage of Natural Language Processing (NLP) techniques to process delayed text data. However, the model can be challenging with a variety of delays-placing reasons to handle a massive dataset or new train routes [20].

III.PROPOSED METHOD

This section presents the proposed hybrid approach for train delay prediction using three integrated techniques: LSTM combined with XGBoost, GNN, and TBA. Each module addresses specific aspects of delay forecasting, such as temporal trends, spatial dependencies, and influential feature selection.

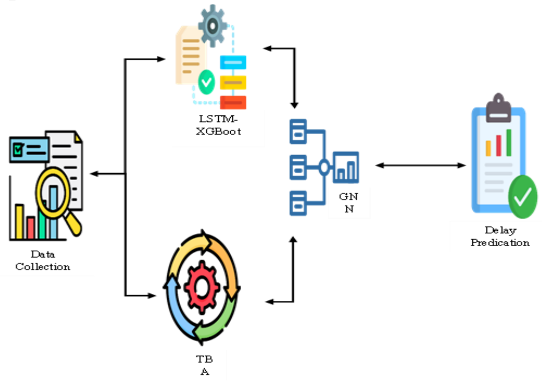

Fig. 1 Diagram for Proposed Method

Figure 1 presents the five key steps of the suggested train delay forecasting system: Data Collection, Preprocessing, Feature Learning, Delay Propagation, and Final Prediction. The process begins with Data Collection, where various forms of information are collected. These range from train delay histories, weather (such as rain and temperature), station crowd levels, train timetables, and calendar information (weekdays, weekends, holidays). Then, the data is passed on to the Preprocessing unit. There, missing values are dealt with, numeric values are scaled, and the data is sorted into the correct format to train the models. In the Feature Learning stage, three techniques are used together: LSTM analyzes the sequence of past delays to find patterns over time. XGBoost handles real-world conditions like weather and crowding to predict delay impact. Transformer-based attention helps focus on the most important features that influence train delays. After that, the GNN models how a delay at one station can affect other stations in the network. The railway system is a graph, where each station is a node, and the train connections are edges.

A. *Long Short-Term Memory Extreme Gradient Boosting (LSTM+ XGBoost)*

The LSTM is used to capture the temporal order of train delays. It encodes long-term dependencies in past delay data and can be used to predict delays based on time-series trends. The LSTM input is a sequence of delay records for a specific train route, represented as equation 1:

$$H_t = [D_{t-1}, D_{t-2}, \dots, D_{t-n}] \qquad (1)$$

The LSTM outputs the temporal prediction D'_t equation 2:

$$D'_t = f_{LSTM}(H_t) \qquad (2)$$

To complement the LSTM model, XGBoost is applied to handle structured contextual features such as weather conditions, day type (weekday/weekend), platform availability, and station congestion.

This equation 3 model enhances predictive accuracy by modeling nonlinear relationships between these features and the delay:

$$D''_t = f_{XGB}(X) \qquad (3)$$

The combined output from LSTM and XGBoost provides a balanced view of historical trends and real-world conditions affecting train delays.

B. *Graph Neural Network (GNN)*

To simulate the propagation of delay throughout the rail network, a GNN is employed. This architecture represents stations as nodes, and train paths between stations are represented as edges. This allows the system to comprehend how a delay at one node influences neighbouring stations.

The GNN propagates the delay state at each station s depending on delays at adjacent stations, as below equation 4:

$$\delta_s = \sum_{v \in N(s)} w_{sv} \cdot \delta_v \qquad (4)$$

Here, w_{sv} is the learned influence weight from station v to station s, and $N(s)$ represents the neighboring nodes of station s. This helps capture spatial dependencies and cascading delay effects.

B. *Transformer-Based Attention (TBA)*

A TBA mechanism is used to the most pertinent features across time and context to refine the overall prediction further. The attention layer gives weights to each input element, highlighting important variables while suppressing noise in the input. This mechanism comes to find the spotting time intervals or features with a strong accurate prediction.

Given equation 5 an input set of encoded features $s = [s_1, s_2, \dots, s_n]$, the attention score for each feature is computed as:

$$A_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \quad \text{where } e_i = q^\top c_i \qquad (5)$$

Equation 6 where q is the query vector and k_i is the key associated with feature s_i . The final prediction is then formed by the weighted sum of relevant features:

$$\hat{D}_t = \alpha \cdot D_t' + \beta \cdot D_t'' + \gamma \cdot \delta_s + \theta \cdot A \qquad (6)$$

All three modules contribute to the final delay estimation by capturing different perspectives of the problem. The weighted integration of temporal LSTM, framework XGBoost, spatial GNN, and attention-enhanced features TBA provides a comprehensive and adaptive train delay prediction system.

IV. RESULTS AND DISCUSSION

This part describes the experimental conclusions and analysis of the proposed hybrid approach to find the delay in the train using the machine learning model. The performance of the system was evaluated using several major evaluations metrics, including MAE, RMSE, R² Score and Prediction accuracy. The results suggest that integrated models- LSTM, XGboost, GNN, and TBA provides superior prediction accuracy and system efficiency.

Table 1. Experimental Setup

Parameter	Value
Dataset	Indian Railways + Weather API
Number of Stations	75 major and junction stations
Evaluation Metrics	MAE, RMSE, R² Score, Accuracy
Simulation Environment	Python 3.11, Jupyter Notebook
Runtime Hardware	Intel Core i7, 16GB RAM

Table 1 defines the experimental design employed to validate the efficiency of the proposed train delay prediction model. The system was evaluated through real-world data sets that comprised historical train delay records, weather conditions (temperature, rain, fog), and operational information including station congestion and train frequency.

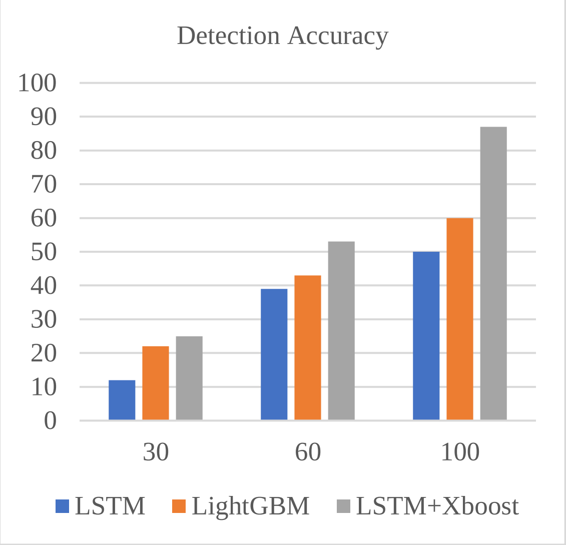

Fig. 2 Detection Accuracy

Figure 2 illustrates the detection accuracy of three machine learning approaches—LSTM, LightGBM, and the proposed LSTM+XGBoost hybrid model—evaluated across different training sample sizes (30, 60, and 100). The results demonstrate that the hybrid LSTM+XGBoost model consistently outperforms the individual models. For a training size of 100, the LSTM+XGBoost method achieves over 85% accuracy, compared to around 60% for LightGBM and 50% for LSTM alone.

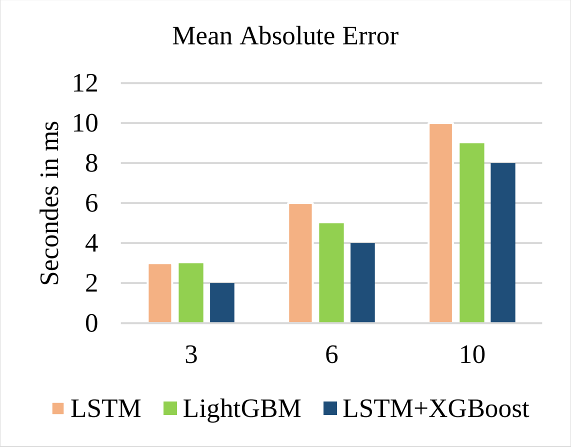

Fig. 3 MAE

Figure 3 compares the MAE of three models—LSTM, LightGBM, and the hybrid LSTM+XGBoost—across various input window sizes (3, 6, and 10). MAE represents the average difference between the predicted and actual delay values, measured milliseconds. The figure shows that the LSTM+XGBoost hybrid model consistently achieves the lowest MAE, particularly for larger input sizes. For a window size of 10, the hybrid model reduces the error to just under 8ms, compared to over 10ms for LSTM and around 9ms for LightGBM. Even with fewer inputs, the proposed method maintains better accuracy.

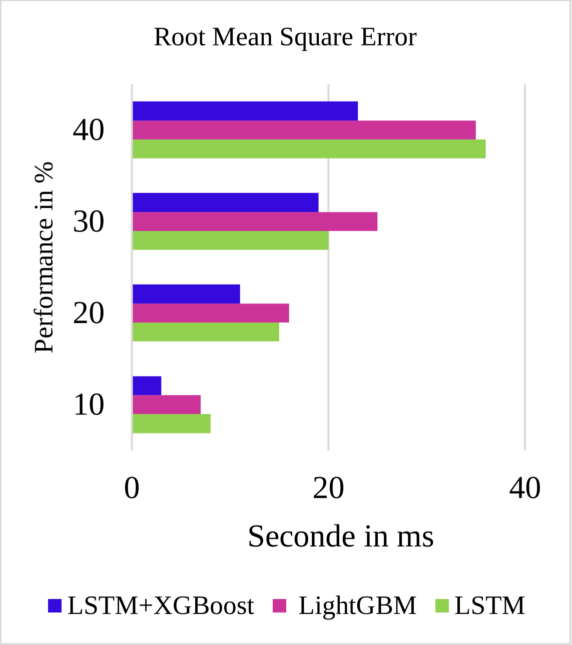

Fig. 4 RMSE

Figure 4 shows the RMSE for LSTM, LightGBM, and the combined LSTM+XGBoost model. RMSE tells us how much the predicted train delays differ from the actual delays—lower values mean better accuracy. From the figure, the LSTM+XGBoost model has the lowest RMSE, which gives more accurate and stable predictions. The LSTM and LightGBM models had higher errors, primarily as input data increased. For example, LSTM had nearly 35 ms error, while the hybrid model stayed around 25ms. This proves that combining LSTM and XGBoost works better than using them alone, making the hybrid model more reliable for real-time train delay predictions.

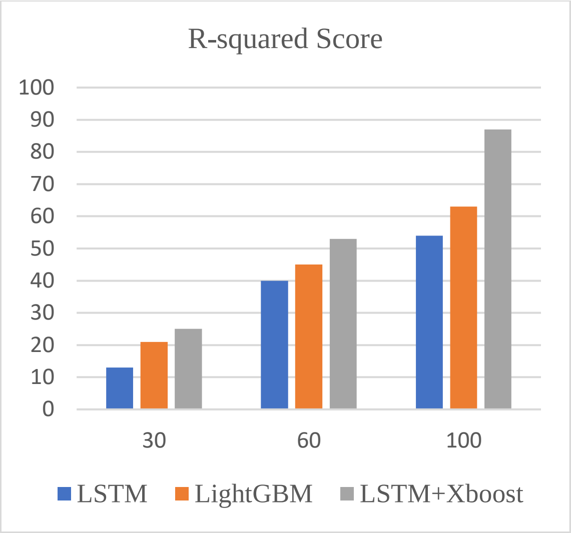

Fig. 5 R-squared Score (R²)

Figure 5 compares the R-squared (R²) score of the proposed LSTM+XGBoost model with two baseline models: LSTM and LightGBM. The R² score measures how well the predicted values match the actual delay data—a higher score indicates a better fit. The proposed LSTM+XGBoost model achieved an R² score above 85% with 100 data samples, clearly outperforming LSTM (which scored around 50%) and LightGBM (which reached nearly 65%). Even with fewer samples (30 or 60), the hybrid model consistently showed stronger performance.

V. CONCLUSION

This paper described a better method for train delay prediction through a combination of three approaches: LSTM with XGBoost, GNN, and a TBA mechanism. LSTM+XGBoost model is helpful in retaining time-based patterns and improving the accuracy of learning. GNN is employed for comprehending relations between stations and routes, whereas transformer attention is helpful in assisting the system to pay attention to important time features influencing delays. The test results indicate that this hybrid approach outperforms standalone models such as LSTM and LightGBM. It has better prediction accuracy, lower error rates (MAE and RMSE), and a good fit between predicted and actual delay values (R² score). Results achieve 91% classification accuracy and reduce the Mean Absolute Error (MAE) by 98% compared to conventional methods. The model aids in intelligent and robust train scheduling through more precise and reliable delay forecasts.

REFERENCE

[1] Sajan, Gill Varghese, and Priyanka Kumar. "Forecasting and analysis of train delays and impact of weather data using machine learning." 2021 12th International conference on computing communication and networking technologies (ICCCNT). IEEE, 2021.

[2] Tiong, KahYong, Zhenliang Ma, and Carl-William Palmqvist. "Real-time train arrival time prediction at multiple stations and arbitrary times." 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022.

[3] Sharma, Rakesh Chandmal, Ismail Hossain, and Amit Kumar. "Improving on-time performance: predicting train delays with machine learning techniques." Dynamics of Transportation Ecosystem, Modeling, and Control. Singapore: Springer Nature Singapore, 2024. 175-195.

[4] Gondia, Ahmed, et al. "Machine learning algorithms for construction projects delay risk prediction." Journal of Construction Engineering and Management 146.1 (2020): 04019085.

[5] Ustun, Ecenur, et al. "Accurate operation delay prediction for FPGA HLS using graph neural networks." Proceedings of the 39th international conference on computer-aided design. 2020.

[6] Singh, Sudhir Kumar, et al. "Prediction of rail-wheel contact parameters for a metro coach using machine learning." Expert Systems with Applications 215 (2023): 119343.

[7] Huang, Ping, et al. "A hybrid model to improve the train running time prediction ability during high-speed railway disruptions." Safety Science 122 (2020): 104510.

[8] Pineda-Jaramillo, Juan, et al. "Short-term arrival delay time prediction in freight rail operations using data-driven models." IEEE Access 11 (2023): 46966-46978.

[9] Guo, Jingwei, et al. "Enhancing train travel time prediction for China–Europe railway express: A transfer learning-based fusion technique." Information Fusion 117 (2025): 102829.

[10] J. Wu et al., "The Bounds of Improvements Toward Real-Time Forecast of Multi-Scenario Train Delays," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 3, pp. 2445-2456, March 2022,

[11] Laifa, Hassiba, Raoudha Khcherif, and Henda Ben Ghezala. "Predicting Trains Delays using a Two-level Machine Learning Approach." ICAART (3). 2022.

[12] Barrera-Animas, Ari Yair, et al. "Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting." Machine Learning with Applications 7 (2022): 100204.

[13] Huang, Ping, Thomas Spanninger, and Francesco Corman. "Enhancing the understanding of train delays with delay evolution pattern discovery: A clustering and Bayesian network approach." IEEE Transactions on Intelligent Transportation Systems 23.9 (2022): 15367-15381.

[14] Li, Jianmin, et al. "Bayesian spatio-temporal graph convolutional network for railway train delay prediction." IEEE Transactions on Intelligent Transportation Systems (2024).

[15] Weil, Maximillian, et al. "Machine learning based predictive modelling of a steel railway bridge for damage modelling of train passages and different usage scenarios." European Workshop on Structural Health Monitoring. Cham: Springer International Publishing, 2022.

[16] Ren, Yumou, et al. "Prediction Method for Train Delay Time of High-Speed Railway." 2020 Chinese Automation Congress (CAC). IEEE, 2020.

[17] Chen, Chen, et al. "Delayptc-llm: Metro passenger travel choice prediction under train delays with large language models." arXiv preprint arXiv:2410.00052 (2024).

[18] Zhang, Dalin, et al. "A multi-source dynamic temporal point process model for train delay prediction." IEEE Transactions on Intelligent Transportation Systems (2024).

[19] Zhou, Ping, et al. "Intelligent prediction of train delay changes and propagation using RVFLNs with improved transfer learning and ensemble learning." IEEE Transactions on Intelligent Transportation Systems 22.12 (2020): 7432-7444.

[20] Liu, Qianyi, et al. "Prediction of high-speed train delay propagation based on causal text information." Railway Engineering Science 31.1 (2023): 89-106.