Infrastructure – Aware Railway Delay Prediction Using Machine Learning and Civil Engineering Features

Lenin Dhal[1*], M. Mariappan[2*]

[1] Assistant Professor, Department of Civil Engineering, Easwari Engineering College, Chennai – 600089, Tamil Nadu, India.
Email: llenindhal@gmail.com

[2] Associate Professor, Department of Civil Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi – 642003, Tamil Nadu, India.
Email: mariappan.m2009@gmail.com

Abstract: Predicting railway delays is crucial because they can impact operations, lower passenger satisfaction, and pose safety risks. A machine learning system that forecasts train delays and identifies potential technical causes is presented in this study. Six thousand train records with sixteen features—train type, journey distance, schedule time, speed, bridge type, soil condition, season, flood risk, and rail vibration—were used to train the model. A variety of machine learning models, such as Support Vector Regressor, Random Forest, Gradient Boosting, Decision Tree, and Linear Regression, were examined. With a mean absolute error of 1.45 minutes and an R2 score of 0.95, the Gradient Boosting model produced the best results out of all of them, attaining an accuracy of almost 95%. A rule-based module of the system also describes potential engineering reasons for delays, including bridge speed limitations, soil settlement challenges, flood hazards, heat-induced rail expansion, and track vibration problems. To assist railway employees in making better decisions on traffic control and maintenance, the anticipated delays are divided into Low, Medium, and High categories. All things considered, this method improves railway reliability and allows for real-time railway monitoring by combining precise prediction with understandable explanations..

Keywords: Railway Delay Prediction, Machine Learning, Intelligent Transportation Systems.

INTRODUCTION

The both domestic and regional travel, railroad transit is crucial. It offers large-capacity, reasonably priced, and energy-efficient passenger and cargo transportation. Railways are essential to industry, tourism, social interaction, and economic progress in nations like India. Millions of people travel locally and across vast distances on trains every day, and many industries rely on freight trains to move goods and raw materials. As a result, safety, dependability, and timeliness are essential components of an effective railway system. However, despite recent advancements, train delays continue to be a prevalent issue. Passenger discontent, subpar operations, increased fuel consumption, personnel scheduling problems, and interruptions on linked routes can all result from delays. Even a minor delay can have a rapid impact on numerous trains and stops on congested railway networks. Complex operations and shifting infrastructure and environmental circumstances have made managing delays more challenging.

The mainstays of traditional railway systems are manual monitoring and set timetables. These techniques work well for simple planning, but they are unable to address challenges that arise in real time, such as weather variations, track deterioration, seasonal impacts, or unforeseen technical difficulties. Additionally, rather than anticipating and mitigating delays, the majority of typical delay analysis is carried out after they occur. It is more difficult to prevent significant disruptions with this reactive strategy. This describes how the railway delay prediction and engineering diagnostic system was designed, constructed, and tested. From gathering the data to getting it ready for analysis, it explains every stage of the process in detail. To deal with missing values and eliminate errors, the data is cleaned and processed. To make them..

understandable to machine learning models, key features are chosen and formatted appropriately. Various machine learning models are created and evaluated to forecast train delay time following preprocessing. To guarantee accurate and trustworthy outcomes, their performance is thoroughly assessed using the appropriate criteria. The system is more reliable, consistent, and simple to replicate when the technique is well-defined and structured. The suggested approach adheres to a sequential pipeline. Data preprocessing and feature encoding come after dataset gathering. The prepared dataset is then used to train and test the models. After that, the system classifies delays and determines potential engineering reasons for them. Lastly, the results are displayed in an understandable manner using clear visualizations. Every step is intended to handle the intricacy of railway operational data and identify significant trends that raise prediction accuracy. In order to support the system's performance, this chapter also outlines the hardware and software requirements, specifies the various system modules, explains the algorithm workflow, enumerates the tools utilized, and presents experimental findings..

significance

A lot of railway data is now gathered by sophisticated technologies like sensors and digital management systems. We now have information on soil conditions, weather, bridge types, train speeds, track vibration, and previous delays. Artificial intelligence and machine learning are also capable of analyzing big datasets and identifying intricate patterns. This presents a chance to develop sophisticated systems that can precisely pinpoint the reasons of train delays and forecast them.

Using operational, environmental, and infrastructural data, this research suggests a machine learning-based approach to forecast train delay duration and identify potential engineering causes. Prediction models and a rule-based diagnostic module are used in the system to deliver precise results and understandable technical explanations. In order to assist railway officials in making better decisions on traffic control, maintenance, and safety, it also divides delays into Low, Medium, and High categories. The backdrop of railway delay prediction, the issue addressed, the project goals, the extent of the work, and the report format are all explained in the remaining sections of this chapter.

1.2 Data collection

16 distinct features and 6,000 train records make up the system's well-structured dataset. Train ID, train type, source and destination stations, distance traveled, departure and arrival times, and average speed are some of these features. Information about bridge type, soil type, season, flood danger, vibration level, and actual and present delays are also included. All the dataset were taken from the NTES as well as manual also. These characteristics together characterize the infrastructure, environmental, and operational elements that influence railway operations.

1.3 Data Preprocessing Module

Numerical values with varying ranges, category-based information, and occasionally errors or missing data are all included in raw railway data. Missing values are dealt with during preprocessing in order to complete the data. Label encoding is used to translate categorical characteristics, such as soil type, season, train type, and bridge type, into numerical values. To help the model learn, numerical features like vibration intensity, speed, distance, and delay time are standardized. In order to assess the model's performance, the dataset is finally separated into training and testing sets.

1.4 Predictive Modelling Module

The length of a train's delay is predicted using a number of supervised regression models. These consist of Support Vector Regression, Gradient Boosting Regression, Random Forest Regression, Decision Tree Regression, and Linear Regression. The cleaned and prepared dataset is used to train each model, and common assessment criteria are used to gauge each model's performance. Because they can comprehend intricate and non-linear correlations between the features, Random Forest and Gradient Boosting outperform the others.

1.5 Delay Classification Module

The predicted delay duration is categorized into three operational classes:

* Low Delay: Less than 8 minutes

* Medium Delay: Between 8 and 18 minutes

* High Delay: Greater than 18 minutes

This classification simplifies interpretation and enables railway operators to prioritize corrective actions based on severity levels.

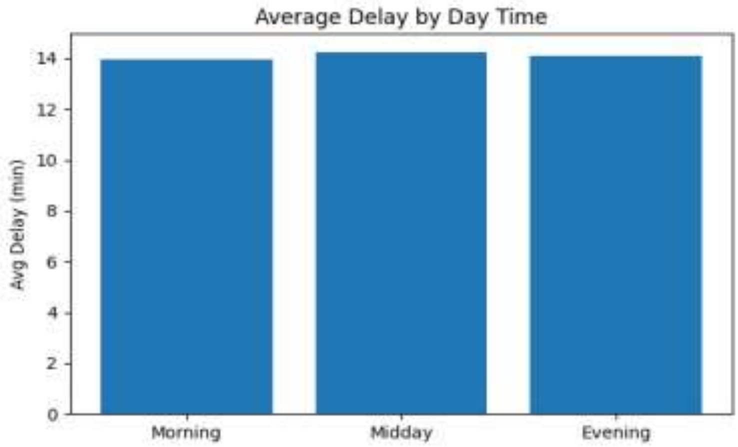

Figure 1 Average Delay Graph

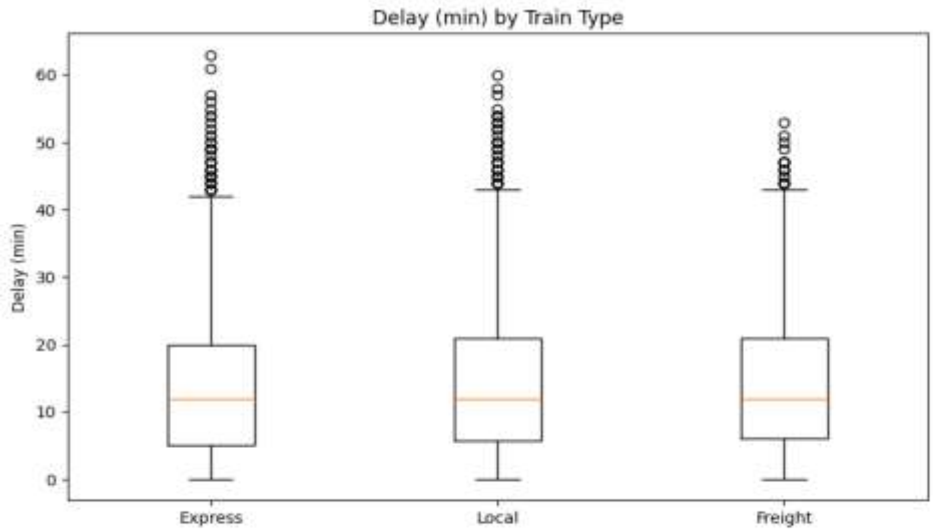

Figure 2 Delay by Train type

2.1 Modular Identification

The proposed system is divided into several functional modules to ensure modularity, clarity, and ease of maintenance. The major modules identified in the system are:

1. Data Collection Module
2. Data Preprocessing Module
3. Feature Encoding and Normalization Module
4. Model Training and Testing Module
5. Delay Classification Module
6. Engineering Diagnostic Module
7. Result Visualization Module

Each module performs a specific task in the overall pipeline and interacts with other modules through well-defined interfaces.

2.2 Hardware Specifications

The implementation and experimentation of the proposed system were carried out on a standard personal computing platform with the following configuration:

* Processor: Intel Core i5 / i7 or equivalent
* Clock Speed: 2.5 GHz or higher
* RAM: 8 GB minimum (16 GB recommended)
* Storage: 256 GB SSD / 1 TB HDD

* Graphics: Integrated / Dedicated GPU (optional)

* Operating System: Windows 10 / Linux

These specifications are sufficient to handle medium-scale datasets and to train ensemble machine learning models efficiently.

2.3 Software Specifications

* Programming Language: Python 3.8 or higher

* Development Environment: Jupyter Notebook / VS Code

* Libraries and Framework: NumPy for numerical computation * Pandas for data manipulation and preprocessing.

* Scikit-learn for machine learning algorithms * Matplotlib and Seaborn for visualization

These tools provide flexibility, scalability, and ease of implementation for data-driven modeling and experimentation.

3.1 model design

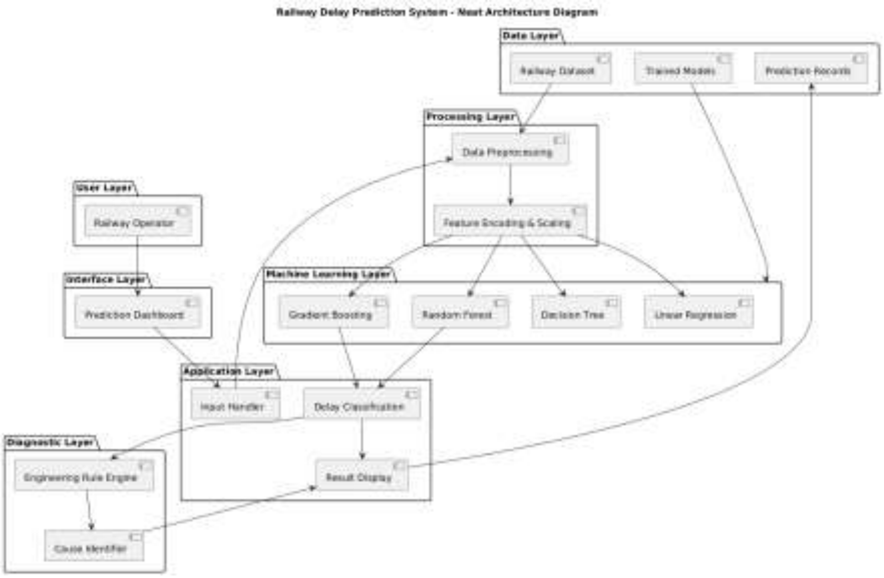

Figure 3 Architecture Diagram

An essential step in preparing raw data for model training is the preprocessing module. In order to prevent repetition of the same data, which could skew the results, it first eliminates duplicate records. After that, it verifies that the data is accurate and consistent and looks for any missing numbers. Any missing or inaccurate values are handled appropriately. The data types are then transformed into the appropriate formats, such as converting dates into a standard format or text to numbers. Lastly, the target variable (the outcome the model must predict) is separated from the input characteristics (the data needed to make predictions). The entire procedure guarantees that the model is trained and enhances the quality of the data.

3.2 Normalaization Module

One can express the Laplacian, a scalar second- The various data kinds are prepared in this step so that the model can appropriately comprehend them. Since most machine learning models are in text form, they are unable to directly employ categorical data like soil type, season, train type, bridge type, etc. Therefore, a technique known as label encoding is used to transform them into numbers. As a result, every category is assigned a distinct number value. However, the ranges of numerical attributes like vibration levels, speed, distance, and delay minutes might change greatly. They are normalized—that is, scaled to a common range—to facilitate processing by the model. This enhances the model's stability during training, speeds up learning, and raises prediction accuracy

4.1 Testing and Training

A number of regression models are trained to generate predictions after the dataset has been prepared. Eighty percent of the data is utilized to train the models, and the remaining twenty percent is used to evaluate how well they perform. This procedure makes use of the following models: Support Vector Regressor, Gradient Boosting Regressor, Random Forest Regressor, Decision Tree Regressor, and Linear Regression. After identifying trends in the training data, each model forecasts the testing data. A variety of metrics are used to assess their success, including the R2 score, which indicates how well the model explains the data, the Mean Absolute Error, which displays the average difference between the predicted and actual values, and the Root Mean Square Error, which assigns greater weight to higher errors. Pattern Recognition. In order to recognize patterns and objects inside an image, features must be extracted. The RF processing technique and the image-processing approach are usually the two primary ways to tackle this problem. The target is identified using the RF processing approach by sending electromagnetic waves in its direction and analyzing the reflected waves. This technique is frequently used when the object is out of sight or at a significant distance away. In contrast, the image-processing method involves taking a picture or an image of the area, then processing and analyzing it to determine which things are there.

4.2 Engineering Diagonastic Module

The diagnostic module uses particular railway system principles to determine potential engineering causes for train delays. It closely looks at a number of crucial elements. For instance, for safety considerations, trains may need to slow down on some bridge types. Track stability may be impacted by ground settlement caused by certain soil types. Railway rails may expand or contract due to seasonal variations like intense heat or cold. Additionally, the module determines whether the route passes through flood-prone areas, where trains may be delayed by safety precautions. High vibration levels may also be a sign of potential track flaws or maintenance problems. The module creates a comprehensive list of probable technical reasons associated with each anticipated delay by examining each of these factors.

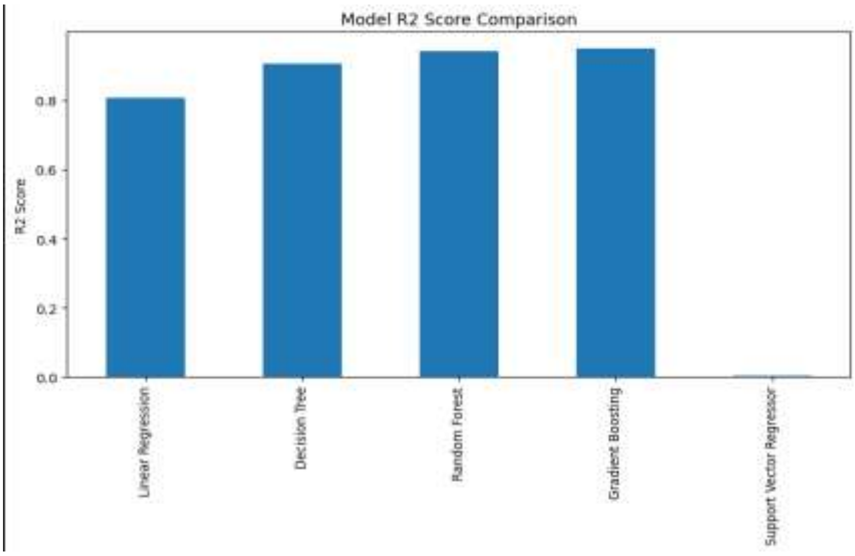

Figure 4 Algorithm Selection

5 results and discussions

There are sixteen distinct characteristics and 6,000 train records in the dataset utilized for this research. Train ID, train type, source and destination stations, distance in kilometers, scheduled departure and arrival times, average speed, and current delay minutes are some of the details that are included in these features. Bridge type, soil type, season, flood-prone indicator, vibration level, actual delay minutes, and delay explanation are among the environmental and infrastructure details that are included. After a thorough review, it was determined that there were no missing values in the data and that the major categories were evenly distributed. The goal variable for prediction is the actual number of delay minutes.

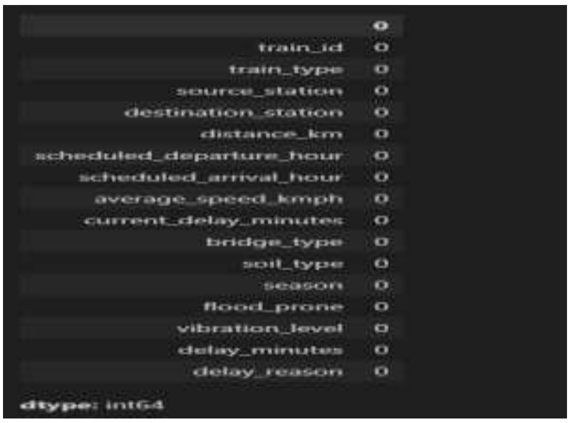

Figure 5 Data

```
sample_input_1 = pd.DataFrame([{
    "train_type": encoders["train_type"].transform(["Passenger"])[0],
    "distance_km": 520,
    "average_speed_kmph": 70,
    "current_delay_minutes": 0,
    "bridge_type": encoders["bridge_type"].transform(["river"])[0],
    "soil_type": encoders["soil_type"].transform(["clay"])[0],
    "season": encoders["season"].transform(["monsoon"])[0],
    "flood_prone": 1,
    "vibration_level": 9.2,
    "scheduled_departure_hour": 10,
    "scheduled_arrival_hour": 20
}])

predicted_delay_1 = round(model.predict(sample_input_1)[0], 2)

print("Predicted Delay Minutes:", predicted_delay_1)
print("Delay Class:", delay_class(predicted_delay_1))
print("Engineering Causes:", detect_engineering_causes(sample_input_1.iloc[0]))
```

```
Predicted Delay Minutes: 41.54
Delay Class: High
Engineering Causes: ['Bridge Speed Restriction', 'Soil Settlement Risk', 'Flood Safety Restriction', 'Track Vibration Issue']
```

Figure 6 Code

5.1 Algorithm and Pseudocode

1. Input railway dataset
2. Perform data preprocessing and cleaning
3. Encode categorical features using label encoding
4. Normalize numerical features
5. Split dataset into training and testing sets
6. Train regression models on training data
7. Evaluate models using MAE, RMSE, and R^2
8. Select best-performing model
9. Input new train record
10. Predict delay duration
11. Classify delay into Low, Medium, or High
12. Apply engineering diagnostic rules.

6 concluding remarks

Using machine learning techniques, an intelligent system was created in this project to forecast railway delays and determine their potential engineering reasons. The system combines a rule-based diagnostic module that explains the technical causes of the delays with regression models for precise delay prediction. Several models, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression, were trained and tested on a dataset of 6,000 train records that included operational, infrastructure, and environmental details Mean Absolute Error, Root Mean Square Error, and R2 score were used to gauge their performance. With an R2 score of almost 0.95, the Gradient Boosting model outperformed the others in terms of accuracy. To aid in improved decision-making, the method additionally divides delays into Low, Medium, and High categories. The diagnostic module also identifies potential causes, such as soil settlement, flooding hazards, bridge speed limitations, temperature effects on tracks, and vibration-related track problems The findings demonstrate the effectiveness of machine learning, particularly ensemble approaches, in managing intricate railway data. All things considered, the system offers railway authorities a workable and trustworthy means to enhance operating effectiveness, repair scheduling, and delay management..

REFERENCES

Hansen, I. A., and Pachl, J., "Railway Timetable & Traffic: Analysis, Modelling, Simulation," Eurailpress, Hamburg, Germany, 2008.

Kecman, P., and Goverde, R. M. P., "Predictive modelling of running and dwell times in railway traffic," *Transportation

Research Part C: Emerging Technologies*, vol. 50, pp. 1–15, 2015.

Zhou, Y., Wang, J., and Sun, H., "Train arrivaldelay prediction using support vector regression," Journal of

Advanced

Transportation*, vol. 2017, pp. 1–10, 2017.

Zhang, H., Wen, C., and Zhao, Y., "Prediction orailway delays using decision tree and random forest models," *IEEE

Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3331–3341, 2019.

Sun, L., Jin, J. G., and Axhausen,W.,"Understanding urban rail transit delays using data mining techniques,"

Transportation Research Part C: Emerging Technologies, vol. 98, pp. 135–151, 2019.

Li, X., Chen, Y., and Wang, Z., "Gradient boosting based railway delay prediction with weather and traffic data," *Expert

Systems with Applications*, vol. 173, pp. 114–125, 2021.

Rezaei, M., and Khademi, M., "Deep learning approach for short-term railway delay prediction," *Engineering

Applications of Artificial Intelligence*, vol. 104, pp. 104–118, 2022.

Chowdhury, A., et al. "RSTGCN: Railway-centric Spatio-Temporal Graph Convolutional Network for Train Delay

Prediction." (Focuses on large-scale, complex networks using graph-based AI).

Chen, et al. "Combining Transformer and 3D Convolutional Neural Networks on 1.7 million high-speed rail records from

Beijing." (Uses advanced AI for high-speed rail operational disruptions).

Huang, P., et al. "Modeling train timetables as images: A cost-sensitive deep learning framework for delay propagation

pattern recognition." Expert Systems with Applications

Bhattacharjee, S. "Railway delay predictor & letter generator system." Institute of Engineering and Management (Covers

feature engineering, random forest, and data preprocessing).

SCiOpen "Short-term train arrival delay prediction: a data-driven approach." (Analyzes, full data vs. late-arrival-only

data).

Liu, et al. "Real-Time delay prediction for single-track intercity railways." ScienceDirect (Uses SHAP analysis with random

forest for 2.53 min MAE accuracy).

Zhang et al. (2018): "Prediction of arrival times of freight traffic on US railroads using support vector regression".

A Review of Train Delay Prediction Approaches (Spanninger et al., 2022): Provides a comprehensive overview of 92 sources, covering both data-driven and event-driven methods