Machine learning model for railway reservation

Divith P
Department of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India
cddivith89@gmail.com

Shridhar Devaramane
Department of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India
devamaneshridhar@gmail.com

Vikas Shanabhog
Department of Computer Science and Engineering
Global Academy of Technology
Bengaluru, India
vikasshanabhog0@gmail.com

Mamatha R
Department of Computer Science and Engineering Global Academy of Technology Bengaluru, India
mamatharmanjunath@gmail.com

Shreyas M
Department of Computer Science and Engineering Global Academy of Technology Bengaluru, India
mshreyas336@gmail.com

Deeksha R
Department of Computer Science and Engineering Global Academy of Technology Bengaluru, India
deeksha.ramdas@gmail.com

***Abstract*— The evolution of train reservation systems—from manual booking to sophisticated digital platforms—is examined in this study. It looks at how important these systems are to improving operational effectiveness and passenger comfort. The article also explores future trends in railway reservation systems, issues encountered, and technology improvements.**

***Keywords*— Railway reservation, Train Passenger**

I. INTRODUCTION

The railway reservation system is a critical component of modern transportation infrastructure, facilitating the efficient movement of millions of passengers daily. In the digital age, the reliance on data-driven solutions has become paramount for optimizing operations and enhancing customer experience. This project aims to develop a robust railway reservation system utilizing a comprehensive dataset encompassing various aspects of railway operations, including train schedules, passenger information, seat availability, and ticketing details.

By leveraging this dataset, our objective is to design and implement a user-friendly application that streamlines the process of booking tickets, checking seat availability, and managing reservations. Through intelligent algorithms and data analysis, we seek to provide real-time updates on train status, optimize seat allocation, and offer personalized recommendations to passengers. Additionally, this system will incorporate features such as online payment integration, ticket cancellation, and booking modifications to ensure a seamless booking experience for users.

Furthermore, this project endeavors to contribute to the advancement of railway management systems by harnessing the power of data analytics and machine learning techniques. By analyzing historical booking patterns, passenger preferences, and travel trends, we aim to identify insights that can inform strategic decision-making and resource allocation for railway operators.

In summary, this railway reservation project represents a fusion of technology and transportation, with the overarching goal of enhancing efficiency, convenience, and customer satisfaction in the realm of railway travel. Through the utilization of a comprehensive dataset and innovative software solutions, we aim to redefine the passenger experience and set new standards for modern railway reservation system.

II. RELATED WORK

In [1], Comprehensively examines the integration of AI technologies in the railway sector. It highlights the significant roles of AI in enhancing operational efficiency, predictive maintenance, and safety. AI techniques such as machine learning and neural networks have been instrumental in optimizing train scheduling, route planning, and traffic management. Predictive maintenance, powered by AI, has reduced Downtime and maintenance costs.

In [2] Explores the application of Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) in maintaining high-speed railway power equipment. The study demonstrates how LSTM-RNN models effectively handle time-series data to predict equipment failures and recommend proactive maintenance measures. By accurately forecasting potential issues, the model reduces downtime and enhances equipment reliability. The paper includes case studies showing improved maintenance accuracy and efficiency using this approach. Challenges such as data quality, computational demands, and integration with current systems are discussed, along with future research directions to further optimize railway maintenance strategies.

In [3] explores the use of Genetic Algorithms (GA) to optimize train schedules. It addresses the complexity of scheduling by formulating it as an optimization problem with constraints like track availability and time windows. The GA is implemented with key operations including population initialization, fitness evaluation, selection, crossover, and mutation. Simulation results demonstrate that the GA significantly improves scheduling efficiency and reduces delays compared to traditional methods. The paper also presents case studies showing successful real-world applications.

In [4] Offers a comprehensive examination of various models and algorithms designed to manage railway traffic in real-time. The review emphasizes the need for robust systems to handle disruptions and maintain operational efficiency. It covers different dynamic models that capture the stochastic nature of railway networks and discusses optimization techniques and heuristics used to address real-time traffic issues. The paper compares online dynamic approaches with traditional offline methods, showcasing the benefits of real-time management in improving efficiency and passenger satisfaction.

In [5] Railway systems, station safety is a critical aspect of the overall structure, and yet, accidents at stations still occur. It is time to learn from these errors and improve conventional methods by utilising the latest technology, such as machine learning (ML), to Analyse accidents and enhance safety systems.

In [6] propose a novel approach to enhance the efficiency of reinforcement learning (RL) algorithms for train rescheduling. The paper addresses the challenge of limited sample efficiency and the need for effective multi-agent communication in complex railway systems. They introduce a hierarchical architecture that incorporates a centralized critic and decentralized actors to improve coordination among multiple agents.

In [7] The approach leverages floating train data, extracting key features to identify patterns preceding infrequent events. It addresses challenges like the rarity and complexity of these events, using advanced machine learning models tailored to the data. The study demonstrates improved prediction accuracy over traditional methods, highlighting the importance of feature selection. This methodology has practical implications for enhancing predictive maintenance, leading to reduced downtime, better resource allocation, and lower maintenance costs in the railway industry.

In [8] The study addresses the challenges posed by dynamic and unpredictable rail traffic, aiming to minimize delays and improve overall efficiency. They employ RL algorithms to learn optimal policies that can adapt to real-time changes in train movements. The approach involves modeling the rail network as a Markov Decision Process (MDP), where states represent the positions of trains and actions correspond to routing decisions. Their findings demonstrate that RL can significantly enhance the performance of traditional scheduling methods, offering a robust framework for real-time decision-making in railway operations.

In [9] The goal is to optimize the alignment of mountain railways, which are characterized by complex terrain and significant construction challenges. The authors utilize a stepwise optimization process to break down the alignment problem into manageable segments, allowing for detailed local adjustments. Hybrid PSO is then applied to these segments to explore a wide range of potential solutions efficiently. The incorporation of genetic operators, such as crossover and mutation, enhances the optimization by introducing diversity and avoiding local minima.

In [10] Current research in automatic train operation concentrates on optimizing an energy-efficient speed profile and designing control algorithms to track the speed profile, which may reduce the comfort of passengers and impair the intelligence of train operation. Different from previous studies, this paper presents two intelligent train operation (ITO) algorithms without using precise train model information and offline optimized speed profiles. The first algorithm, i.e., ITOe, is based on an expert system that contains expert rules and a heuristic expert inference method. Then, in order to minimize the energy consumption of train operation online, an ITOr algorithm based on reinforcement learning (RL) is developed via designing an RL policy, reward, and value function. In addition, from the field data in the Yizhuang Line of the Beijing Subway, we choose the manual driving data with the best performance as ITOm.

In [11] A collision avoidance system for railroad vehicles needs to determine their location in the railroad network precisely and reliably. For a vehicle-based system, that is independent from the infrastructure, it is vital to determine the direction a railroad vehicle turns at switches. In this paper a vision based approach is presented that allows to achieve this reliably, even under difficult conditions. In the images of a camera that observes the area in front of a railroad vehicle the rail tracks are detected in real-time. From the perspective of the moving railroad vehicle rail tracks branch and join from/to the currently travelled rail track. By tracking these rail tracks in the images, switches are detected as they pass.

In [12] The study uses historical passenger flow data to establish temporal and spatial relationships between different locations in the station, enabling more accurate and dynamic forecasting. The authors developed a prediction model that combines machine learning techniques with statistical analysis to capture the complexities and variability of passenger flow patterns. The model consider factors such as the time of day, day of the week, and specific events that may affect passenger numbers. By integrating these time-space correlations, the model can predict short-term fluctuations in passenger flow, which is crucial for managing congestion and improving the efficiency of subway operations. The findings demonstrate that considering spatial relationships between multiple sites enhances the accuracy of passenger flow predictions compared to traditional single-site models.

`

In [13] These clevises are critical components that connect various parts of the catenary structure, and their failure can lead to significant operational issues. The method employs a visual ensemble approach, combining multiple deep learning models to enhance detection accuracy and robustness.

The process begins with the extraction of clevis images captured by an inspection vehicle. A convolutional neural network (CNN), specifically a faster region-based CNN (R-CNN), is utilized to identify and isolate clevises from the background. Once isolated, the ensemble method, which integrates several deep learning models, is applied to detect fractures within the clevises.

In [14] Fasteners are crucial components that keep rails in place, and their failure can lead to derailments, making their inspection vital for railway safety. The authors propose a method that leverages Histograms of Oriented Gradients (HOG) features and a linear Support Vector Machine (SVM) classifier to detect and classify fasteners. This approach is particularly robust against clutter and background noise, which are common challenges in the railway environment. The system operates by scanning predefined regions of interest on railway ties, detecting fasteners, and then classifying them into categories such as good, missing, or broken.

In [15] The railway scheduling problem involves optimizing train timetables to minimize conflicts and ensure efficient utilization of track resources. The proposed genetic algorithm is designed to handle the NP-hard nature of the problem by evolving a population of potential solutions through processes analogous to natural selection, crossover, and mutation.

The study demonstrates how GAs can effectively explore the vast search space of possible schedules and iteratively improve solutions. The algorithm's performance is evaluated on various scheduling scenarios, showing significant improvements in timetable efficiency and conflict reduction compared to traditional methods. Additionally, the paper discusses the initial population generation strategies and fitness evaluation mechanisms crucial for the GA's success in solving real-world railway scheduling problems.

III. DATASETS

In the context of a railway reservation system, robust and comprehensive datasets are crucial for developing, testing, and validating various functionalities such as booking, scheduling, and customer service optimization. This section delineates the types of datasets employed in a railway reservation project, outlining their structure, sources, and significance in ensuring an effective and reliable system.

1) Passenger Information Dataset

This dataset contains personal and demographic information about passengers. It includes fields such as:

Passenger ID: A unique identifier for each passenger.

Name: Full name of the passenger.

Age: Age of the passenger.

Gender: Gender of the passenger.

2) Booking and Reservation Dataset

This dataset captures the details of ticket bookings and reservations. Key fields include:

Booking ID: A unique identifier for each booking.

Passenger ID: Links to the Passenger Information dataset.

Train Number: The train for which the booking is made.

Date of Journey: The scheduled date of travel.

Booking Date: The date when the reservation was made.

Class: The travel class (e.g., sleeper, AC, first class).

Seat Number: The allocated seat or berth number.

Booking Status: Status of the booking (e.g., confirmed, waitlisted, cancelled).

This dataset is crucial for managing seat availability, planning, and revenue management.

3) Train Schedule Dataset

The train schedule dataset provides comprehensive details about train operations. It includes:

Train Number: A unique identifier for each train.

Train Name: The name of the train.

Source Station Code: The station code where the train originates.

Station Code: The station code where the train terminates.

Departure Time: Scheduled departure time from the source station.

Arrival Time: Scheduled arrival time at the destination station.

Intermediate Stops: List of all intermediate stops with arrival and departure times.

Frequency: Days of operation (e.g., daily, weekdays, weekends).

Accurate scheduling data ensures timely operations and effective passenger communication.

4) Operational Dataset

This dataset includes operational metrics and performance indicators:

Train Number: A unique identifier for each train.

Date: Date of operation.

Punctuality: On-time performance data.

Delays: Details of delays, if any, including reasons.

Cancellations: Records of cancelled services with reasons.

5) Station Information Dataset

This dataset contains information about the railway stations, including:

Station Code: A unique code assigned to each station.

Station Name: The name of the station.

Location: Geographic coordinates (latitude and longitude).

Facilities: Available facilities at the station (e.g., waiting rooms, restrooms, food courts).

Zone/Division: The railway zone or division to which the station belongs.

Station information is pivotal for journey planning and enhancing passenger experience.

IV. ALGORITHMS

A decision tree is a powerful and intuitive machine learning algorithm used for both classification and regression tasks. It operates by mimicking the human decision-making process, which involves breaking down a complex decision into a series of simpler decisions. This model consists of nodes representing decisions, outcomes, or chances, connected by branches that outline the decision paths.

A decision tree typically has three types of nodes: the root node, internal nodes, and leaf nodes. The root node represents the entire dataset and is the starting point of the decision-making process. Internal nodes are decision points within the tree where the data is further split based on specific features. Finally, leaf nodes are terminal nodes that provide the final decision or outcome, which could be a class label for classification tasks or a continuous value for regression tasks.

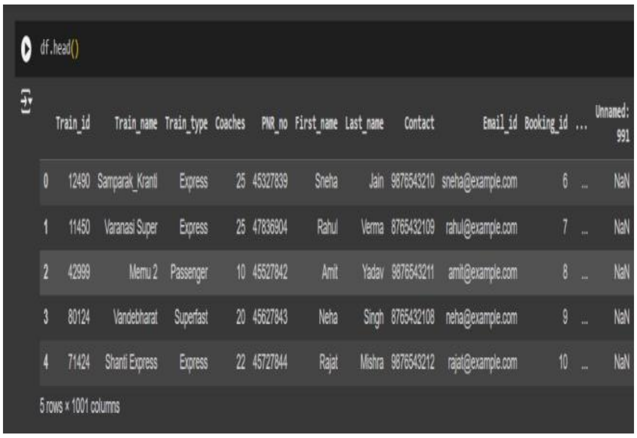
Building a decision tree involves selecting the best feature to split the data at each node. This selection is based on criteria such as Gini impurity, information gain, or variance reduction. Gini impurity measures the frequency of incorrect labeling if a random element was labeled according to the distribution of labels in the subset. Information gain, based on entropy, quantifies the reduction in uncertainty about the target variable after the dataset is split on a particular feature. Variance reduction is used in regression trees and measures the reduction in variance after splitting. Once the best feature is selected, the dataset is split into subsets, and the process is repeated recursively for each subset until a stopping criterion is met, such as maximum depth, minimum samples required to split, or minimum samples in a leaf node.

Decision trees offer several advantages. They are highly interpretable and easy to visualize, making them useful for understanding the model's decision-making process. They can capture non-linear relationships between features and the target variable, and they provide insights into the importance of different features in predicting the target variable. Additionally, decision trees require minimal data preprocessing, as they do not require normalization or scaling of data.

Decision trees are widely used in various applications due to their simplicity and interpretability. In medical diagnosis, they help classify patients based on symptoms and test results. In credit scoring, they evaluate the creditworthiness of loan applicants. In marketing, they segment customers based on behavior and preferences, enabling targeted marketing strategies.

Decision trees are a versatile and widely-used machine learning algorithm that provides a clear and interpretable model structure. While they have limitations, such as the tendency to overfit and instability, techniques like pruning and ensemble methods (e.g., Random Forests) can mitigate these issues. Overall, decision trees remain a valuable tool in the data scientist's toolkit for both classification and regression tasks.
Decision trees are widely used in various applications due to their simplicity and interpretability.

V. RESULTS



```
df.head()

   Train_id     Train_name Train_type Coaches    PNR_no First_name Last_name     Contact          Email_id Booking_id ... Unnamed:
                                                                                                                              991
0     12490 Samparak_Kranti    Express      25  45327839      Sneha      Jain  9876543210 sneha@example.com          6 ...      NaN
1     11450  Varanasi Super    Express      25  47836904      Rahul     Verma  8765432109 rahul@example.com          7 ...      NaN
2     42999          Memu 2  Passenger      10  45527842       Amit     Yadav  9876543211  amit@example.com          8 ...      NaN
3     80124     Vandebharat  Superfast      20  45627843       Neha     Singh  8765432108  neha@example.com          9 ...      NaN
4     71424   Shanti Express   Express      22  45727844      Rajat    Mishra  9876543212 rajat@example.com         10 ...      NaN

5 rows × 1001 columns
```

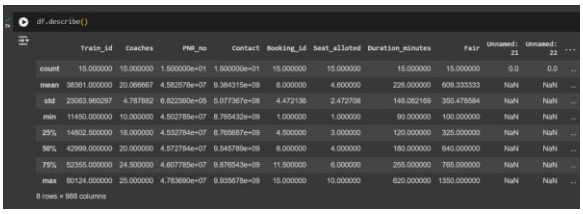
The table consists of several attributes related to train reservations, each serving a specific purpose in the context of managing and tracking bookings.

The **train_id** is a unique identifier assigned to each train, ensuring distinct recognition within the database. Complementing this is the **train_name**, which provides the official name of the train, aiding in easy identification by passengers and staff. The **train_type** attribute categorizes the train based on its service type, such as express, local, or freight, which is crucial for operational logistics and passenger information.

The **coaches** attribute details the number and types of coaches attached to the train, such as sleeper, AC, or general, providing essential information for both inventory management and seat allocation. The **pnr_no** (Passenger Name Record number) is a unique identifier for each booking, enabling easy tracking and retrieval of reservation details.

Passenger information is captured through several attributes. **First_name** and **last_name** record the personal names of passengers, ensuring that reservations are accurately associated with individual travelers. The **contact** attribute stores the passenger's phone number, facilitating direct communication for updates or emergencies. The **email_id** captures the passenger's email address, used for sending electronic tickets, booking confirmations, and other correspondence.

The **booking_id** is another unique identifier but is specific to the reservation transaction. It helps in managing and referencing individual bookings within the system, ensuring that each booking is distinct and traceable. Together, these attributes create a comprehensive dataset that supports efficient train reservation management, ensuring accurate tracking of trains, coaches, and passenger bookings, while also facilitating effective communication with travelers.

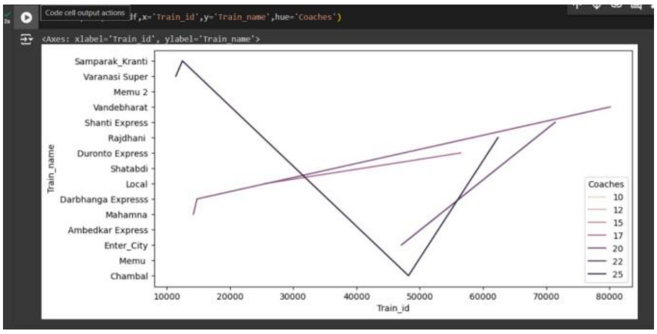


```
df.describe()

          Train_id    Coaches        PNR_no       Contact  Booking_id  Seat_alloted  Duration_minutes         Fair  Unnamed:  Unnamed:  ...
                                                                                                                          21        22
count    15.000000  15.000000  1.500000e+01  1.500000e+01   15.000000     15.000000         15.000000    15.000000       0.0       0.0  ...
mean  38361.000000  20.066667  4.582578e+07  9.384315e+09    8.000000      4.600000        226.000000   609.333333       NaN       NaN  ...
std   23063.960297   4.787882  6.822360e+05  5.077367e+08    4.472136      2.472708        146.062169   350.478584       NaN       NaN  ...
min   11450.000000  10.000000  4.502785e+07  8.765432e+09    1.000000      1.000000         90.000000   100.000000       NaN       NaN  ...
25%   14802.500000  18.000000  4.532784e+07  8.765687e+09    4.500000      3.000000        120.000000   325.000000       NaN       NaN  ...
50%   42999.000000  20.000000  4.572784e+07  9.545789e+09    8.000000      4.000000        180.000000   640.000000       NaN       NaN  ...
75%   52355.000000  24.500000  4.607785e+07  9.876543e+09   11.500000      6.000000        255.000000   765.000000       NaN       NaN  ...
max   80124.000000  25.000000  4.783690e+07  9.935678e+09   15.000000     10.000000        620.000000  1350.000000       NaN       NaN  ...

8 rows × 988 columns
```

The table includes attributes crucial for managing seat allocations and fare calculations in a train reservation system.

The **seat_alloted** attribute specifies the seat or berth assigned to a passenger, ensuring clarity in seating arrangements and helping passengers locate their seats easily. The **duration_minutes** attribute records the total travel time in minutes for each journey, which is essential for scheduling, operational planning, and providing passengers with accurate travel information. The **fare** attribute denotes the cost of the ticket for the journey, which is calculated based on factors such as distance, class of travel, and additional services.

These attributes together provide a clear and organized way to manage seat assignments, calculate travel durations, and ensure accurate fare computations, contributing to an efficient and user-friendly train reservation system.

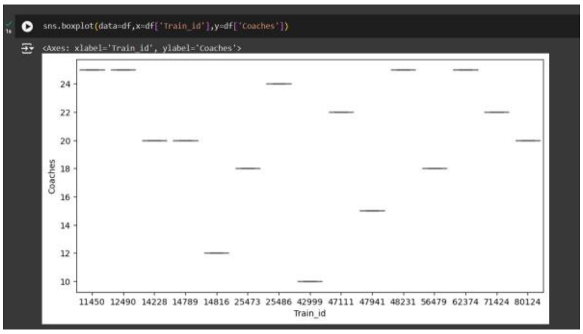


```
f,x='Train_id',y='Train_name',hue='Coaches')

<Axes: xlabel='Train_id', ylabel='Train_name'>
```

The graph features **train_names** on the y-axis and **train_id** on the y-axis, which suggests a unique structure intended to display the relationship between train names and their corresponding IDs.

In this graph, each point represents a train, with its position indicating both its name and its unique identifier. Since both axes contain similar data (train names and IDs), the graph likely highlights a comparison or mapping between these two attributes. This setup is useful in ensuring that each train ID is correctly associated with its respective train name, validating the accuracy and consistency of the dataset.

Such a graph can be particularly beneficial for database administrators and analysts in verifying the integrity of the train records. It ensures that every train name is uniquely matched to a train ID without duplication or errors. This visualization might also help in identifying any discrepancies or anomalies where a train name might be incorrectly assigned to multiple IDs or vice versa.



```
sns.boxplot(data=df,x=df['Train_id'],y=df['Coaches'])
<Axes: xlabel='Train_id', ylabel='Coaches'>
```

The graph features **coaches** on the y-axis and **train_id** on the x-axis, illustrating the relationship between different trains and the number of coaches each train has.

Each point on the graph represents a specific train, identified by its unique train_id on the x-axis, and shows the corresponding number of coaches on the y-axis. This setup allows for a clear comparison across various trains in terms of their capacity and coach configuration.

By examining the distribution of points, one can easily identify trends and patterns. For instance, certain train_ids may have a higher number of coaches, indicating they are designed for higher passenger capacity or longer routes. Conversely, trains with fewer coaches may be intended for shorter distances or less populated routes.

This graph is particularly useful for railway operations and logistics planning. It helps in assessing whether the allocation of coaches per train aligns with the demand and operational requirements. It can also highlight the need for reconfiguration or redistribution of coaches to optimize the use of resources.

Furthermore, the visualization can assist in identifying outliers or inconsistencies, such as a train_id with an unusually high or low number of coaches, which may warrant further investigation or correction in the dataset.

Overall, this graph provides valuable insights into the composition of train sets within the railway network, aiding in efficient management and strategic planning of train services.

REFERENCES

[1] Ruifan Tanga, Lorenzo De Donatob, Nikola Besinović, Francesco Flammini, Rob M.P. Goverdec, Zhiyuan Lina, Ronghui Liu, Tianli Tang, Valeria Vittorini, Ziyulong Wang (2022). A literature review of Artificial Intelligence applications in railway systems.

[2] Qi Wang, Siki Bu, Zhengyou He, 2020c. Achieving predictive and proactive maintenance for high-speed railway power equipment with LSTM-RNN. IEEE Trans. Ind. Inf. 16 (10), 6509–6517

[3] Rahul Barman, Chandra Jyoti Baishya, Bandonlang Kharmalki, Aphibakordor Syiemlieh, Kriti Bikash Pegu Tanuja Das [2016], Automated Train Scheduling System using Genetic Algorithm.

[4] Kecman P, Goverde, R.M.P., 2015. Predictive modelling of running and dwell times in railway traffic. Public Transp. 7 (3), 295–319

[5] Alawad, HAH, Kaewunruen, S., An, M., 2020. Learning from accidents: Machine learning for safety at railway stations. IEEE Access 8, 633–648.

[6] Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., Kautz, J., 2017. Reinforcement learning through asynchronous advantage actor-critic on a GPU.

[7] Wissam Sammouri, Etienne Come, Latifa Oukhellou, Patrice Aknin. Floating Train Data Systems For Preventive Maintanence : A Data Mining Approach

[8] Shripad Salsingikar, Narayan Rangaraj 2020, Reinforcement Learning for Train Movement Planning at Railway Stations Adaptive and Learning Agents Workshop.

[9] Paul Schonfeld, Taoran Sang, Hong Zhang [2019]. Mountain railway alignment optimization using stepwise & hybrid particle swarm optimization incorporating genetic operators. Appl. Soft Comput. 78, 41–57

[10] Jiateng Yin, Dewang Chen, Lingxi Li [2014] Intelligent Train Operation Algorithms for Subway by Expert System and Reinforcement Learning.

[11] Jürgen Wohlfeil [2011] Vision based rail track and switch recognition for self-localization of trains in a rail network.

[12] Dewei Li, Chen Zhang, Jinming Cao [2016]. Short-Term Passenger Flow Prediction of a Passageway in a Subway Station Using Time Space Correlations Between Multi Sites

[13] Ye Han, Zhigang Liu, Kai Liu, Yang Liu [2020]. Deep Learning-based Visual Ensemble Method for High-speed Railway Catenary Clevis Fracture Detection.

[14] Xavier Gibert, Vishal M. Patel, Rama Chellappa [2015]. Robust Fastener Detection for Autonomous Visual Railway Track Inspection

[15] María Pilar Tormos, A. Lova, F. Barber, Laura Paola Ingolotti [2008]. A Genetic Algorithm for Railway Scheduling Problems.

VI. CONCLUSION

The railway reservation data science project demonstrates the profound impact of leveraging data analytics to optimize and enhance the railway reservation system. By analyzing vast datasets encompassing ticket bookings, passenger demographics, train schedules, and peak travel times, significant insights were derived to improve operational efficiency and passenger satisfaction. Predictive models were developed to forecast demand, enabling better resource allocation and dynamic pricing strategies. Additionally, the project identified patterns in booking cancellations and no-shows, allowing for more effective overbooking policies. The implementation of these data-driven strategies not only aims to increase revenue but also to minimize congestion and improve the overall travel experience. Ultimately, this project underscores the potential of data science in transforming traditional transportation systems into smarter, more efficient networks.