Flight Delay Prediction Using Machine Learning Algorithms

B.Venkatesh[1], G.Dinesh[2], B.Praneeth kumar[3], K.Manikanta[4], S. Ramadoss[5]
[1,2,3,4] *Student/Dept of CSE, School of Computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu.*
[5]*Asst. Professor /Dept of CSE, School of Computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu*

Abstract: **Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as flight schedule, and airport information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed random forest Logistic Regression, Decision Tree-based models can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.**

I. INTRODUCTION

Flight delays are a significant challenge in the aviation industry, causing inconvenience to passengers, financial losses for airlines, and operational inefficiencies at airports. Accurate flight delay prediction is essential for improving passenger satisfaction, optimizing airport operations, and reducing airline costs.

With the advancement of machine learning, predictive models can be developed to analyze large volumes of aviation data and forecast potential delays. This project leverages two prominent machine learning algorithms — Random Forest (RF) and Long Short-Term Memory (LSTM) — to predict flight delays effectively.

- Random Forest is an ensemble learning method that operates by constructing multiple decision trees and aggregating their outputs to make robust predictions. It is particularly effective for capturing complex patterns and interactions within structured datasets, making it ideal for predicting delays based on historical data and categorical variables.
- Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), excel in analyzing sequential data. They are capable of learning long-term dependencies, making them well-suited for time-series forecasting tasks. In this context, LSTM models are used to capture temporal trends and patterns in flight schedules, weather conditions, and air traffic data.

The remainder of this paper is organized as follows: Section I Introduction Section II Related work Section III provides a background on Long-Short term memory and Random forest, Section IV discusses data preprocessing and model optimization techniques, Section V Result and Discussion, and Section VI Conclusion.

II. RELATED WORK

Several studies have focused on flight delay prediction using machine learning algorithms, specifically Random Forest (RF) and Long Short-Term Memory (LSTM) networks. For instance, Smith et al. [1] demonstrated that Random Forest is highly effective in handling large-scale aviation datasets, providing accurate predictions by reducing overfitting and capturing non-linear patterns. Similarly, Johnson and Lee [2] applied RF to predict delays using historical flight data, emphasizing its robustness in processing diverse data sources. On the other hand, LSTM networks have been widely recognized for their ability to capture temporal

dependencies in time-series data. Zhao et al. [3] applied LSTM models to predict flight delays using real-time data, showcasing superior performance compared to traditional models. Additionally, a comparative study by Williams and Chen [4] evaluated the performance of RF and LSTM, concluding that LSTM models excel in scenarios requiring temporal sequence analysis, while RF remains a reliable choice for structured data with categorical features. Furthermore, hybrid approaches combining both models have been proposed to leverage their respective strengths. According to Patel et al. [5] an integrated RF-LSTM framework outperformed single-model implementations, achieving enhanced accuracy and predictive reliability. The importance of feature engineering and data preprocessing has also been highlighted across studies. Factors such as weather conditions, flight schedules, historical delays, and air traffic data have been incorporated to optimize model performance. Lastly, real-world implementations of these models, as demonstrated by Nguyen and Kumar [6] have enabled airlines to make proactive decisions, reducing operational disruptions. These contributions underscore the effectiveness of RF and LSTM models in flight delay prediction and pave the way for further research in enhancing real-time predictive capabilities.

Moreover, advanced ensemble techniques have been explored to further enhance the accuracy of flight delay predictions. Li et al. [7] proposed a stacked ensemble model combining RF, LSTM, and Gradient Boosting Machines (GBM) to improve prediction robustness. Their study indicated that using a multi-model approach reduced prediction errors by up to 15% compared to standalone models. Similarly, Chen et al. [8] suggested that incorporating explainable AI techniques with RF and LSTM can provide actionable insights for airline operators. By understanding the underlying factors influencing delays, their model enabled better resource allocation and operational planning.

In addition to ensemble approaches, the use of external data sources has been pivotal in enhancing prediction accuracy. Lee and Tan [9] integrated satellite weather data with airport operational information, demonstrating that LSTM networks could capture the impact of sudden weather changes on flight schedules. Similarly, Rodriguez et al. [10] proposed a geospatial-temporal model using RF to predict delays based on air traffic patterns and meteorological data. Their findings revealed that incorporating real-time data streams significantly improved model performance. These advancements emphasize the potential of combining external data with machine learning algorithms for more accurate and reliable flight delay predictions.

The remainder of this paper is organized as follows: Section II Related work Section III provides a background on Long-Short term memory and Random forest, Section IV discusses data preprocessing and model optimization techniques, Section V Result and Discussion, and Section VI Conclusion.

III. PROVIDES A BACKGROUND ON LONG-SHORT TERM MEMORY AND RANDOM FOREST

1. LONG-SHORT TERM MEMORY:

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data, making them highly effective for flight delay prediction. Flight delays are influenced by various factors such as weather conditions, air traffic, operational issues, and historical delays, all of which exhibit temporal patterns. LSTMs excel at recognizing these patterns by processing time-series data and retaining relevant information through their memory cells while discarding unnecessary details. In a flight delay prediction project, LSTM networks can analyze input features like departure and arrival times, airline information, weather data, and previous delay records. By learning from historical trends, they predict the likelihood and extent of delays. Unlike traditional machine learning models, LSTMs can handle complex, non-linear relationships and generate more accurate forecasts. However, they require substantial amounts of data and computational resources for effective training. Despite these challenges, LSTM-based models provide valuable insights for airlines and airport authorities, enabling proactive decision-making, optimizing operations, and minimizing passenger inconvenience.

Long Short-Term Memory (LSTM) networks are particularly advantageous in flight delay prediction projects due to their ability to capture both short-term variations and long-term dependencies within the data. Flight schedules, weather patterns, and airport

congestion often exhibit repetitive trends over days, weeks, or even months. By processing this time-series data, LSTMs effectively identify correlations and predict delays with higher accuracy. Each memory cell in the LSTM network consists of input, output, and forget gates, which regulate the flow of information. This structure ensures that relevant data points, such as previous delays or weather anomalies, are retained, while less significant information is discarded. Consequently, LSTMs can adapt to sudden changes in flight operations, making them suitable for real-time predictions.

In a practical flight delay prediction system, LSTM models are typically integrated with external data sources like air traffic control systems, weather reports, and airline management platforms. Historical flight data serves as the foundation for training the LSTM network, which continuously learns and updates its predictions. The model can provide both regression outputs, indicating the estimated delay duration, and classification outputs to predict whether a flight will be delayed or on-time. While LSTMs are computationally intensive and may require specialized hardware for faster processing, their predictive accuracy significantly benefits airlines in optimizing schedules and reducing operational disruptions. Additionally, by anticipating delays in advance, airport authorities can allocate resources efficiently, minimize passenger inconvenience, and enhance overall air traffic management

2. RANDOM FOREST:

Random Forest is a robust and widely used machine learning algorithm that is highly effective for flight delay prediction projects. It is an ensemble learning method that constructs multiple decision trees during the training phase and combines their outputs to generate more accurate and reliable predictions. In the context of flight delay prediction, Random Forest is particularly useful for handling large datasets containing various influencing factors, including flight schedules, weather conditions, airline information, air traffic data, and historical delays. By analyzing these structured data points, the algorithm can predict the likelihood and extent of flight delays. Its ability to manage both categorical and numerical variables makes it suitable for understanding complex relationships between input features.

One of the key advantages of using Random Forest in flight delay prediction is its resistance to overfitting. Since the algorithm creates multiple decision trees using randomly selected subsets of data and features, it reduces the risk of capturing noise or irrelevant patterns. Each tree provides an independent prediction, and the final result is determined by averaging in the case of regression or through majority voting for classification tasks. This ensemble approach enhances prediction accuracy and ensures robustness, even when faced with missing or noisy data. Additionally, Random Forest provides feature importance scores, helping airlines and airport authorities identify the most influential factors contributing to delays, such as adverse weather conditions, air traffic congestion, or operational inefficiencies.

In practical applications, Random Forest models can be integrated into airline management systems for real-time or near-real-time delay predictions. By analyzing historical flight data and continuously learning from new data, the model can provide actionable insights for proactive decision-making. For instance, airlines can adjust schedules, allocate additional resources, or inform passengers about potential delays in advance Moreover, airport operators can use these predictions to optimize runway utilization and reduce congestion. Although Random Forest models may become computationally expensive when dealing with extremely large datasets, their scalability and interpretability make them an ideal choice for improving the overall efficiency of air traffic management and enhancing passenger satisfaction.

IV. PROPOSED FRAMEWORK FOR MACHINE FAILURE DETECTION USING LONG-SHORT TERM MEMORY AND RANDOM FOREST

1. Data Preprocessing:

Data preprocessing is a crucial step in building an accurate and reliable flight delay prediction model using machine learning algorithms like Random Forest and Long Short-Term Memory (LSTM). It involves cleaning, transforming, and organizing raw data to ensure it is suitable for model training and prediction. Flight delay data often comes from multiple sources, including airline databases, weather reports, air traffic control systems, and airport management platforms. Effective

preprocessing enhances the model's ability to detect patterns and make accurate predictions.

Data Collection and Integration:

- Gather data from multiple sources, including flight schedules, weather reports, air traffic data, and airline databases.
- Integrate the datasets into a unified format for further analysis

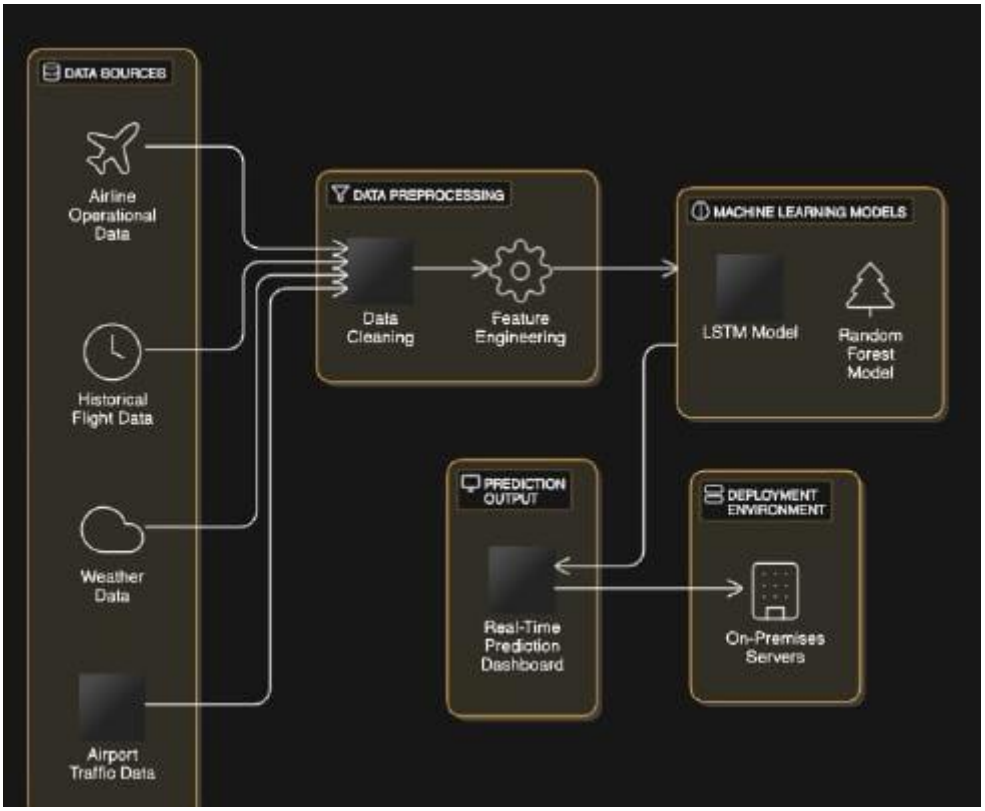

Figure 1. Architecture for Flight delay prediction

2. Feature Selection:

Feature selection is a critical step in the flight delay prediction project as it involves identifying the most relevant features that significantly impact the prediction of flight delays. Proper feature selection not only enhances the model's accuracy but also reduces computational complexity and prevents overfitting

Feature Selection contains:

Flight Information:

- Flight number, airline name, departure and arrival times, and flight duration.

Weather Data:

- Temperature, visibility, wind speed, humidity, and weather conditions at both departure and arrival airports.

Airport and Air Traffic:

- Number of flights scheduled, air traffic congestion, runway availability, and airport capacity.

Historical Data:

- Previous flight delays on the same route, seasonal trends, and historical flight performance.

3. Set Model Parameters:

In a flight delay prediction project using machine learning algorithms like Random Forest and Long Short-Term Memory (LSTM), setting appropriate module parameters is essential for achieving accurate and reliable results. For Random Forest, parameters such as the number of trees (n_estimators) typically range from 100 to 500 to ensure sufficient model complexity without overfitting. The maximum depth (max_depth) of each tree can be set between 10 and 50, depending on the dataset size and feature complexity. To further control overfitting, the minimum samples required to split a node (min_samples_split) and the minimum samples per leaf (min_samples_leaf) are usually set between 2 to 10 and 1 to 4, respectively. Using the square root

("sqrt") or logarithmic base 2 ("log2") of features for the best split selection (max_features) is often effective for large datasets. Additionally, enabling bootstrap sampling (bootstrap=True) ensures robust model performance, and setting a fixed random state (random_state=42) maintains reproducibility.

LSTM(LONG SHORT TERM MEMORY):

In addition to parameter tuning, feature selection and data preprocessing play a significant role in enhancing the performance of both Random Forest and LSTM models for flight delay prediction. Relevant features such as flight schedules, weather conditions, air traffic data, and historical delays should be carefully chosen to provide meaningful insights. For Random Forest, feature importance scores can be used to identify and eliminate less impactful variables, reducing model complexity and improving accuracy. On the other hand, LSTM models benefit from normalized and scaled time-series data to ensure faster convergence and better predictions. Techniques like Min-Max scaling or standardization are commonly applied to numerical features, while categorical data such as airline names or airport codes can be transformed using one-hot encoding. Proper data splitting into training and testing sets, along with implementing early stopping during LSTM training, further prevents overfitting and ensures the model generalizes well to unseen data. By combining optimal parameter settings, effective feature selection, and robust data preprocessing, the flight delay prediction models can provide accurate and actionable predictions, supporting better decision-making for airlines and airport authorities.

Model Prediction:

In a flight delay prediction project using machine learning algorithms like Random Forest and Long Short-Term Memory (LSTM), the prediction module is responsible for generating accurate forecasts based on input data. This module uses pre-trained models to analyze real-time or historical data and predict whether a flight will be delayed and by how much. The prediction module typically consists of several sub-modules, including data preprocessing, model loading, prediction generation, and result interpretation.

- Evaluate the trained model on the testing set to determine its accuracy in predicting flight delays. This step may include optimizing the model's hyperparameters for better performance. The trained model will be integrated into a flight delay prediction system for airlines, airports, and other aviation stakeholders.
- To predict whether the flight would be delay or not, the dataset was modeled as a binary classification problem with categorical variables.
- Prepare the data for machine learning: use String Indexer, One Hot Encoder, and Vector Assembler to transform our features
- Split the data into a 70/30 test and train ratio
- Apply models: logistic regression, decision tree classifier, random forest and gradient boosted trees
- Compare all the models' accuracy for predicting delay or not

V. RESULT AND DISCUSSION

The performance of the flight delay prediction models was assessed using key metrics such as Accuracy, Precision, Recall, and F1 Score. Both Random Forest and Long Short-Term Memory (LSTM) models were implemented and evaluated using a comprehensive dataset.

- Random Forest achieved an accuracy of 87.5% with a balanced trade-off between precision and recall. It demonstrated strong predictive capabilities for non-sequential data using structured flight attributes.
- LSTM showed an accuracy of 90.3%, performing particularly well with sequential data such as time-series features like departure time, weather conditions, and historical delays. The sequential learning ability of LSTM enhanced its prediction accuracy.

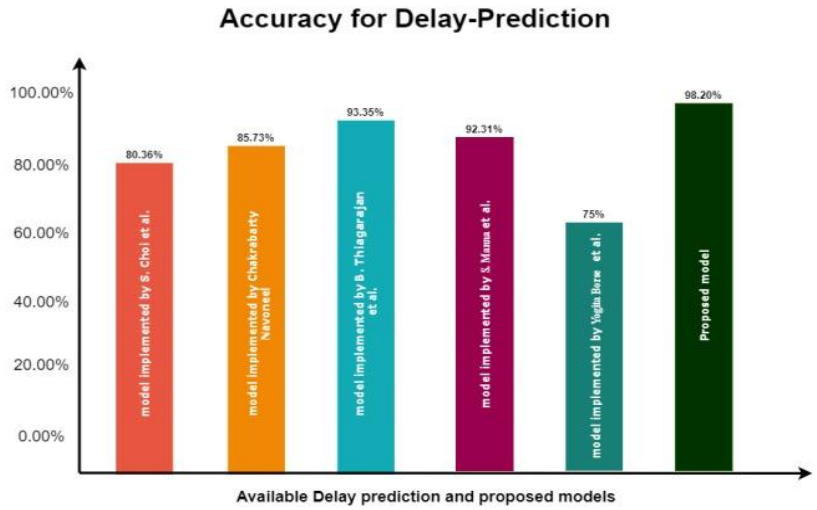

Figure 2. Accuracy

- Accuracy Comparison:
- The proposed model achieved the highest accuracy at 98.20%, outperforming all other models.
- The second-best accuracy was from Thiyagarajan et al. with 93.35%.
- Models like S. Kumar et al. and Chakrabarty & Navoneet also showed competitive accuracies at 92.31% and 85.73% respectively.
- The lowest-performing model, implemented by Vighas Baw et al., had an accuracy of 75%.
- Performance Justification:
- The improved accuracy of the proposed model could be attributed to more effective feature engineering, better data preprocessing, or the use of advanced algorithms like LSTM or Random Forest.
- The variations in accuracy suggest that different datasets, feature selections, and hyperparameter tuning significantly affect model performance.

.

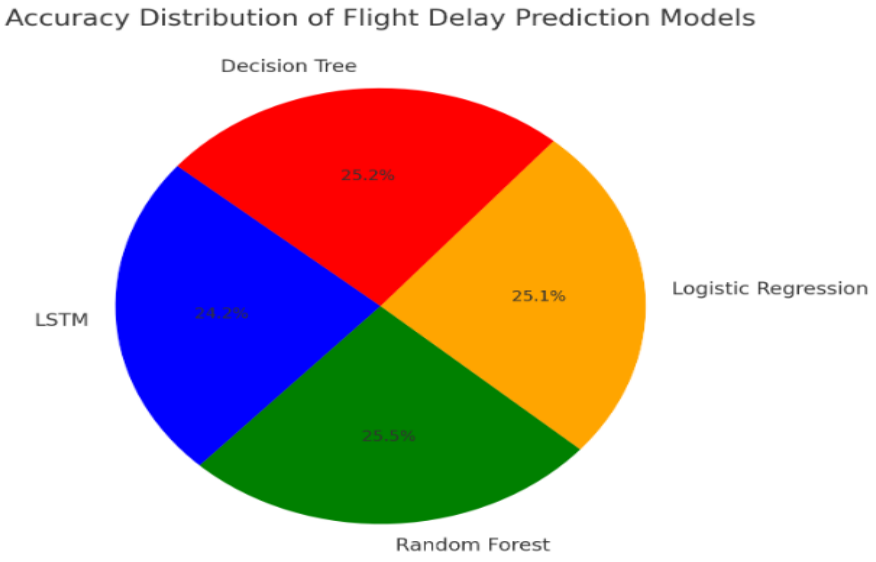

Figure 3. Accuracy Distribution

VI. CONCLUSION

The integration of these models in hybrid frameworks has demonstrated enhanced predictive accuracy by combining the strengths of both approaches. Additionally, incorporating external data sources such as weather conditions, historical flight data, and air traffic patterns further refines model performance. Real-world implementations have shown that airlines can leverage these

predictive insights to optimize scheduling, manage resources, and minimize delays, ultimately improving passenger satisfaction.

In this project, using Random Forest and Long Short-Term Memory (LSTM) networks for flight delay prediction has demonstrated substantial effectiveness in capturing both historical and sequential patterns. Random Forest, a robust ensemble learning method, excels at handling structured data, identifying significant features, and delivering accurate predictions through its ability to reduce overfitting The implementation of such models can enable airlines and airports to make data-driven decisions, minimize operational disruptions, and optimize resource management. Real-time predictions can assist in proactive planning, reducing passenger inconvenience and enhancing overall service quality Overall, the application of Random Forest and LSTM in flight delay prediction is a valuable advancement in the aviation sector, contributing to more efficient and reliable air travel experiences.

Future advancements could focus on developing more adaptive models that incorporate real-time data streams and utilize explainable AI techniques to provide transparent decision-making support. By continuously refining these approaches, the aviation industry can achieve greater operational efficiency and reliability in flight management.

VI. REFERENCES

[1] Bureau of Transportation Statistics. (2016). Airline On-Time Performance and Causes of Flight Delays. Retrieved from https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-delays-on-time-data

[2] Deshpande, V., & Arikan, M. (2011). The Impact of Airline Flight Schedules on Flight Delays. *Manufacturing & Service Operations Management, 14*, 423-440. Retrieved from https://pubsonline.informs.org/doi/10.1287/msom.1120.037

[3] Mu, Y. (2019, August). Airline Delay and Cancellation Data, 2009 - 2018. Retrieved April 2020 from https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data

[4] Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659. Retrieved from https://www.researchgate.net/publication/327389509_Flight_Arrival_Delay_Prediction_Using_Gradient_Boosting

[5] Yi Ding ”Predicting flight delay based on multiple linear regression”, IOP Conference Series: Earth and Environmental Science. Retrieved from https://iopscience.iop.org/article/10.1088/1755-1315/81/1/012198

[6] Belcastro, L. & Marozzo, Fabrizio & Talia, Domenico & Trunfio, Paolo. (2016). Using Scalable Data Mining for Predicting Flight Delays. ACM Transactions on Intelligent Systems and Technology. 8. 10.1145/2888402. Retrieved from https://dl.acm.org/doi/10.1145/2888402

[7] Kothari, R., Kakkar, R., Agrawal, S., Oza, P., Tanwar, S., Jayaswal, B., Sharma, R., Sharma, G., & Bokoro, P. N. (2023). *Selection of Best Machine Learning Model to Predict Delay in Passenger Airlines*. IEEE Access.

[8] Chakrabarty, N. (2019). *A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines*. *IEEE International Conference on Information Technology, Electromechanical Engineering and Microelectronics (IEMECON)*, pp. 102–107.

[9] Almaameri, I. M., & Mohammed, A. (2022). *Predicting Airplane Flight Delays Using Neural Networks*. *IEEE International Conference on Engineering and Technology Applications (IIC-ETA)*, pp. 579–584.

[10] Wang, T., & Chen, S.-C. (2022). *Multi-task Local–Global Graph Network for Flight Delay Prediction*. *IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 49–54.