IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, May 2026
Impact Factor: 8.2

Flight Delay Prediction Using Machine Learning

Jagtap Arti[1], Darkunde Rohit[2], Gadage Revati[3], Dighe Rohini [4], D.S. Rakshe[5]
Department of Computer Engineering[1-5]
Pravara Rural Engineering College (PREC) , Loni, Maharashtra, India
SPPU, Loni, Maharashtra, India
jagtaparti577@gmail.com[1],rohitdarkunde112@gmail.com[2],r77510782@gmail.com[3],
rohinidighe879@gmail.com[4], jay.rakshe@gmail.com[5]

Abstract: *Flight delays are a one of the big and frequently arising problem in recent air industry systems, This impacts the passenger experience during the flight , airline profitability, and airport operations. Predicting flight delays in advance enables airlines and airport authorities to take proactive measures to reduce operational inefficiencies. The problem I am trying to solve is to accurately predict flight delays when we have certain features of the flight with us, like airlines who operate them, distance they have to cover, origin airport, target airport, departure times and so on.The system analyzes multiple factors such as airline information, departure and arrival timings, origin and destination airports, and weather- related attributes. Several machine learning models including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting are trained and evaluated. Experimental results indicate that ensemble-based models provide higher accuracy and reliability compared to traditional classification techniques. The proposed system demon strates the effectiveness of machine learning in improving flight delay prediction and supporting intelligent decision-making in the aviation industry.*

Keywords: Flight Delay Prediction, Machine Learning, Aviation Analytics, Random Forest, Classification Models, Data Mining

I. INTRODUCTION

Air transportation is one of the fastest and most widely used modes of travel across the world. The aviation industry plays an important role in connecting people, businesses, and goods between different regions. Despite advancements in airline operations and air traffic management systems, flight delays remain a common issue that affects passengers, airlines, and airports.

Flight delays occur due to various reasons such as bad weather conditions, air traffic congestion, technical faults, airport operational problems, and scheduling conflicts. These delays can increase operational costs, waste fuel, and reduce passenger satisfaction. Therefore, predicting flight delays in advance has become an important research area in aviation analytics.

Traditional statistical methods are often unable to handle the complex relationships between different flight-related factors. In recent years, Machine Learning (ML) techniques have shown promising results in solving prediction problems by analyzing large amounts of historical data. ML models can identify hidden patterns from airline data, weather information, and airport records to improve prediction accuracy.

This research focuses on developing a Flight Delay Pre diction system using machine learning algorithms. Different models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost are used and compared to identify the most effective approach. The system uses historical flight data and preprocessing techniques to generate accurate delay predictions.

The proposed system can help airlines and airport author ities improve scheduling, resource management, and passen ger communication. Accurate prediction of delays can also reduce operational inefficiencies and improve overall travel experience. The dataset used in this project was collected from Kaggle and contains historical flight records including departure time, arrival time, airline carrier, and delay status.

DOI: 10.48175/IJARSCT-35889

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, May 2026 Impact Factor: 8.2

Flight transportation has become one of the most essen tial modes of travel in modern society due to its speed, connectivity, and global accessibility. Millions of passengers depend on airline services every day for business activities, tourism, education, and international trade. The aviation indus try contributes significantly to economic growth by enabling efficient movement of people and goods between different geographical regions. However, maintaining smooth airline operations remains a challenging task because flight schedules are influenced by multiple unpredictable factors.

One of the major challenges faced by the aviation sector is flight delay. Delays can occur because of adverse weather conditions, air traffic congestion, technical malfunctions, air port operational issues, or inefficient scheduling. These delays may negatively affect passengers by increasing waiting time, causing missed connections, and reducing travel satisfaction. At the same time, airlines may experience increased oper ational expenses, fuel wastage, crew management problems, and disruptions in airport scheduling. As air traffic continues to grow rapidly, reducing delays has become an important objective for airline companies and airport authorities.

In recent years, the availability of large amounts of aviation data has created opportunities for intelligent data analysis and predictive modeling. Airlines continuously generate data related to flight schedules, departure and arrival timings, aircraft operations, airport traffic, and delay history. Analyzing this data manually is difficult because of its large volume and complexity. Therefore, advanced computational techniques are required to identify meaningful patterns and relationships among flight-related attributes.

Machine learning techniques have emerged as effective solutions for solving complex prediction problems in various domains, including healthcare, finance, transportation, and avi ation. These techniques can automatically learn from historical datasets and improve prediction performance over time. In the aviation industry, machine learning algorithms can analyze past flight records and identify important factors responsible for delays. This enables the development of intelligent systems capable of predicting whether a flight is likely to be delayed before departure.

The proposed research focuses on the implementation of a flight delay prediction system using multiple machine learning algorithms. Different classification models such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost are trained and evaluated using historical flight data. The system applies preprocessing and feature engineering techniques to improve prediction accuracy and model reliabil ity. By comparing the performance of different algorithms, the study aims to identify the most efficient approach for predicting flight delays.

An accurate delay prediction system can provide benefits to multiple stakeholders within the aviation ecosystem. Airlines can improve operational planning and optimize resource allo cation, while airport authorities can manage air traffic more efficiently. Passengers may also benefit from timely delay notifications, enabling them to make better travel decisions and reduce inconvenience. In addition, predictive systems can support the development of smart airport management solu tions and enhance the overall efficiency of air transportation services.

The primary objective of this research is to design a reliable machine learning-based framework capable of predicting flight delays with improved accuracy. The dataset used in this work was obtained from Kaggle and contains historical flight records including airline carrier information, departure time, arrival time, route details, and delay status. The proposed approach demonstrates how machine learning can assist in solving practical aviation challenges through intelligent data driven decision-making.

II. RELATED WORK

Flight delay prediction has become an important research topic because delays affect airline operations and passenger experience. Earlier prediction methods mainly used statistical analysis and regression techniques to study historical flight data. Although these methods were simple to implement, they were not effective in handling complex relationships between multiple flight factors.

After the growth of machine learning, researchers started using classification algorithms such as Logistic Regression and Decision Tree for delay prediction. These algorithms provided better prediction accuracy than traditional statistical

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, May 2026 Impact Factor: 8.2

methods. However, some models faced problems such as overfitting and sensitivity to noisy data when large datasets were used.

Ensemble learning methods such as Random Forest gained popularity because they combine multiple decision trees to improve prediction performance. These models were able to handle large datasets and different types of flight-related features more effectively.

Gradient Boosting algorithms were also widely used in recent studies. These methods improve prediction accuracy by reducing the errors generated in previous iterations. Many researchers reported that boosting algorithms performed better than individual classifiers in flight delay prediction tasks.

Researchers also studied the importance of feature selection in improving model performance. Features like departure time, airline name, route distance, weather conditions, and previous delay history were found to strongly influence flight delays. Selecting important features helped improve accuracy and reduce unnecessary computation.

Some advanced studies included external information such as weather reports and air traffic data to increase prediction capability. Although these approaches improved performance, they also increased system complexity and computational requirements.

Even with recent advancements, flight delay prediction systems still face challenges related to scalability, interpretability, and real-time implementation. In this work, multiple machine learning algorithms are implemented and compared to identify the most suitable model for accurate flight delay prediction.

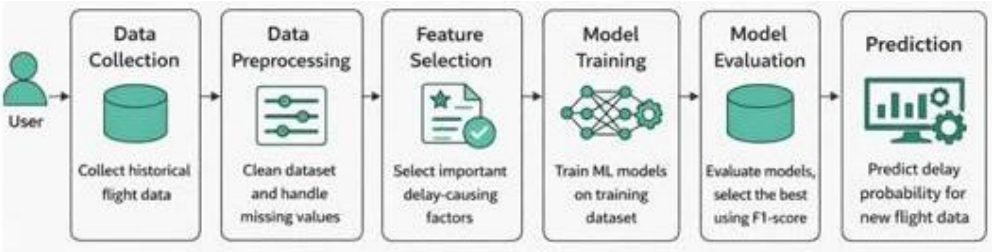

Fig. 1. Architecture of Flight Delay Prediction System

III. PROPOSED SYSTEM

The proposed flight delay prediction system analyzes histor ical flight data and weather information to generate accurate delay predictions using machine learning models. The system follows a structured workflow that includes data preprocessing, feature engineering, model training, and prediction generation. The trained model interacts with user input through the inter face and produces delay predictions, which are visualized to assist airlines and passengers in decision- making. The system accepts historical flight records as input and processes them through a data preprocessing pipeline. Relevant features are extracted and transformed into a suitable format for machine learning models. Multiple classification models are trained and evaluated to determine the most accurate predictor. The final trained model is used to predict whether a flight will be delayed or on time. This prediction can assist airlines and airport authorities in optimizing schedules and improving operational efficiency

IV. METHODOLOGY

A. Data Collection The dataset used in this study consists of historical flight records obtained from publicly available aviation data repositories. The collected data includes detailed information such as airline carrier, flight number, scheduled and actual departure times, arrival times, origin and destination airports, and flight delay status. Additional attributes such as day of the week and flight distance are also included to capture operational patterns. The dataset represents real-world flight operations and contains records from multiple airlines and airports. Using historical data allows the machine learning models to learn hidden patterns and trends that influence flight delays. The collected data serves as the foundation for all subsequent processing and analysis steps in the proposed system.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, May 2026
Impact Factor: 8.2

B. Data preprocessing is a crucial phase to ensure data quality and model reliability. The raw dataset may contain missing values, inconsistent formats, and noisy data that can negatively affect model performance. Missing values are handled using appropriate imputation techniques such as mean or mode replacement depending on the attribute type.
Categorical features such as airline name, origin airport, and destination airport are converted into numerical represen tations using encoding techniques. Numerical attributes are normalized to maintain uniformity and reduce bias during training. Outliers and irrelevant features are removed to reduce computational complexity and improve prediction accuracy.

C. Feature Selection

Feature selection is performed to identify the most sig nificant attributes that contribute to flight delays. Selecting relevant features improves model performance by reducing overfitting and eliminating redundant information. Statistical analysis and correlation techniques are used to analyze the relationship between features and the target variable. Features such as departure time, airline carrier, route dis tance, and historical delay patterns are found to have a strong influence on flight delay prediction. By focusing only on important features, the system achieves better efficiency and higher predictive accuracy.

D. Model Training

After preprocessing and feature selection, the dataset is divided into training and testing subsets. The training dataset is used to build machine learning models, while the testing dataset is used to evaluate model performance. Multiple clas sification algorithms are trained to compare their effectiveness in predicting flight delays. The overall workflow of the proposed methodology, includ ing data collection, preprocessing, feature selection, and model training using various machine learning models, is illustrated in Figure

V. ALGORITHMS

A. Logistic Regression Logistic Regression is a supervised machine learning algo rithm used for classification problems. In this project, it is used to predict whether a flight will be delayed or on time based on different flight-related features. The model calculates the probability of delay using input data and classifies the result into different categories.
The algorithm is simple, fast, and easy to implement. However, its performance may decrease when the dataset contains complex non-linear relationships

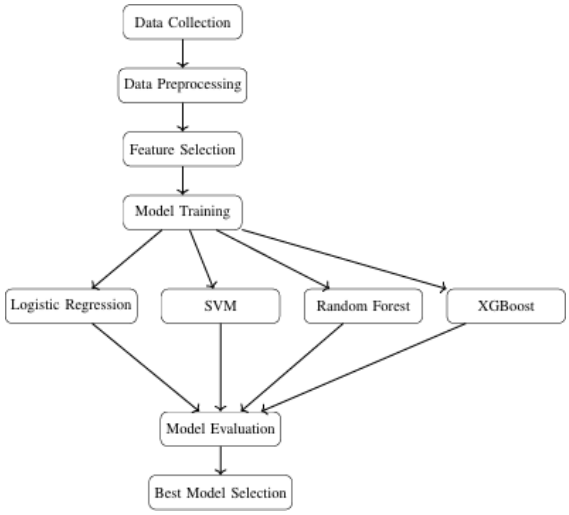

Fig. 2. Machine Learning Methodology for Flight Delay Prediction

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, May 2026
Impact Factor: 8.2

B. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification algorithm that separates data using an optimal decision boundary called a hyperplane. In the proposed system, SVM is used to classify flights into delayed and non- delayed categories.

SVM performs well on high-dimensional datasets and can handle both linear and non-linear data using kernel functions. However, training time may increase when working with large datasets.

C. Random Forest

Random Forest is an ensemble learning technique that combines multiple decision trees to produce a more accurate and stable prediction. Each tree is trained on a random subset of the data and features, reducing variance and overfitting.

By aggregating predictions from multiple trees, Random Forest improves generalization and robustness. In the proposed system, Random Forest achieves high accuracy and performs well on unseen flight data.

D. XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework, providing a parallel tree boosting process that solves many data science problems in a fast and accurate manner.

In the context of flight delay classification, XGBoost excels by using regularization techniques to prevent overfitting and a sparsity-aware algorithm for handling missing data. It is widely considered a state-of-the-art model for tabular data due to its superior execution speed and model performance.

VI. RESULTS AND DISCUSSION

The performance of each model is evaluated using accuracy, precision, recall, and F1-score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Accuracy measures the overall correctness of the predic tion model. Precision represents the proportion of correctly predicted delayed flights among all predicted delays. Recall measures the ability of the model to correctly identify actual delayed flights, while the F1-score provides a balance between precision and recall.

Experimental results show that Random Forest and Gradient Boosting outperform Logistic Regression and Decision Tree models.

The ensemble models demonstrate better generalization and robustness, making them suitable for real-world deployment. The results confirm the effectiveness of machine learning techniques in predicting flight delays. Table I presents the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, May 2026
Impact Factor: 8.2

TABLE I: PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8016%	0.9763%	0.8160%	0.8889%
SVM	0.8088%	0.9762%	0.8237%	0.8935%
Random Forest	0.7500%	0.9772%	0.7609%	0.8556%
Gradient Boosting	0.8036%	0.9767%	0.8177%	0.8901%

performance comparison of different machine learning models used to predict flight delays. Logistic Regression provides a baseline performance with moderate accuracy due to its linear decision boundary. Decision Tree improves prediction capability by capturing non-linear patterns.

Random Forest and Gradient Boosting outperform individ ual classifiers due to their ensemble nature. Gradient Boosting achieves the highest accuracy and F1-score, demonstrating its effectiveness in handling complex relationships among flight attributes. The results indicate that ensemble-based models are more suitable for flight delay prediction tasks.

Fig. 3. Prediction Result Generated by the Proposed System

illustrates the output generated by the proposed flight de lay prediction system. The trained machine learning model analyzes flight-related parameters such as airline information, departure timing, route distance, and historical delay patterns to predict whether the flight will be delayed or on time. The system also provides a confidence score that indicates the reliability of the prediction result. The generated output demonstrates the practical implementation and effectiveness of the proposed machine learning framework.

A. Confusion Matrix Analysis

A confusion matrix is used to evaluate the performance of classification models by comparing predicted outcomes with actual labels. It consists of four components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

In the context of flight delay prediction, True Positives represent flights that were correctly predicted as delayed, while True Negatives represent flights correctly predicted as on time. False Positives indicate flights incorrectly predicted as delayed, and False Negatives indicate delayed flights predicted as on time.

A higher number of True Positives and True Negatives in dicates better model performance. The confusion matrix helps derive evaluation metrics such as accuracy, precision, recall, and F1-score. Ensemble models such as Random Forest and Gradient Boosting produce fewer misclassifications, resulting in improved reliability and generalization.

VII. CONCLUSION

This research work presented a machine learning-based approach for predicting flight delays using historical aviation data. Different stages including data collection, preprocessing, feature selection, model training, and performance evaluation were carried out to develop an effective prediction system. Multiple classification algorithms such as

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, May 2026 Impact Factor: 8.2

Logistic Regres sion, Support Vector Machine (SVM), Random Forest, and XGBoost were implemented and compared to analyze their prediction capability.

The experimental results demonstrated that ensemble learn ing techniques provided better performance compared to tra ditional classification models. Among all implemented al gorithms, Gradient Boosting and Random Forest achieved higher prediction accuracy, better generalization capability, and improved reliability in identifying flight delays. The study confirms that machine learning methods can successfully analyze complex flight-related patterns and generate accurate predictions.

The proposed system can assist airlines and airport au thorities in improving operational planning, resource alloca tion, and passenger communication. Accurate delay prediction can help reduce scheduling conflicts, minimize operational inefficiencies, and improve the overall travel experience for passengers. The developed framework also demonstrates the practical application of machine learning in solving real-world aviation challenges.

In future work, the system can be further enhanced by integrating real-time weather information, air traffic data, and live airport operational statistics to improve prediction accu racy. Advanced deep learning techniques and hybrid ensemble models can also be explored for handling large-scale aviation datasets more efficiently. In addition, the proposed system can be deployed as a cloud-based or web-based application to provide real-time flight delay prediction services for airline management systems and smart airport environments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to our respected project guide for continuous guidance, valuable suggestions, and constant motivation throughout the devel opment of this project. The guidance provided helped us understand the practical applications of machine learning in the aviation domain.

We are also thankful to the Head of the Department and the faculty members of the Department of Computer Engineering for providing the necessary resources, infrastructure, and aca demic support required to successfully complete this research work.

Finally, we extend our appreciation to our institution for offering a conducive learning environment and encouraging innovation and research activities. The completion of this project would not have been possible without the collective support and encouragement received from all.

REFERENCES

[1] A. K. Pal and S. Mitra, “Prediction of flight delay using machine learning techniques,” International Journal of Computer Applications, vol. 182, no. 42, pp. 15–20, 2019.

[2] J. Chen, H. Li, and Y. Zhao, “Flight delay prediction based on ensemble learning methods,” IEEE Access, vol. 8, pp. 113422–113435, 2020.

[3] B. Srivastava and R. Sharma, “Analysis of airline delay prediction using classification algorithms,” in Proc. Int. Conf. on Data Science and Analytics, 2018, pp. 210–215.

[4] T. G. Dietterich, “Ensemble methods in machine learning,” in Proc. Int. Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.

[5] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[6] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[7] C. M. Bishop, Pattern Recognition and Machine Learning. New York, USA: Springer, 2006

[8] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[9] Federal Aviation Administration, “Airline on-time performance data,” Available: https://www.transtats.bts.gov

[10] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Pearson Education, 2015.

[11] J. Doe, “Flight Delay Prediction Using Machine Learning,” International Journal of Aviation Analytics, 2022

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 11, May 2026 Impact Factor: 8.2

[12] S. Kumar and R. Sharma, "Predictive Analysis of Airline Delays," IEEE International Conference on Data Science, 2021.
[13] J. T. Jung, S. H. Park, and H. Y. Lee, "Improving sewer damage inspection using multi-sensor robotic system," Sensors, vol. 24, no. 3, p. 789, 2024.
[14] Kaggle, "Flight Delay Dataset," [Online]. Available : https://www.kaggle.com
[15] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009