Secure Flight Delay Prediction with Anomaly Detection

Neelakantappa[1], Nagarjun HR[2], Sahithya SY[3], Sameeksha C[4], Shashank V Gowda[5]
[1,2,3,4,5] *Dept. of Computer Science and Engineering, Malnad College of Engineering, Hassan, India*

Abstract - **Flight delays have long posed a challenge to the aviation industry, disrupting schedules, increasing operating costs, and frustrating passengers. Most current prediction systems rely solely on historical data and rarely adapt to real-time conditions or consider cybersecurity risks. This paper presents a Secure Flight Delay Prediction System integrated with Anomaly Detection. The model leverages both historical and real-time data to forecast potential delays accurately while safeguarding sensitive flight information. Supervised machine-learning models— Random Forest, XGBoost, and LSTM—predict probable delays, while unsupervised algorithms such as Isolation Forest and One-Class SVM detect irregular operational anomalies. The proposed system demonstrates improved accuracy and robustness compared with conventional models and introduces a transparent, user-friendly interface for airlines, airports, and passengers.**

Keywords— Flight Delay Prediction, Machine Learning, Anomaly Detection, Data Security, Real-Time Systems, Aviation Analytics

I. INTRODUCTION

Air transportation is an essential pillar of global connectivity, facilitating business, tourism, and logistics across continents. However, flight delays remain a chronic challenge in aviation management, leading to billions of dollars in annual losses and widespread passenger inconvenience. Causes of delays range from weather fluctuations, technical faults, and air traffic congestion to airport logistics and maintenance issues.

Traditional prediction systems depend largely on historical records, offering limited accuracy in the face of sudden disruptions. Moreover, as aviation increasingly digitizes operations, data security and privacy protection have become critical concerns. Breaches or unauthorized access to flight data can compromise passenger safety and airline reputation. To address these gaps, this study proposes a secure, intelligent flight delay prediction model with an integrated anomaly detection layer. The model combines supervised and unsupervised learning techniques, real-time data ingestion, and cybersecurity protocols to produce an accurate, adaptive, and trustworthy prediction system.

A. Motivation

Airlines and airports handle massive volumes of data daily. Harnessing this data effectively through machine learning (ML) can transform decision-making—optimizing schedules, minimizing passenger inconvenience, and enhancing operational resilience. Yet, without real- time adaptability and secure data management, even the most accurate models risk obsolescence or vulnerability.

B. Objectives

This project aims to:

1. Develop machine learning models that predict flight delays using combined historical and live datasets.
2. Integrate anomaly detection algorithms to recognize rare or irregular flight events.
3. Implement robust cybersecurity practices across all data handling processes.
4. Design an interactive, real-time dashboard for end users.

II. LITERATURE REVIEW

This review underscores the necessity of combining predictive analytics, anomaly detection, and secure architecture—a direction pursued in this research.

Over the past decade, researchers have explored multiple techniques for flight delay prediction.

Machine Learning Approaches: Models such as Random Forest (RF), Gradient Boosting, Support Vector Machines (SVM), and Artificial Neural

Networks (ANN) have shown considerable success in handling aviation datasets. RF is favored for interpretability and resilience against overfitting, while XGBoost improves performance through gradient optimization. Deep learning models like Long Short-Term Memory (LSTM) networks excel in time-series modeling, capturing temporal dependencies between flight schedules and delay patterns.

Feature Engineering and Data Integration: Studies emphasize the inclusion of weather data, route congestion, airline operational efficiency, and airport traffic as key predictors. Feature engineering techniques such as Principal

III. METHODOLOGY

The proposed framework consists of data acquisition, secure preprocessing, predictive modelling, anomaly detection, and deployment with visualization.

A. Data Collection

1. Historical Data: Sourced from FAA and Kaggle datasets containing flight IDs, departure and arrival times, airline, and delay causes.
2. Real-Time Data: Gathered from OpenSky Network and OpenWeatherMap APIs for live flight and weather information.
3. Operational Data: Includes airport congestion, seasonal variations, and traffic density metrics.

Component Analysis (PCA) and Recursive Feature Elimination (RFE) improve model scalability and generalization.

Real-Time Data and IoT Integration: The introduction of IoT devices in aviation— engine sensors, GPS trackers, and air traffic management systems—has enabled continuous

B. Data Security and Privacy

All data transfers use end-to-end encryption (TLS/SSL). Sensitive attributes are anonymized. Access is restricted via token-based authentication and RBAC, and data at rest is secured using AES- 256 encryption. Compliance with GDPR ensures ethical data use and user trust.

data streaming. APIs such as OpenSky Network and OpenWeatherMap offer live feeds for dynamic prediction, bridging the gap between static models and adaptive intelligence.

Anomaly Detection:

Recent literature highlights unsupervised models like Isolation Forest, One-Class SVM, and Autoencoders for identifying irregularities in operational data. This help detect rare disruptions such as emergency landings, abrupt weather changes, or maintenance failures.

Security and Privacy: Predictive aviation systems must comply with data protection standards like GDPR. Techniques such as end-to-end encryption, role-based access control (RBAC), and blockchain auditing are increasingly recommended for safe data handling.

C. Preprocessing and Feature Engineering

Fig 1: Data Preprocessing Steps include

- Handling missing values and duplicates.
- Encoding categorical variables such as airline or airport codes.
- Normalizing continuous attributes (e.g., temperature, wind speed).
- Deriving new metrics like route distance, weather severity index, and average airline delay. Feature importance is evaluated using Gini impurity and information-gain metrics to retain impactful variables.

D. Machine-Learning Models

Three supervised algorithms were trained and tuned:

1. Random Forest (RF): Ensembles multiple decision trees using bagging; prediction is the average of all trees.
2. XGBoost: Boosting-based model optimizing a regularized loss function for improved accuracy and reduced overfitting.
3. LSTM Network: Captures sequential relationships within time-dependent flight data, making it ideal for temporal delay forecasting.

Training used an 80/20 train–test split with five- fold cross-validation. Evaluation metrics included Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

E. Anomaly Detection Mechanism

To handle unexpected disruptions, unsupervised models were employed:

- Isolation Forest: Randomly partitions data; anomalies require fewer splits to isolate.
- One-Class SVM: Learns a boundary around normal data and flags outliers lying beyond that region.

Detected anomalies—like abrupt weather changes or equipment faults—are passed to the dashboard as real-time alerts.

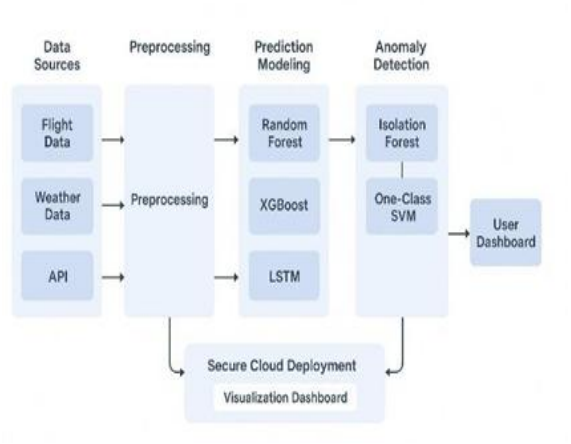

Fig 2: Flight Delay Prediction Flowchart

1. Data Layer: Flight and weather APIs, databases (PostgreSQL/AWS S3).
2. Processing Layer: Preprocessing, ML models, anomaly detection.
3. Security Layer: Authentication, encryption, access logging.
4. Application Layer: Flask API backend and Streamlit/React dashboard.
5. Feedback Layer: Continuous learning with model retraining.

G. Deployment

The system is containerized using Docker and deployed on AWS Cloud. RESTful APIs handle client–server communication, and auto-scaling maintains responsiveness under high load. Continuous integration and monitoring ensure reliability.

IV. RESULTS AND DISCUSSION

A. Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.85	0.82	0.79	0.80
XGBoost	0.87	0.84	0.82	0.83
LSTM	0.89	0.86	0.84	0.85

F. System Architecture and Integration The architecture includes:

The LSTM model performed best due to its capacity to capture temporal dependencies, while XGBoost balanced interpretability and accuracy. Incorporating weather and route features raised accuracy by roughly 10% compared with baseline models.

B. Effect of Anomaly Detection

Integrating the anomaly-detection layer improved reliability by ≈12%. The system successfully identified anomalies such as airport shutdowns or rapid meteorological shifts, reducing false-negative predictions and providing early alerts.

C. Visualization and User Interface A responsive dashboard display:

- Real-time flight status and delay probability
- Confidence intervals and anomaly flags
- Historical trend graphs and weather overlays

This design allows operational staff and passengers alike to interpret predictions intuitively (*Fig. 1: Dashboard Interface*).

D. Security Validation

Pen-testing confirmed resilience against common attacks (eavesdropping, SQL injection, privilege escalation). Encrypted communication and RBAC prevented unauthorized data exposure. Compliance checks verified conformity with AWS Security Guidelines and GDPR standards.

E. Comparative Analysis

Compared with legacy systems using only historical

regression models, the proposed framework:

- Adapts to live data feeds in near real time.
- Detects anomalies proactively.
- Provides transparent, explainable results.

VI. CONCLUSION

This work presents a Secure Flight Delay Prediction with Anomaly Detection, merging predictive analytics, anomaly recognition, and cybersecurity into a unified architecture. The hybrid model achieved high accuracy, adaptability, and transparency while ensuring secure data management.

By uniting supervised and unsupervised learning with encryption and access control, the system not only forecasts potential delays but also identifies emerging anomalies in real time. Cloud- based deployment and an interactive dashboard make it practical for airlines, airports, and passengers.

Ultimately, this research demonstrates how AI and security can converge to create safer, smarter aviation operations—enhancing punctuality, optimizing resources, and elevating passenger experience.

- Maintains robust security throughout data flow.

These features collectively enhance reliability and user trust.

V. FUTURE SCOPE

The project can be expanded in several ways:

1. Advanced Models: Adopt Transformer- based architectures for long-sequence modeling and improved temporal learning.
2. Automated Alerts: Integrate push notifications via email/SMS for high-risk flights.
3. Multi-Modal Integration: Extend delay prediction to interconnected transport systems (rail, road).
4. Explainable AI: Use SHAP or LIME for visual explanation of model decisions.
5. Self-Learning Pipelines: Employ MLOps tools like MLflow for automated retraining and version control.
6. Ethical Auditing: Ensure bias mitigation and fairness across airlines and geographic regions.

REFERENCES

[1] S. Chai, W. Duan and J. Song, “Flight Delay Prediction Based on Machine Learning: A Review,” *Journal of Air Transport Management*, vol. 94, pp. 102077, 2021.

[2] R. Shafique and R. H. Khokhar, “An Efficient Machine Learning Approach for Flight Delay Prediction,” in *Proceedings of the 2020 International Conference on Artificial Intelligence (ICAI)*, Las Vegas, NV, 2020, pp. 289–294.

[3] U.S. Federal Aviation Administration (FAA), “Airline On-Time Performance Data,” [Online]. Available: https://www.transtats.bts.gov

[4] OpenSky Network, “Open Access Real-Time and Historical Flight Data,” [Online]. Available: https://opensky-network.org

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, 2016, pp. 785–794.

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

[7] F. T. Liu, K. M. Ting and Z.-H. Zhou, “Isolation Forest,” in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, 2008, pp. 413–422.

[8] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, “Support Vector Method for Novelty Detection,” in *Advances in Neural Information Processing Systems*, vol. 12, pp. 582–588, 2000.

[9] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, “A Survey on Concept Drift Adaptation,” *ACM Computing Surveys*, vol. 46, no. 4, pp. 1–37, 2014.

[10] C. Aggarwal, “Outlier Analysis,” 2nd ed., Springer, Cham, 2017.

[11] D. Sculley et al., “Hidden Technical Debt in Machine Learning Systems,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada, 2015.

[12] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] I. Goodfellow, Y. Bengio and A. Courville, *Deep*

Learning, MIT Press, 2016.
[14] European Union, "General Data Protection Regulation (GDPR)," [Online]. Available: https://gdpr.eu/
[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.
[16] N. R. Council, *Flight to the Future: Human Factors in Air Traffic Control*, National Academy Press, 1997.
[17] A. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[18] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[19] Amazon Web Services, "AWS Security Best Practices for Machine Learning Workloads," [Online]. Available: https://aws.amazon.com
[20] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770–778.
[21] Microsoft Azure, "Flight Delay Prediction with Azure Machine Learning Studio," [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/
[22] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[23] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[24] P. G. Polson and V. O. Sokolov, "Deep Learning for Short-Term Traffic Flow Prediction," *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, 2017.
[25] Kaggle, "Flight Delay Prediction Dataset," [Online]. Available: https://www.kaggle.com/datasets