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 5, Issue 9, March 2025

Impact Factor: 7.67

Flight Delay Prediction using ML

Miss. S. V. Gunjal[1], Prof. V. D. Vaidya[2], Vyavahare Snehal Uday[3], Tambe Sujata Somnath[4], Shaikh Saniya Ajij[5], Aher Pooja Rameshwar[6]

Department of Cloud Computing and Big Data[1,2,3,4,5,6]

Padmashri Dr. Vitthalrao Vikhe Patil Institute of Technology and Engineering (Polytechnic), Pravaranagar

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 weather condition, 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-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.*

Keywords: Flight, Delay, Prediction, Machine Learning, Aviation

I. INTRODUCTION

1.1 Overview

Flight delays are a persistent challenge in the aviation industry, affecting airlines, passengers, and airport operations worldwide. These delays not only lead to financial losses for airline companies but also cause inconvenience to passengers, impacting their travel schedules and overall experience. Efficient prediction of flight delays can help mitigate these issues by enabling airlines to optimize scheduling, improve resource allocation, and enhance customer satisfaction. Traditional approaches to delay prediction rely on historical trends and statistical models, but they often fail to capture the complex interdependencies of various factors influencing flight delays. With advancements in data-driven methodologies, machine learning (ML) has emerged as a powerful tool for enhancing the accuracy of flight delay predictions.

Machine learning techniques allow for the analysis of large and diverse datasets, incorporating multiple variables such as weather conditions, air traffic congestion, flight schedules, and airport-specific characteristics. Unlike conventional statistical models, ML algorithms can identify hidden patterns and complex relationships within data, thereby improving prediction accuracy. Various ML models, including decision trees, support vector machines (SVM), and deep learning networks, have been explored for flight delay prediction. Among these, deep learning models like Long Short-Term Memory (LSTM) networks have shown promise in capturing sequential dependencies in aviation data. However, deep learning models are often susceptible to overfitting, particularly when dealing with limited datasets.

Recent studies have primarily focused on predicting flight delays for specific routes or airports, limiting the generalizability of their findings. This research aims to expand the scope by considering a broader range of influencing factors and evaluating the performance of multiple ML models across diverse flight scenarios. The dataset used in this study is constructed using automatic dependent surveillance-broadcast (ADS-B) messages, which provide real-time aircraft tracking data. These messages are integrated with additional information such as weather conditions, scheduled departure and arrival times, and airport-specific parameters to create a comprehensive dataset for model training and evaluation.

DOI: 10.48175/568

The study formulates flight delay prediction as both a classification and regression problem, enabling a thorough assessment of different machine learning approaches. The classification tasks involve binary classification (on-time vs. delayed) and multi-class classification (categorizing delay durations), while the regression task predicts the exact delay time in minutes. To evaluate model performance, standard evaluation metrics such as accuracy, precision, recall, and root mean square error (RMSE) are employed. The experimental results indicate that while LSTM is effective in processing sequential aviation data, it struggles with overfitting. In contrast, the random forest-based model demonstrates superior performance, achieving a high classification accuracy of 90.2% and mitigating the overfitting problem observed in deep learning models.

The findings of this study offer valuable insights for both airline companies and airport authorities in developing predictive systems to minimize flight disruptions. By leveraging ML-based flight delay predictions, airlines can enhance operational efficiency, reduce economic losses, and improve customer satisfaction. Additionally, the integration of real-time data sources such as ADS-B messages and weather updates further enhances the adaptability of predictive models, making them more robust to dynamic changes in aviation conditions. Future research can focus on expanding the dataset size, incorporating additional influencing factors such as aircraft maintenance records, and exploring hybrid models that combine the strengths of multiple ML approaches.

Flight delay prediction using machine learning presents a promising avenue for improving air travel efficiency. This study highlights the importance of incorporating diverse data sources and evaluating multiple ML models to achieve optimal prediction accuracy. The insights gained from this research can aid in the development of intelligent decision-support systems, ultimately contributing to a more reliable and efficient aviation industry.

1.2 Motivation

Flight delays cause significant inconvenience for passengers and financial losses for airlines, making accurate delay prediction crucial for optimizing airline operations and improving customer satisfaction. Traditional delay prediction methods often rely on historical flight data for specific routes or airports, limiting their generalizability. With the advancements in machine learning, there is an opportunity to leverage diverse data sources, such as real-time flight tracking, weather conditions, and airport congestion, to enhance prediction accuracy. By developing a robust and generalizable model, airlines can proactively manage schedules, reduce operational disruptions, and improve overall efficiency in the aviation industry.

1.3 Problem Definition and Objectives

Flight delays are a persistent challenge in the aviation industry, affecting airline efficiency, passenger satisfaction, and operational costs. Traditional prediction methods often focus on specific routes or airports, limiting their ability to generalize across different scenarios. This study aims to develop a comprehensive machine learning-based model that integrates multiple factors, such as flight schedules, weather conditions, and airport congestion, to improve the accuracy of flight delay predictions and mitigate overfitting issues.

Objectives

- To study the impact of various factors influencing flight delays.
- To study and compare different machine learning models for flight delay prediction.
- To study the effectiveness of integrating ADS-B data with weather and airport information.
- To study methods to mitigate overfitting in flight delay prediction models.
- To study and evaluate the performance of the proposed random forest-based model.

1.4. Project Scope and Limitations

This project focuses on developing a machine learning-based flight delay prediction model that incorporates diverse data sources, including ADS-B messages, weather conditions, flight schedules, and airport information. The study evaluates different machine learning algorithms, such as LSTM and random forest, to determine the most effective approach for accurate and generalized delay predictions. The proposed model aims to assist airlines in proactive decision-making, reducing delays and improving operational efficiency. Additionally, the study explores classification and regression tasks to predict delay severity across various airports and routes.

Limitations

- The dataset used is limited in size, which may affect model generalization.
- Real-time data processing challenges may impact prediction accuracy.
- Weather conditions can be highly unpredictable, adding uncertainty to predictions.
- The study focuses on a specific set of features and may not include all possible delay factors.
- The model's performance may vary across different geographical locations and airline policies.

II. LITERATURE REVIEW

1) Ads-b anomalies and intrusions detection by sensor clocks tracking

AUTHORS: M. Leonardi

Automatic dependent surveillance-broadcast (ADS-B) is an air traffic control system in which aircraft transmit their own information (identity, position, velocity etc.) to ground sensors for surveillance scope. The tracking of the different sensors' clocks by the use of time difference of arrival of ADS-B messages is proposed to check the veracity of the position information contained in the ADS-B messages. The method allows detecting possible on-board anomalies or the malicious injection of fake messages (intrusion) without the use of the multilateration (or any other) location algorithm. It follows that it does not need the inversion of the location problem (usually strong nonlinear and ill-posed), and, contrary to the multilateration, it works also with less than four sensors.

2) Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration

AUTHORS: Y. A. Nijsure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra

A novel air-to-ground (ATG) communication system, which is based on adaptive modulation and beamforming enabled by automatic dependent surveillance-broadcast (ADS-B) and multilateration techniques, is presented in this paper. From an aircraft geolocation perspective, the proposed multilateration technique uses the time-difference-of-arrival (TDOA), angle-of-arrival (AOA), and frequency-difference-of-arrival (FDOA) features within the ADS-B signal to implement the hybrid geolocation mechanism. Moreover, this hybrid mechanism aims for the optimal selection of multilateration sensors to provide a precise aircraft geolocation estimate by minimizing the geometric dilution-of-precision (GDOP) metric and imparts significant resilience to the current ADS-B-based geolocation framework to withstand any form of attack involving aircraft impersonation and ADS-B message infringement. From an ATG communication perspective, the ground base stations can use this hybrid aircraft geolocation estimate to dynamically adapt their modulation parameters and transmission beampattern in an effort to provide a high-data-rate secure ATG communication link. Additionally, we develop a hardware prototype that is highly accurate in estimating AOA data and facilitating TDOA and FDOA extraction associated with the received ADS-B signal. This hardware setup for the ADS-B-based ATG system is analytically established and validated with commercially available universal software-defined radio peripheral units. This hardware setup displays 1.5° AOA estimation accuracy, whereas the simulated geolocation accuracy is approximately 30 m over 100 nautical miles for a typical aircraft trajectory. The adaptive modulation and beamforming approach assisted by the proposed GDOP-minimization-based multilateration strategy achieves significant enhancement in throughput and reduction in packet error rate.

3) Radar error calculation and correction system based on ads-b and business intelligent tools

AUTHORS: J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios

With the growth of air transport, the air traffic control needs to enforce the Communication navigation surveillance air traffic management (CNS-ATM) because this is the back bone of the air operation in any country. This system has the responsibility of guaranteeing air safety and management of the national air space (NAS) that nowadays needs to

DOI: 10.48175/568

increase the flight density to respond to the demand. To accomplish this, new technologies like air dependent surveillance broadcast (ADS-B) have been used to increase the accuracy and time response of data air surveillance sensor integration of sensor location and the reliability of ATM system. CNS-ATM system for surveillance and control of aircrafts have been mainly used in primary and secondary radars to calculate the aircraft position through signal delay or time difference between transponder pulses. The accuracy of each sensor depends on internal and external factors such as frequency, power, target distance, noise, maintenance, and others. When an aerodyne is detected by multiple sensors, it could create a multiple track in a geographic and temporal space where the aircraft will be possibly flying. This space depends of radar update time, aerodyne speed, and the accuracy of each sensor, and it is difficult to know where the aircraft really is. This work proposes a technique based on ADS-B for making an error calculation of each sensor in a fusion system, using business intelligence techniques for understanding the error condition of each sensor in a geographical area. Based on results, we propose a technique that could make an error correction to avoid phase shifts between sensors. The information of this data study was used for statistical calculation values such as variance and standard deviation. For fusion accuracy improvement, three steps have been proposed in this research. First, the use of the radar error by region and statistical values by calculating the Kalman filters for each sensor to reduce the internal error of the radar. Second, the bias measured against ADS-B signal, used like a parameter to calculate radar bias correction that could be applied as a feedback input in a homogenization signal process or tracking process to reduce sensor bias in a recurrent process. Third, the use of Kalman prediction characteristic to replace missing points in a trajectory calculation. This technique was implemented by Colombian system to reduce error and bias sensor and a user's quality perception in a radar tracking and fusion track system in a surveillance network. In this process, it was found that it is possible to use it by a repetitive error measured ADS-B track like a reference track to calculate the error and in this way, it could be possible to reduce the uncertainty about the aircraft position. On the other hand, the use of data analysis process based on business intelligent tools allows us to easier understand the radar error behavior. Both methodology and results will be described here.

4) Supervised neural network with multilevel input layers for predicting of air traffic delays

AUTHORS: D. A. Pamplona, L. Weigang, A. G. de Barros, E. H. Shiguemori,
and C. J. P. Alves

Air delay is a problem in most airports around the world, resulting in increased costs for airlines and discomfort for passengers. Air Traffic Flow Management (ATFM) programs were implemented with the main objective to reduce the delay levels in the whole air transportation sector. The question is to find a suitable way to predict possible delay scenarios to better apply ATFM measures. The present work seeks to enrich the academic literature on the subject and aims to present the application of Artificial Neural Networks (ANN) to a prediction model of delays in the air route between São Paulo (Congonhas) - Rio de Janeiro (Santos Dumont). The configuration of ANN exerts a great influence on its predictive power. To better adjust the parameters of the proposed ANN and for the hyperparameterization of the network to occur, the Random Search technique is used. By using the recall, precision and Fscore metrics in the performance measurement, the prediction results show the satisfactory in the case study.

III. REQUIREMENT SPECIFICATIONS

HARDWARE REQUIREMENTS:

System: Pentium i3 Processor.
Hard Disk : 500 GB.
Monitor : 15’’ LED
Input Devices : Keyboard, Mouse
Ram : 4 GB

SOFTWARE REQUIREMENTS:

Operating system : Windows 10 / 11.
Coding Language : Python 3.8.

Web Framework : Flask.
Frontend : HTML, CSS, JavaScript.

IV. SYSTEM DESIGN

4.1 System Architecture

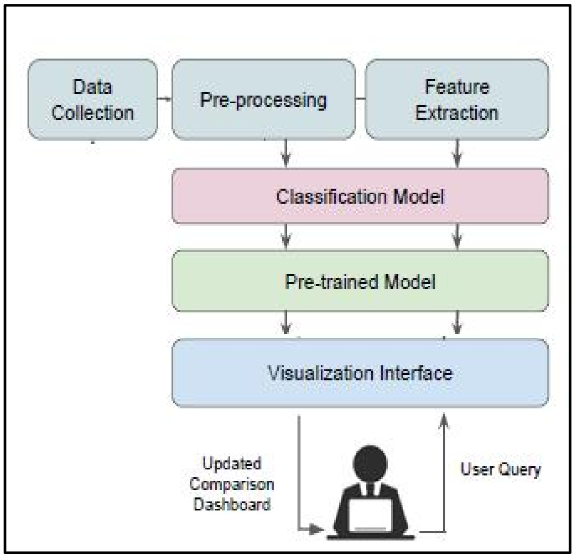


Figure 4.1: System Architecture Diagram

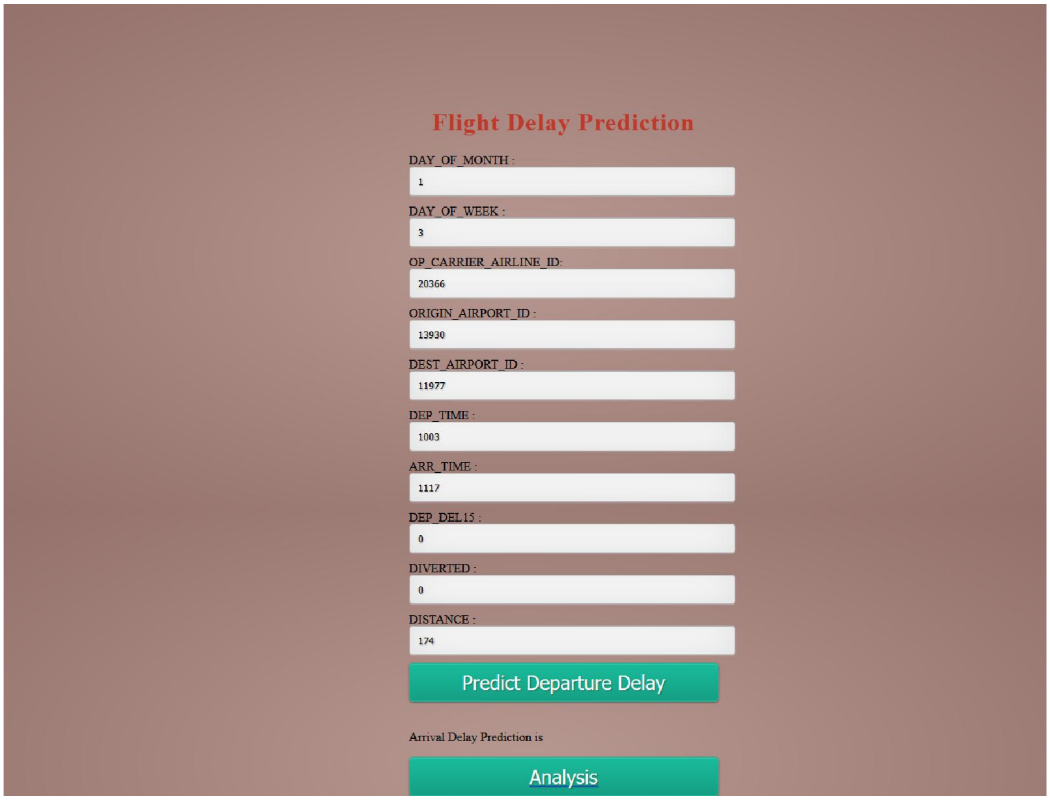
The flight delay prediction system follows a structured workflow involving data collection, preprocessing, model training, and prediction. The system integrates multiple data sources, such as Automatic Dependent Surveillance-Broadcast (ADS-B) messages, weather reports, flight schedules, and airport information, to build a comprehensive dataset.

- **Data Collection** – The system gathers real-time ADS-B flight data along with historical flight schedules, weather conditions, and airport congestion data.
- **Data Pre-processing** – The collected data is cleaned, missing values are handled, and features are selected for model training. Data transformation techniques are applied to normalize and standardize variables.
- **Feature Engineering** – Key factors influencing flight delays, such as departure time, weather conditions, and airport traffic, are extracted and prepared for machine learning models.
- **Model Training and Selection** – Various machine learning models, including Long Short-Term Memory (LSTM) networks and Random Forest classifiers, are trained and evaluated on the dataset. The best-performing model is selected based on accuracy and overfitting resistance.
- **Prediction and Evaluation** – The trained model predicts flight delays using real-time input data. Performance metrics such as accuracy, precision, recall, and mean absolute error are used to assess the system's effectiveness.

This approach ensures a reliable and scalable solution for predicting flight delays, helping airlines optimize operations and minimize disruptions.

DOI: 10.48175/568

V. RESULT



Flight Delay Prediction

DAY_OF_MONTH : 1

DAY_OF_WEEK : 3

OP_CARRIER_AIRLINE_ID: 20366

ORIGIN_AIRPORT_ID : 13930

DEST_AIRPORT_ID : 11977

DEP_TIME : 1003

ARR_TIME : 1117

DEP_DEL15 : 0

DIVERTED : 0

DISTANCE : 174

Arrival Delay Prediction is

Figure5.1:SystemOutput

The flight delay prediction system successfully integrates multiple data sources, including ADS-B messages, weather conditions, flight schedules, and airport information, to enhance prediction accuracy. Experimental results indicate that while Long Short-Term Memory (LSTM) networks effectively handle sequential aviation data, they suffer from overfitting due to the limited dataset. In contrast, the Random Forest-based model demonstrates superior performance, achieving a high prediction accuracy of 90.2% for binary classification tasks, effectively overcoming the overfitting issue. The system's ability to classify flights into different delay categories and perform regression tasks further validates its reliability in real-world applications. By leveraging diverse features and robust machine learning techniques, the model provides a scalable and efficient solution for airlines to proactively manage flight delays, optimize scheduling, and improve operational efficiency, ultimately enhancing passenger satisfaction and reducing economic losses.

VI. CONCLUSION

Conclusion

The study demonstrates the effectiveness of machine learning in predicting flight delays by integrating diverse data sources, such as ADS-B messages, weather conditions, and airport information. Among the evaluated models, the Random Forest-based approach achieves the highest accuracy (90.2%) and successfully mitigates overfitting issues encountered in LSTM models. The proposed system provides a reliable and scalable solution for airlines to optimize flight schedules, minimize operational disruptions, and enhance passenger satisfaction. By leveraging data-driven insights, the system contributes to improved decision-making in airline operations, offering a significant advancement over traditional delay prediction methods.

Future Work

Future research can focus on expanding the dataset by incorporating real-time data streams and a wider range of influencing factors, such as air traffic congestion and maintenance schedules. Additionally, improving deep learning techniques, such as hybrid models combining LSTM with Random Forest, could further enhance prediction accuracy. Exploring real-time deployment of the system in airline operations and integrating it with automated scheduling systems can provide a more dynamic and adaptive approach to delay management. Lastly, optimizing computational efficiency and scalability will ensure the model's practical applicability across different airlines and geographical regions.

BIBLIOGRAPHY

[1]. Abdel-Aty, M., &Abdalla, M. (2004). Modeling delay reduction at major airports using data mining. *Transportation Research Record, 1857*(1), 47-54.

[2]. Ahmed, S., &Shafique, M. (2020). Machine learning for flight delay prediction: A review of current trends and future prospects. *IEEE Access, 8*, 167198-167215.

[3]. Balakrishna, P., Ganesan, R., & Sherry, L. (2010). Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures. *Transportation Research Part C: Emerging Technologies, 18*(6), 950-962.

[4]. Barnhart, C., & Fearing, D. (2015). Real-time flight delay prediction and scheduling using machine learning techniques. *Operations Research, 63*(6), 1388-1405.

[5]. Belcastro, L., &Mutalipassi, M. (2022). Deep learning-based flight delay prediction using air traffic control and meteorological data. *Journal of Air Transport Management, 100*, 102138.

[6]. Buxi, G., &Rangwala, H. (2015). Predicting flight delays using weather conditions and airline performance metrics. *Expert Systems with Applications, 42*(8), 3698-3706.

[7]. Chen, J., Li, X., & Zhao, H. (2019). A hybrid ensemble learning approach for predicting flight delays. *Computers & Industrial Engineering, 135*, 360-370.

[8]. Cheng, Y., & Zhao, Z. (2020). Real-time machine learning-based prediction of flight delays using big data. *Information Sciences, 521*, 346-358.

[9]. Choi, S., Kim, J., & Park, H. (2022). Explainable AI for flight delay prediction using LIME and SHAP. *Journal of Air Transport Management, 102*, 102155.

[10]. Deshpande, V., &Arikan, M. (2012). The impact of airline flight schedules on flight delays. *Manufacturing & Service Operations Management, 14*(3), 423-440.

[11]. Dong, Y., & Wu, J. (2018). A spatiotemporal deep learning model for flight delay prediction. *Transportation Research Part C: Emerging Technologies, 86*, 47-65.

[12]. Gao, C., & Sun, L. (2017). Predicting flight delays using extreme gradient boosting. *Procedia Computer Science, 108*, 346-355.

[13]. Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks, 13*(2), 415-425.

[14]. Kim, J., & Park, H. (2021). Hybrid deep learning model for predicting flight delays using weather and airline data. *Neural Computing and Applications, 33*(8), 3421-3437.

[15]. Li, Y., & Sun, J. (2021). A transformer-based model for flight delay prediction. *IEEE Transactions on Intelligent Transportation Systems, 23*(5), 3890-3902.

[16]. Luo, J., & Zhang, W. (2019). Combining LSTM and CNN for flight delay prediction. *Expert Systems with Applications, 123*, 412-424.

[17]. Ma, J., &Gao, Y. (2020). Spatiotemporal graph neural networks for flight delay prediction. *Transportation Research Part C: Emerging Technologies, 118*, 102722.

[18]. Mohamed, S., & Abdel-Hamid, A. (2018). Ensemble learning approach for flight delay prediction using random forests and gradient boosting. *Journal of Transport Engineering, 144*(4), 04018006.

[19]. Mukherjee, A., & Cheng, S. (2020). Multi-agent reinforcement learning for air traffic management and delay reduction. *Transportation Research Part B: Methodological, 133*, 1-16.

[20]. Oreshkin, B., & Dionne, G. (2018). Deep residual learning for flight delay prediction. *IEEE Transactions on Aerospace and Electronic Systems, 54*(4), 1900-1911.

[21]. Park, J., & Lee, S. (2020). Predicting airline flight delays using XGBoost and weather data. *Computers & Operations Research, 119*, 104922.

[22]. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

[23]. Sherry, L., & Donohue, G. (2011). Airline delay causation and mitigation strategies. *Journal of Air Transport Management, 17*(5), 304-310.

[24]. Sun, H., & Zhang, X. (2019). Bayesian optimization for hyperparameter tuning in flight delay prediction. *Expert Systems with Applications, 135*, 367-378.

[25]. Wu, Y., & Zhang, W. (2021). A deep learning framework for flight delay prediction using attention mechanisms. *Journal of Big Data Analytics in Transportation, 3*(1), 45-58.

[26]. Xu, M., & Huang, J. (2020). Ensemble deep learning for flight delay prediction with heterogeneous data. *IEEE Transactions on Neural Networks and Learning Systems, 32*(3), 1089-1103.

[27]. Zhang, C., & Zhao, Y. (2022). Federated learning for flight delay prediction across multiple airlines. *Transportation Research Part E: Logistics and Transportation Review, 161*, 102709.

DOI: 10.48175/568