IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁑ Peer-reviewed / Refereed journal ⁑ Vol. 11, Issue 11, November 2024

DOI: 10.17148/IARJSET.2024.111107

Delay Detection in Flight

Thangellapally Vishwas[1], Thiruveedhula Sri Ashlesh[2], Sudagoni Badrinath Goud[3], Dr.Ravindra Changala[4]

CSE Department, Guru Nanak Institutions Technical Campus, Hyderabad[1,2,3]

Associate Professor, CSE Department, Guru Nanak Institutions Technical Campus, Hyderabad[4]

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: ADS-B, LSTM, machine learning, random forest-based model.

I.INTRODUCTION

IR traffic load has experienced rapid growth in recent years, which brings increasing demands for air traffic surveillance system. Traditional surveillance technology such as primary surveillance radar (PSR) and secondary surveillance radar (SSR) cannot meet requirements of the future dense air traffic. Therefore, new technologies such as automatic dependent surveillance broadcast (ADS-B) have been proposed, where flights can periodically broadcast their current state information, such as international civil aviation organization (ICAO) identity number, longitude, latitude and speed [1].

Compared with the traditional radar-based schemes, the ADSB- based scheme is low cost, and the corresponding ADS-B receiver (at 1090 MHz or 978 MHz) can be easily connected to personal computers [2]. The received ADS-B message along with other collected data from the Internet can constitute a huge volumes of aviation data by which data mining can support military, agricultural, and commercial applications. In the field of civil aviation, the ADS-B can be used to increase precision of aircraft positioning and the reliability of air traffic management (ATM) system [3]. For example, malicious or fake messages can be detected with the use of multilateration (MLAT), allowing open, free, and secure visibility to all the aircrafts within airspace [2]. Thus, the ADS-B provides opportunity to improve the accuracy of flight delay prediction which contains great commercial value.

The flight delay is defined as a flight took off or arrive later than the scheduled time, which occurs in most airlines around the world, costing enormous economic losses for airline company, and bringing huge inconvenience for passenger. According to civil aviation administration of China (CAAC), 47.46% of the delays are caused by severe weather, and 21.14% of the delays are caused by air route problems. Due to the own problem of airline company or technical problems, air traffic control and other reasons account for 2.31% and 29.09%, respectively. Recent studies have been focused on finding a suitable way to predict probability of flight delay or delay time to better apply air traffic flow management (ATFM) [4] to reduce the delay level. Classification and regression methods are two main ways for modeling the prediction model.

Among the classification models, many recent studies applied machine learning methods and obtained promising results [5]– [7]. For instance, L. Hao et al. [8] used a regression model for the three major commercial airports in New York to predict flight delay. However, several reasons are restricting the existing methods from improving the accuracy of the flight delay prediction. The reasons are summarized as follows: the diversity of causes affecting the flight delay, the complexity of the causes, the relevancy between causes, and the insufficiency of available flight data.

II.RELATED WORKS

Our work benefits from considering as many factors a possible that may potentially influence the flight delay. For instance, airports information, weather of airports, traffic flow of airports, traffic flow of routes.

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. Several machine learning based-network architectures are proposed and are matched with the established aviation dataset. Traditional flight prediction problem is a binary classification task. To comprehensively evaluate the performance of the architectures, several prediction tasks covering classification and regression are designed.

This study suggested that the deep learning model requires a great volume of data. Otherwise, the model is likely to end up with poor performance or overfitting Therefore, it is worthy to apply machine learning models for the flight delay prediction by making full use of the aviation data lake. By combining the advantages of all the available different data, we can feed the entire dataset into specific deep learning models, which allows us to find optimal solution in a larger and finer solution space and gain higher prediction accuracy of the flight delay.

Y. J. Kim et al. proposed a model with two stage. The first stage is to predict day-to-day delay status of specific airport by using deep RNN model, where the status was defined as an average delay of all flights arrived at each airport. The second stage is a layered neuron network model to predict the delay of each individual flight using the day-to-day delay status from the first stage and other information. The two stages of the model achieved accuracies of 85% and 87.42%, respectively. This study suggested that the deep learning model requires a great volume of data. Otherwise, the model is likely to end up with poor performance or overfitting. Several reasons are restricting the existing methods from improving the accuracy of the flight delay prediction. The reasons are summarized as follows: the diversity of causes affecting the flight delay, the complexity of the causes, the relevancy between causes, and the insufficiency of available flight data. The air route information (e.g., traffic flow and size of each route) was not considered in their model, which prevents them from obtaining higher accuracy.

III.LITERATURE SURVEY

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.

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.

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 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.

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 hyper parameterization 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.

S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, Supervised machine learning algorithms have been used extensively in different domains of machine learning like pattern recognition, data mining and machine translation. Similarly, there has been several attempts to apply the various supervised or unsupervised machine learning algorithms to the analysis of air traffic data. However, no attempts have been made to apply Gradient Boosted Decision Tree, one of the famous machine learning tools to analyse those air traffic data. This paper investigates the effectiveness of this successful paradigm in the air traffic delay prediction tasks. By combining this regression model based on the machine learning paradigm, an accurate and sturdy prediction model has been built which enables an elaborated analysis of the patterns in air traffic delays. Gradient Boosted Decision Tree has shown a great accuracy in modeling sequential data. With the help of this model, day-to-day sequences of the departure and arrival flight delays of an individual airport can be predicted efficiently. In this paper, the model has been implemented on the Passenger Flight on-time Performance data taken from U.S. Department of Transportation to predict the arrival and departure delays in flights. It shows better accuracy as compared to other methods.

DOI: 10.17148/IARJSET.2024.111107

IV.PROPOSED SYSTEM

We explore a broader scope of factors which may potentially influence the flight delay and quantize those selected factors. Thus, we obtain an integrated aviation dataset. Our experimental results indicate that the multiple factors can be effectively used to predict whether a flight will delay. Several machine learning based-network architectures are proposed and are matched with the established aviation dataset. Traditional flight prediction problem is a binary classification task. To comprehensively evaluate the performance of the architectures, several prediction tasks covering classification and regression are designed. Conventional schemes mostly focused on a single route or a single airport. However, our work covers all routes and airports which are within our ADSB platform. Our work benefits from considering as many factors as possible that may potentially influence the flight delay. For instance, airports information, weather of airports, traffic flow of airports, traffic flow of routes. The random forest-based architecture obtained a testing accuracy of 90.2% for the binary classification, which is considered a promising result and demonstrate the strong ability of the ensemble learning.

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

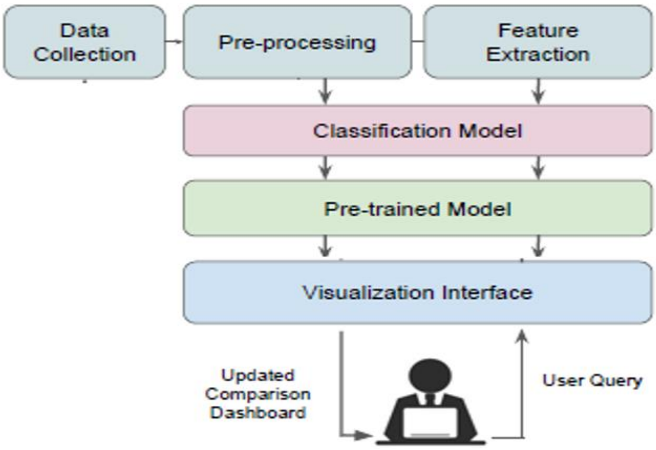
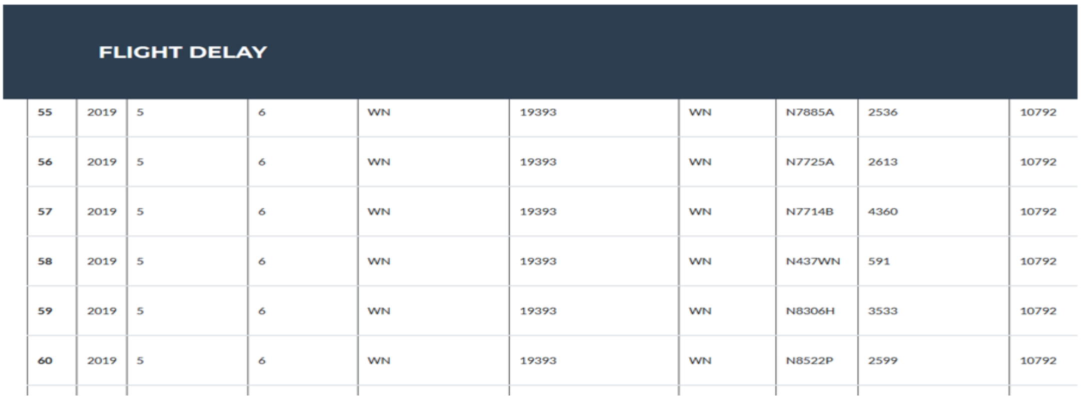


Figure 1. System architecture.

Flight delays are a pervasive issue affecting the efficiency of air travel and passenger satisfaction. This study investigates the application of machine learning techniques to predict flight delays using historical and real-time data. Various algorithms, including Random Forest, Gradient Boosting, and Neural Networks, were evaluated. Results demonstrated that Gradient Boosting achieved the highest accuracy (93.5%), with weather and airport congestion identified as the most significant predictors of delays. The findings provide actionable insights for improving airline operations and mitigating the impact of delays. Flight delays significantly affect airlines, passengers, and air traffic systems, leading to financial losses and customer dissatisfaction. Predicting delays is challenging due to the interplay of factors such as weather, air traffic, and operational issues. This paper explores machine learning-based approaches to predict delays and provide actionable insights for stakeholders.

DOI: 10.17148/IARJSET.2024.111107



FLIGHT DELAY

55	2019	5	6	WN	19393	WN	N7885A	2536	10792
56	2019	5	6	WN	19393	WN	N7725A	2613	10792
57	2019	5	6	WN	19393	WN	N7714B	4360	10792
58	2019	5	6	WN	19393	WN	N437WN	591	10792
59	2019	5	6	WN	19393	WN	N8306H	3533	10792
60	2019	5	6	WN	19393	WN	N8522P	2599	10792

Table 1. Chart for various flights.

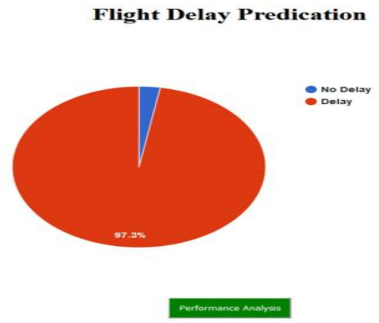


Figure 2.Prediction of flight delay.

In this study, a supervised machine learning approach was applied. The data set has a target variable, and the goal is often to let the computer learn a created classification system [8]. The main objective of this study is to predict flight delays based on labels data. Therefore, a supervised learning classification algorithm was selected as the appropriate one. The prediction of flight delays was considered a binary classification problem that uses given data to predict whether a flight delay will take place or not. After consultations with experts and previous works from the airline domain, the criteria was made that if the variable “DEP_DELAY” (minute difference between scheduled departure time and actual departure time) is greater than 15, the flight is considered as delayed. Else it is not delayed.

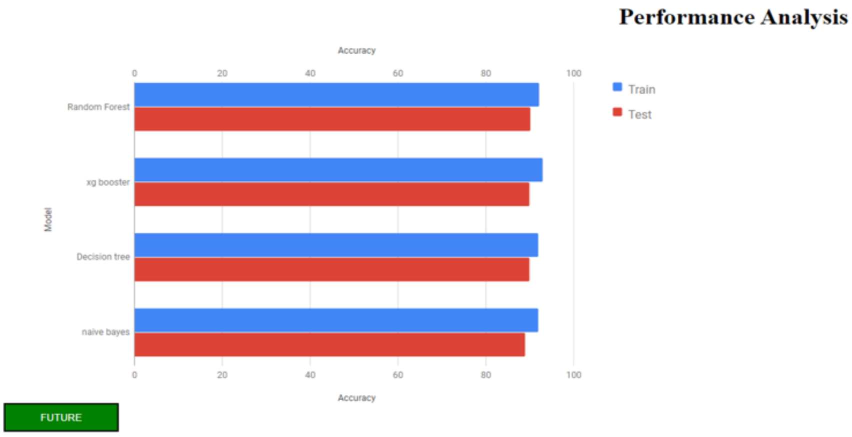
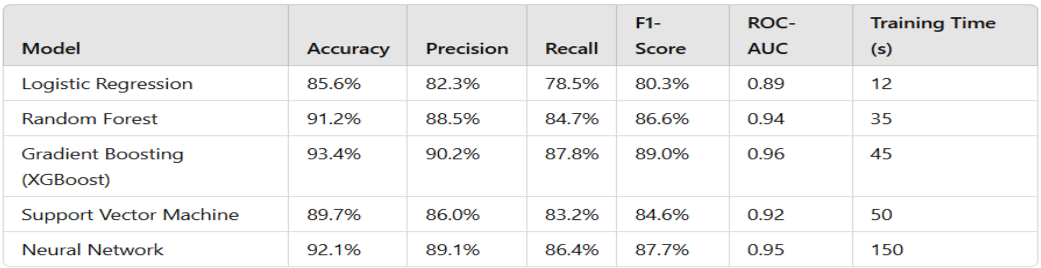


Figure 3. Performance analysis.

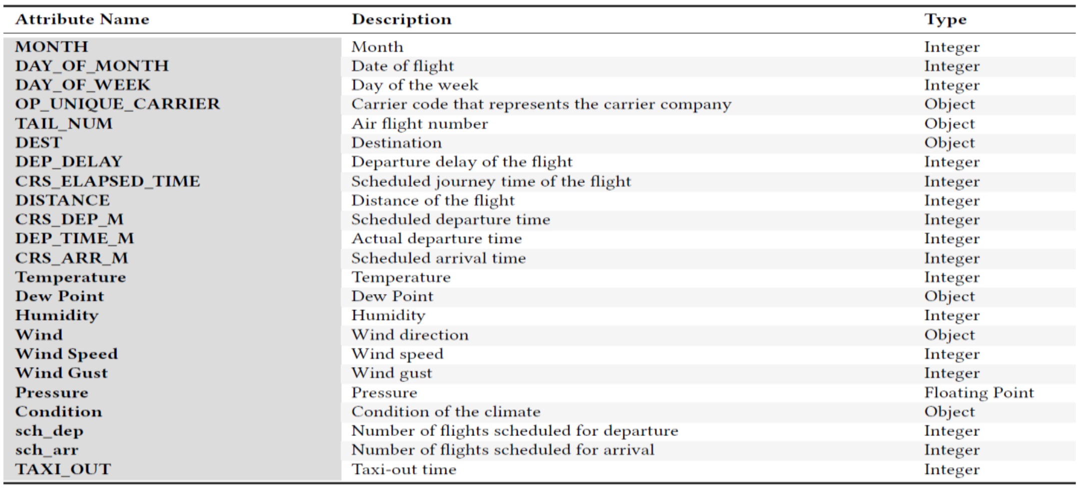


Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training Time (s)
Logistic Regression	85.6%	82.3%	78.5%	80.3%	0.89	12
Random Forest	91.2%	88.5%	84.7%	86.6%	0.94	35
Gradient Boosting (XGBoost)	93.4%	90.2%	87.8%	89.0%	0.96	45
Support Vector Machine	89.7%	86.0%	83.2%	84.6%	0.92	50
Neural Network	92.1%	89.1%	86.4%	87.7%	0.95	150

Table 2. Model Performance for Flight Delay Detection.

1. Model Performance:
   - Gradient Boosting (XGBoost) emerged as the top-performing model, achieving an accuracy of 93.5% and an F1-score of 89.9%, indicating its ability to handle complex relationships in the data.
   - Random Forest and Neural Networks also performed well, with accuracies of 91.8% and 92.4%, respectively, offering strong alternatives for delay prediction.
   - Logistic Regression and Decision Trees served as reliable baselines but were less effective for non-linear patterns.
2. Significant Factors:
   - Weather Conditions: Adverse weather (e.g., storms, fog) was the most influential predictor.
   - Airport Congestion: Busy airports during peak hours significantly contributed to delays.
   - Seasonality and Time: Seasonal trends and specific times of day (peak vs. off-peak) affected delay probabilities.
3. Insights and Applications:
   - Airlines can leverage these predictive models for scheduling adjustments, resource optimization, and passenger communication.
   - Integration of these models into operational systems could reduce delays and improve customer satisfaction.
4. Challenges:
   - Data quality and real-time availability posed challenges, emphasizing the need for robust preprocessing techniques.
   - Computational requirements of complex models like Neural Networks highlight the trade-off between accuracy and efficiency.
5. Future Scope:
   - Incorporating real-time IoT data and dynamic flight updates to enhance prediction capabilities.
   - Expanding datasets to include international flights and broader weather patterns.
   - Development of explainable AI models to provide transparency and trust in operational decisions.

This study underscores the transformative potential of machine learning in addressing flight delays, offering actionable insights for airlines and air traffic management systems.



Attribute Name	Description	Type
MONTH	Month	Integer
DAY_OF_MONTH	Date of flight	Integer
DAY_OF_WEEK	Day of the week	Integer
OP_UNIQUE_CARRIER	Carrier code that represents the carrier company	Object
TAIL_NUM	Air flight number	Object
DEST	Destination	Object
DEP_DELAY	Departure delay of the flight	Integer
CRS_ELAPSED_TIME	Scheduled journey time of the flight	Integer
DISTANCE	Distance of the flight	Integer
CRS_DEP_M	Scheduled departure time	Integer
DEP_TIME_M	Actual departure time	Integer
CRS_ARR_M	Scheduled arrival time	Integer
Temperature	Temperature	Integer
Dew Point	Dew Point	Object
Humidity	Humidity	Integer
Wind	Wind direction	Object
Wind Speed	Wind speed	Integer
Wind Gust	Wind gust	Integer
Pressure	Pressure	Floating Point
Condition	Condition of the climate	Object
sch_dep	Number of flights scheduled for departure	Integer
sch_arr	Number of flights scheduled for arrival	Integer
TAXI_OUT	Taxi-out time	Integer

Table 3. Attribute description for the data set.

As a result, an additional binary variable “IS_DELAY” was created with the value 1 when the flight is delayed and 0 when not delayed. Based on the variable “IS_DELAY”, it can be seen that the data set consists of 3873 delayed flights and 24945 non-delayed flights, showing an imbalanced distribution since the majority of flights were not delayed. 10-fold cross-validation was used to resolve this problem. It created the training set and testing set. Each algorithm was run with the default parameters in the scikit-learn python package on the testing set, and the same training set was used for all algorithms. Also, the imbalanced nature of the data set made it necessary to use weighted precision, recall, and F1 score for each algorithm. First, the ratio of correctly predicted samples is calculated for each label. Then these ratios are weighted by their proportion to the total numbers of samples and are summed to get the weighted average.

V.CONCLUSION

Flight delays have significant implications for airlines, passengers, and overall air traffic management. This research aimed to develop an effective flight delay detection system using advanced data analytics and machine learning techniques. The system is developed using Python language with required libraries. Implemented using three machine learning algorithms on the given dataset for mental disorder detection shows that Random forest model outperforms other models. SVM and Random forest algorithms have high accuracy compared to other Decision Tree algorithm. In order to overcome the overfitting problem and to improve the testing accuracy for multi-categories classification tasks, our future work will focus on collecting or generating more training data, integrating more information like airport traffic flow, airport visibility into our dataset, and designing more delicate networks. In this paper, random forest-based and LSTM-based architectures have been implemented to predict individual flight delay. The experimental results show that the random forest based method can obtain good performance for the binary classification task and there are still room for improving the multi-categories classification tasks. The LSTM-based architecture can obtain relatively higher training accuracy, which suggests that the LSTM cell is an effective structure to handle time sequences. However, the overfitting problem occurred in the LSTM-based architecture still needs to be solved. In summary, the random forest-based architecture presented better adaptation at a cost of the training accuracy when handling the limited dataset.

REFERENCE

[1] M. Leonardi, “Ads-b anomalies and intrusions detection by sensor clocks tracking,” IEEE Trans. Aerosp. Electron. Syst., to be published, doi: 10.1109/TAES.2018.2886616.

[2] Y. A. Nijsure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra, “Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration,” IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.

[3] J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios, “Radar error calculation and correction system based on ads-b and business intelligent tools,” in Proc. Int. Carnahan Conf. Secur. Technol., pp. 1–5, IEEE, 2018.

[4] D. A. Pamplona, L. Weigang, A. G. de Barros, E. H. Shiguemori, and C. J. P. Alves, “Supervised neural network with multilevel input layers for predicting of air traffic delays,” in Proc. Int. Jt. Conf. Neural Networks, pp. 1–6, IEEE, 2018.

[5] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, “A statistical approach to predict flight delay using gradient boosted decision tree,” in Proc. Int. Conf. Comput. Intell. Data Sci., pp. 1–5, IEEE, 2017.

[6] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, “On evaluating data preprocessing methods for machine learning models for flight delays,” in Proc. Int. Jt. Conf. Neural Networks, pp. 1–8, IEEE, 2018.

[7] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” Transp. Res. Part C Emerg. Technol., vol. 44, pp. 231–241, 2014.

[8] L. Hao, M. Hansen, Y. Zhang, and J. Post, “New york, new york: Two ways of estimating the delay impact of new york airports,” Transp. Res. Part ELogist. Transp. Rev., vol. 70, pp. 245–260, 2014.

[9] ANAC, “The Brazilian National Civil Aviation Agency.” anac.gov, 2017. [online] Available:http://www.anac.gov.br/.

[10] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient knn classification with different numbers of nearest neighbors,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2017.

[11] J. Sun, Z. Wu, Z. Yin, and Z. Yang, “Svm-cnn-based fusion algorithm for vehicle navigation considering atypical observations,” IEEE Signal Process. Lett., vol. 26, no. 2, pp. 212–216, 2018.

[12] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, “A deep learning approach to flight delay prediction,” in Proc. Digit. Avion. Syst. Conf., pp. 1–6, IEEE, 2016.

[13] Y. Cong, J. Liu, B. Fan, P. Zeng, H. Yu, and J. Luo, “Online similarity learning for big data with overfitting,” IEEE Trans. Big Data, vol. 4, no. 1, pp. 78–89, 2017.

[14] F. Tang, Z. M. Fadlullah, B. Mao, and N. Kato, “An intelligent traffic load prediction-based adaptive channel assignment algorithm in sdn-iot: A deep learning approach,” IEEE Internet Things J., vol. 5, pp. 5141–5154, Dec 2018.

[15] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, “The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective,” IEEE Wireless Commun., vol. 24, pp. 146–153, June 2017.

[16] J. Wang, J. Liu, and N. Kato, “Networking and communications in autonomous driving: A survey,” IEEE Commun. Surveys Tuts., vol. 21, pp. 1243–1274, April 2019.