IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

FLIGHT DELAY PREDICTION BASED ON AVIATION BIG DATA AND MACHINE LEARNING

Jakkamsetti Kiran Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

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 surveillancebroadcast (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.

Index Terms:

Flight delay prediction, ADS-B, machine learning, LSTM neural network, random forest.

1. INTRODUCTION

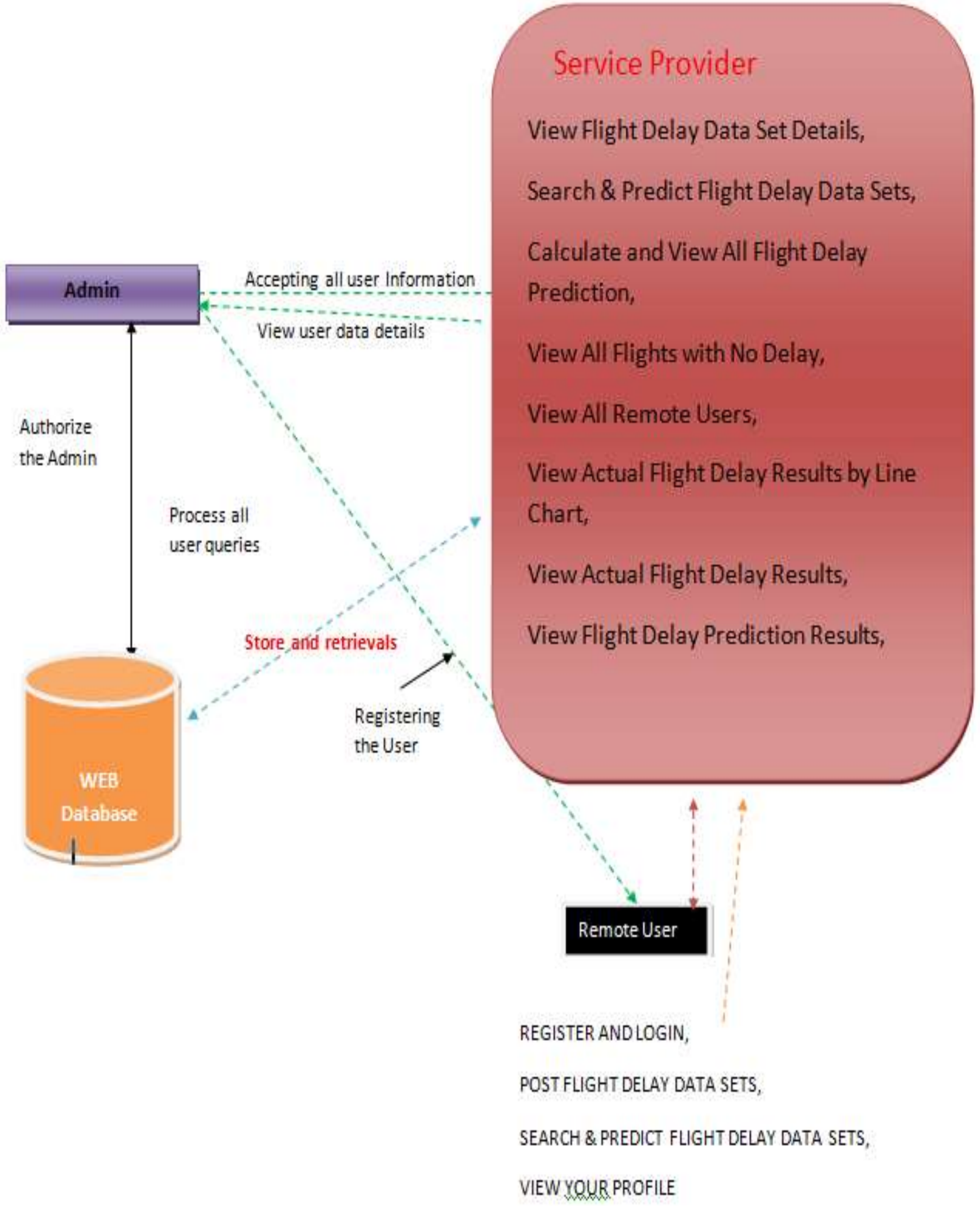
A IR traffic load has experienced rapid growth in recent years, which brings increasing demands for air traffic surveillance system. Traditional surveillance technology such as primary surveillance radar (PSR) and secondary surveillance radar (SSR) cannot meet requirements of the future dense air traffic. 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 This work was supported by the Project Funded by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant TC190A3WZ-2, National Natural Science Foundation of China under Grant 61901228, Jiangsu Specially Appointed Professor Program under Grant RK002STP16001, Summit of the Six Top Talents Program of Jiangsu under Grant XYDXX-010, Program for High-Level Entrepreneurial and Innovative Talents Introduction under Grant CZ0010617002, and 1311 Talent Plan of Nanjing University of Posts and

Telecommunications. (Corresponding authors: Jinlong Sun and Jie Yang) The authors are with College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (E-mails: {guiguan, 1018010402, sunjinlong, jyang, 1218012005, 1218012004}@njupt.edu.cn). 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) [1], 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 arrived 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.

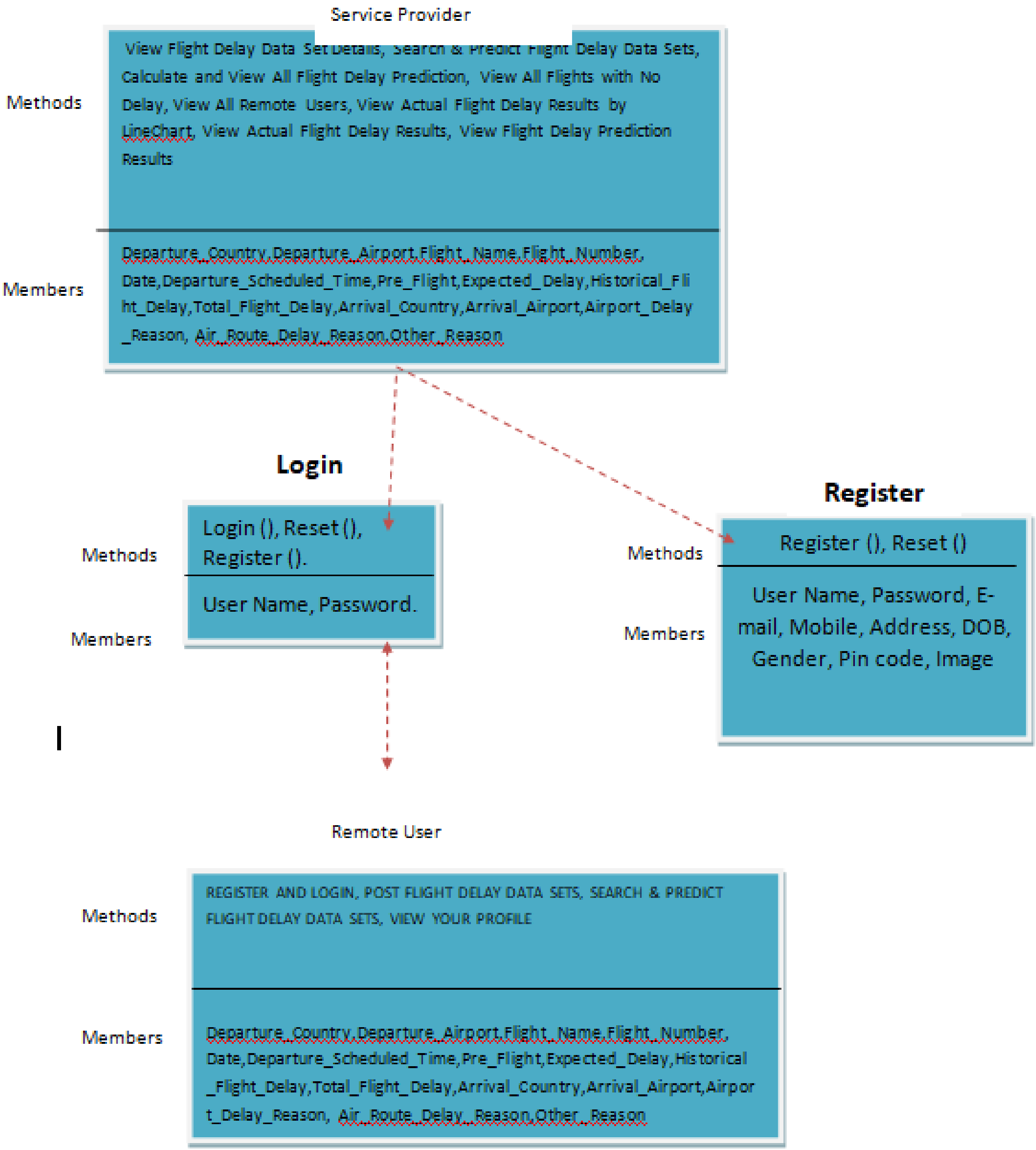
[8] used a regression model for the three major commercial airports in New York to predict flight delay. However, several reasons are restricting the existing methods from improving the accuracy of the flight delay prediction. The reasons are summarized as follows: the diversity of causes affecting the flight delay, the complexity of the causes, the relevancy between causes, and the insufficiency of available flight data. In [6], a public dataset named VRA [9] was used to compare the performance of several machine learning models including k-nearest neighbors (K-NN) [10], support vector machines (SVM) [11], naive Bayes classifier, and random forests for predicting flight delay, and achieved the best accuracy of 78.02% among the four schemes. However, 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. In [4], D. A. Pamplona et al. built an artificial neural network with 4 hidden layers, and achieved the highest accuracy of 87%; their proposed model suggested that the day of the week, block hour, and route has great influence on the flight delay. This model did not consider meteorological factors, so there is room for improvement. Y. J. Kim et al. [12] 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 volumes of data. Otherwise, the model is likely to end up with poor performance or overfitting [13]. To address the problems in ATM, the received ADS-B messages can be coupled with weather information, traffic flow information, and other information to constitute an aviation data lake, which provides an opportunity to find a better approach to accurately predict the flight delay. Meanwhile, machine learning have made great progress and have obtain amazing performance in many domains, such as internet of things [14], heterogeneous network traffic control [15], autonomous driving [16], unmanned aerial vehicle [17]–[21], wireless communications [22]–[28], and cognitive radio [29]–[31]. The above successes motivate us to apply machine learning in the field of air traffic data analytic applications [12], [32]. Compared with the scenarios in wireless communications, the air traffic also faces dynamic environment and can be affected by many dynamic factors. 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. 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 contributions of this paper can be summarized as follows: • 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 [4], [6], [12]. However, our work covers all routes and airports which are within our ADS-B platform.

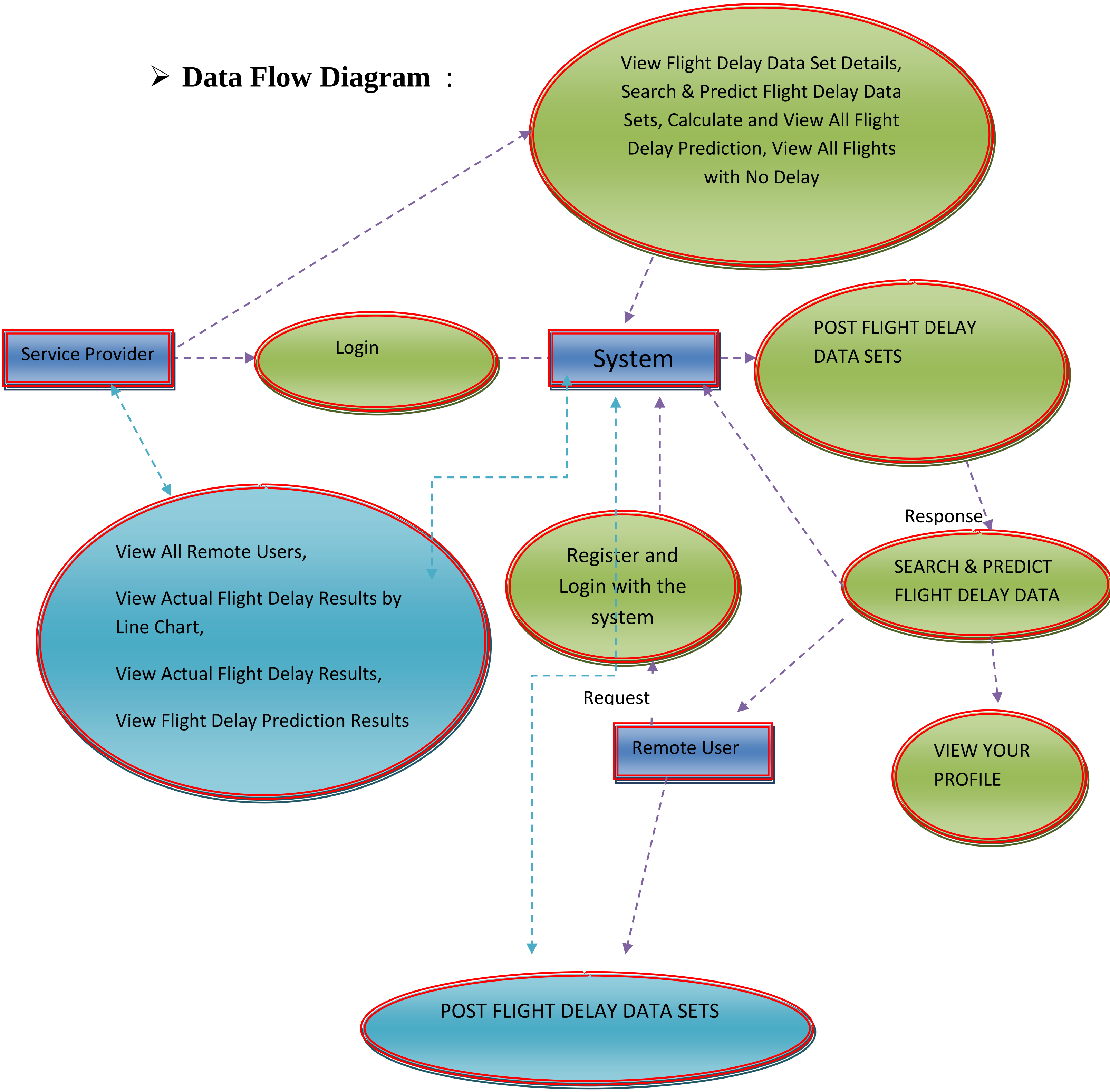
2. ARCHITECTURE DIAGRAM


Service Provider
View Flight Delay Data Set Details,
Search & Predict Flight Delay Data Sets,
Calculate and View All Flight Delay Prediction,
View All Flights with No Delay,
View All Remote Users,
View Actual Flight Delay Results by Line Chart,
View Actual Flight Delay Results,
View Flight Delay Prediction Results,
Admin
Accepting all user Information
View user data details
Authorize the Admin
Process all user queries
Store and retrievals
Registering the User
WEB Database
Remote User
REGISTER AND LOGIN,
POST FLIGHT DELAY DATA SETS,
SEARCH & PREDICT FLIGHT DELAY DATA SETS,
VIEW YOUR PROFILE

➢ Class Diagram :


Service Provider
Methods
View Flight Delay Data Set Details, Search & Predict Flight Delay Data Sets, Calculate and View All Flight Delay Prediction, View All Flights with No Delay, View All Remote Users, View Actual Flight Delay Results by LineChart, View Actual Flight Delay Results, View Flight Delay Prediction Results
Members
Departure_Country,Departure_Airport,Flight_Name,Flight_Number, Date,Departure_Scheduled_Time,Pre_Flight,Expected_Delay,Historical_Flight_Delay,Total_Flight_Delay,Arrival_Country,Arrival_Airport,Airport_Delay_Reason, Air_Route_Delay_Reason,Other_Reason
Login
Methods
Login (), Reset (), Register ().
Members
User Name, Password.
Register
Methods
Register (), Reset ()
Members
User Name, Password, E-mail, Mobile, Address, DOB, Gender, Pin code, Image
Remote User
Methods
REGISTER AND LOGIN, POST FLIGHT DELAY DATA SETS, SEARCH & PREDICT FLIGHT DELAY DATA SETS, VIEW YOUR PROFILE
Members
Departure_Country,Departure_Airport,Flight_Name,Flight_Number, Date,Departure_Scheduled_Time,Pre_Flight,Expected_Delay,Historical_Flight_Delay,Total_Flight_Delay,Arrival_Country,Arrival_Airport,Airport_Delay_Reason, Air_Route_Delay_Reason,Other_Reason

➢ Data Flow Diagram :


View Flight Delay Data Set Details, Search & Predict Flight Delay Data Sets, Calculate and View All Flight Delay Prediction, View All Flights with No Delay
Service Provider
Login
System
POST FLIGHT DELAY DATA SETS
Response
SEARCH & PREDICT FLIGHT DELAY DATA
Register and Login with the system
View All Remote Users,
View Actual Flight Delay Results by Line Chart,
View Actual Flight Delay Results,
View Flight Delay Prediction Results
Request
Remote User
VIEW YOUR PROFILE
POST FLIGHT DELAY DATA SETS

3. CONCLUSIONS

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 over fitting 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. 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.

4. REFERENCES

[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.