Delay Prediction of Aircraft Using Machine Learning Classifiers

S. V. Satya Krishna
Associate professor and
Head of the department
Data Science

L. Gangothri
B. Tech Student
Data science
Siddhartha Institute of
Technology and sciences

Vijay Nithin Kumar
B. Tech Student
Data Science
Siddhartha Institute of
Technology and sciences

K. Venkatesh
B. Tech Student
Data Science
Siddhartha Institute of
Technology and Sciences

K. Pavan Kalyan
B. Tech Student
Data Science
Siddhartha Institute of
Technology and Sciences

M. Akshitha
B. Tech Student
Data Science
Siddhartha Institute of
Technology and Sciences

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 over fitting 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 over fitting problem.

1. INTRODUCTION

AIR 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.

1.1 MOTIVATION

AIR 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.

1.2 PROBLEM STATEMENT

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. 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. 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.

1.3 PURPOSE OF THE SYSTEM

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. For example, malicious or fake messages can be detected with the use of multi alteration (MLAT), allowing open, free, and secure visibility to all the aircrafts within airspace. 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.

1.4 SCOPE OF THE PROJECT

1.5 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.

OBJECTIVE

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 a public dataset named VRA was used to compare the performance of several machine learning models including k-nearest neighbors (K-NN), support vector machines (SVM), 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 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.

2. LITERATURE REVIEW

Here we will elaborate the aspects like the literature survey of the project and what all projects are existing and been actually used in the market which the makers of this project took the inspiration from and thus decided to go ahead with the project covering with the problem statement.

2.1 Literature Survey

Delayed flights are a major problem for the airline industry, causing financial loss and inconvenience to passengers. The rapid growth of the civil aviation industry has led to overcrowding and frequent delays at most major airports worldwide. Several studies have examined a variety of variables that affect flight delays, such as traffic volume, aircraft type, maintenance, airline operations, weather conditions, procedural changes enroute, capacity limitations, customer service issues, and delays due to the late arrival of aircraft or crew. Weather is a significant factor in flight delays, contributing to about 69 per cent of such incidents. In contrast, airport congestion accounts for about 32 per cent of the flight delays. These variables also influence departure times, flight routes and arrival times, leading to greater airport air traffic unpredictability.

This has led to recent studies investigating flight delay problems by applying various machine learning methodologies that integrate mathematics, statistics and computer science concepts. Machine learning can potentially exceed the constraints of mathematical formulas and elevate the precision of predicting flight delays. Various machine learning methods offer distinct attributes, such as supervised, unsupervised, deep reinforcement, and ensemble learning. The appropriate way and algorithm selection are critical for research, as underperforming algorithms result in imprecise outcomes and squander computational power. Hence, the process of algorithm selection holds significance in the domain of machine learning technology. A similar study used a support vector machine model to investigate the non-linear relationship between flight delays using individual flight data from three airports to identify patterns and causes of air traffic delays. The authors examined weather information, airport ground operation, demand capacity, and flow management characteristics. They found that pushback delay, taxi-out delay, ground delay program, and demand-capacity imbalance had the highest probabilities and were significantly related to flight departure delay. These findings provide insights into the causes of flight delays and can guide future research in this area.

To enhance the effectiveness of airlines and airports and mitigate the adverse effects of flight delays on passengers, it is imperative to adopt measures to reduce such delays. An early study developed a model to predict delays in arrivals at Hartsfield-Jackson International Airport, utilising historical flight, weather data, aeroplane information and delay propagation to train the model. Various sampling techniques such as DT, RF, and Multilayer Perceptron were employed to overcome the issue of unbalanced datasets. Among these techniques, the Multilayer Perceptron model was the most accurate.

Research on flight delay prediction has used ensemble learning techniques, but most have focused on a hub and spoke network for a specific route. This approach must capture the factors influencing flight delays across different routes or airports. This study proposes to use ensemble learning techniques to build a flight delay prediction model from data from international and regional airports in a point-to-point network, addressing previous studies' limitations and potentially developing a more accurate and reliable flight delay prediction model. This will allow the model to learn the patterns and relationships between flight delays and various variables, such as departure time, departure time block, arrival time, type of aircraft, air traffic, and airport operations, across a broader range of airports and routes. Ensemble learning techniques typically outperform individual machine learning algorithms on tasks such as flight delay prediction.

2.2 System Analysis

System analysis involves dissecting the flight delay prediction system, examining its components, data flow, processes, and interactions. This includes assessing requirements, data handling, performance, integration, scalability, security, usability, and feedback mechanisms. By thoroughly analyzing these aspects, developers gain insights to optimize the system's design, functionality, and user experience, ensuring it effectively meets the needs of airlines, airports, and passengers for accurate and timely flight delay predictions.

2.2.1 Existing System:

- Nowadays, aircrafts have become a necessity because they easy life. They are efficient in carrying goods and passengers around the world. It also supplies emergencies in warfare and takes a vital role in carrying medical necessities. Hence, advent of airplanes is considered important. Delays in aircrafts can affect thousands of people across the globe either directly or indirectly. There are a lot of reasons of delays in aircrafts such as critical weather, security issues, traffic and many more.

- There are several methods implemented in the existing system to predict the flight delays but due to various complexities of the ATFM and the huge datasets involved, it has become very difficult to find an accurate solution for this complication. Many algorithms have been implemented to forecast flight delays. We are using Python in Visual Studio Code. We implement Binary Classification to prepare a model that can predict the delays.

2.2.2 Disadvantages:

- In the existing system, the system is not using Data Transformation and Balancing.
- This system is less performance due to lack of Data Cleaning and Data Integration.

- **Proposed System:**

The proposed 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:

The system explores 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.

2.2.4 Advantages:

- Proposed methods implementing ADS-B Message Based Aviation Big Data Platform which is more effective and faster.
- ADS-B system is a communication and surveillance integrated system for air traffic management (ATM) where flights periodically broadcast location and other information on the same frequency band.

3. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

3.1. Request Clarification: After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires. Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So, taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

3.2. Feasibility Study: An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- Operational Feasibility
- Economic Feasibility
- Technical Feasibility

3.2.1. Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

3.2.2. Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer-based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at any time. The Virtual Private Network is to be developed using the existing resources of the organization. So, the project is economically feasible.

3.2.3. Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

3.3. Request Approval: Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine

where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

4. MODULES

a. Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as 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 User, View Actual Flight Delay Results by Line Chart, View Actual Flight Delay Results, View Flight Delay Prediction Results.

b. Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like POST FLIGHT DELAY DATA SETS,SEARCH & PREDICT FLIGHT DELAY DATA SETS,VIEW YOUR PROFILE.

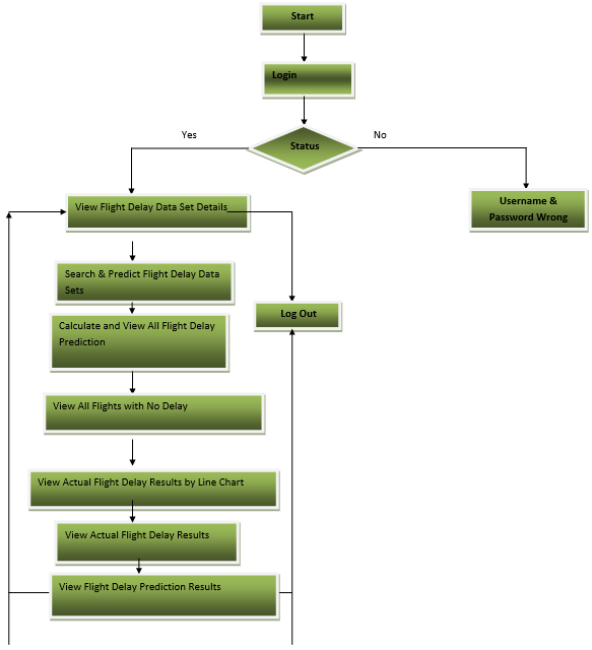


Fig.a. service Provider Flow-Chart

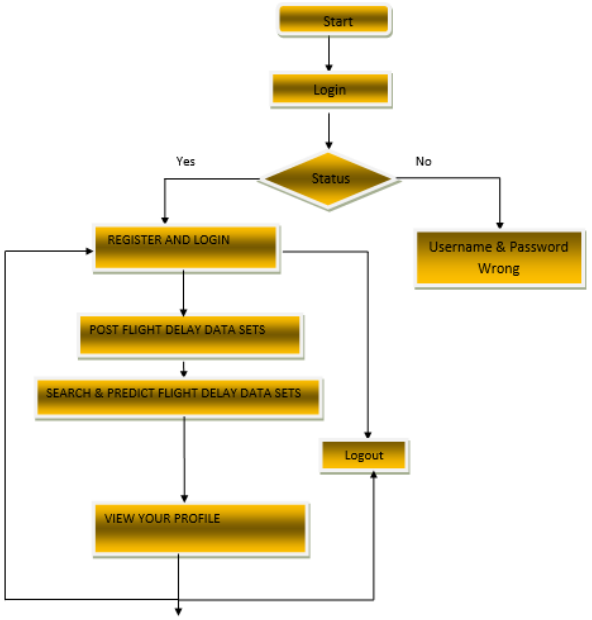
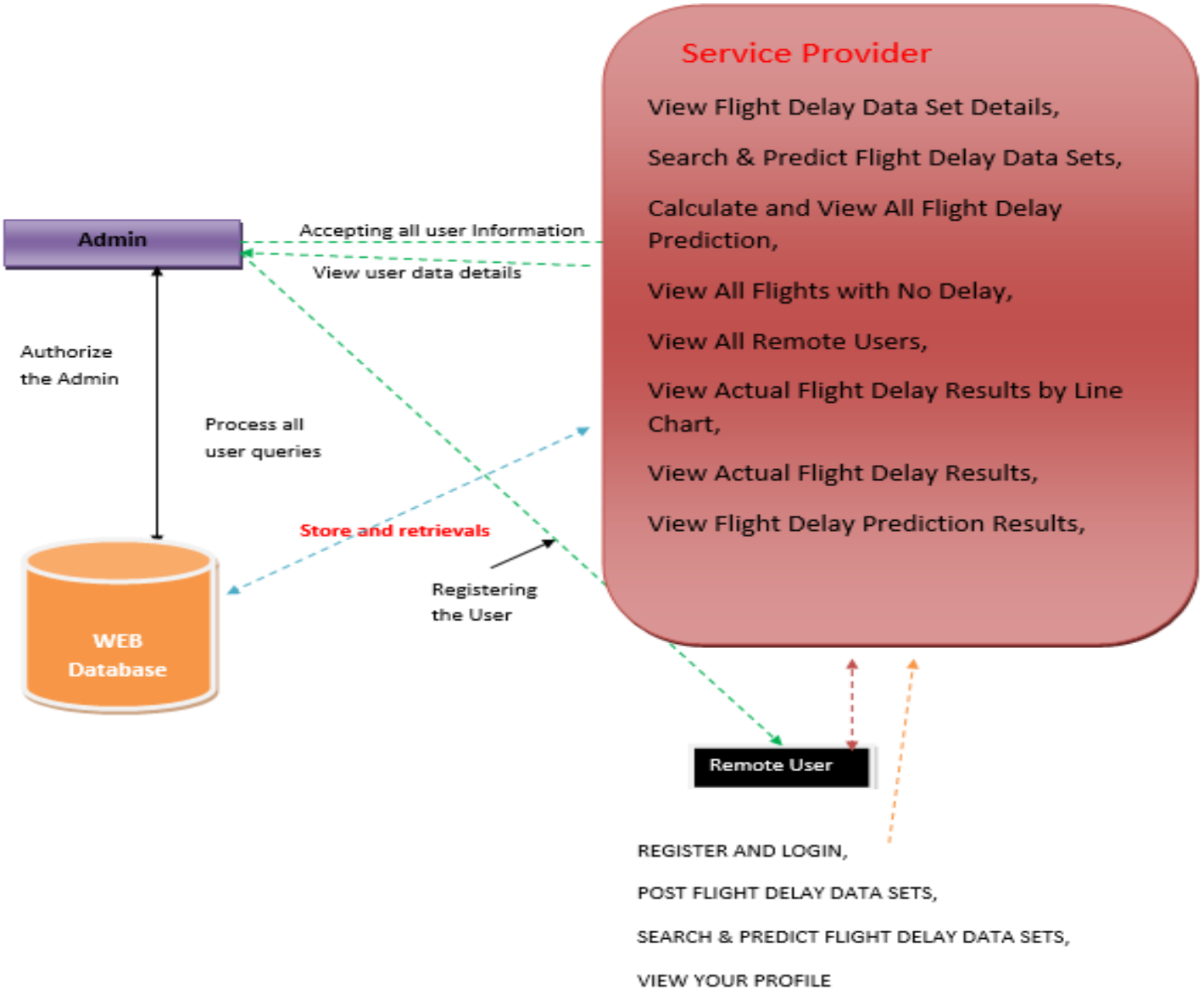


Fig.b. Remote User Flow-Chart

5. ARCHITECTURE DIAGRAM



6. SYSTEM DESIGN AND DEVELOPMENT

6.1 Input Design

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations. This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error is in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases. Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

6.2 Output Design

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client.

User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rest with the administrator only. The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

7. SYSTEM REQUIREMENTS

7.1 Hardware Requirements

- **Processor** - Pentium –IV
- **RAM** - 4 GB (min)
- **Hard Disk** - 20 GB
- **Key Board** - Standard Windows Keyboard
- **Mouse** - Two or Three Button Mouse
- **Monitor** - SVGA

7.2 Software Requirements

- **Operating system** : Windows 7 Ultimate.
- **Coding Language** : Python.
- **Front-End** : Python.
- **Back-End** : Django-ORM
- **Designing** : Html, css, javascript.
- **Data Base** : MySQL (WAMP Server).

8. DATAFLOW DIAGRAM

The flow of data of a system or a process is represented by DFD. **DFD** is the abbreviation for **Data Flow Diagram**. It also gives insight into the inputs and outputs of each entity and the process itself. DFD does not have control flow and no loops or decision rules are present. Specific operations depending on the type of data can be explained by a flowchart. It is a graphical tool, useful for communicating with users ,managers and other personnel. it is useful for analyzing existing as well as proposed system.

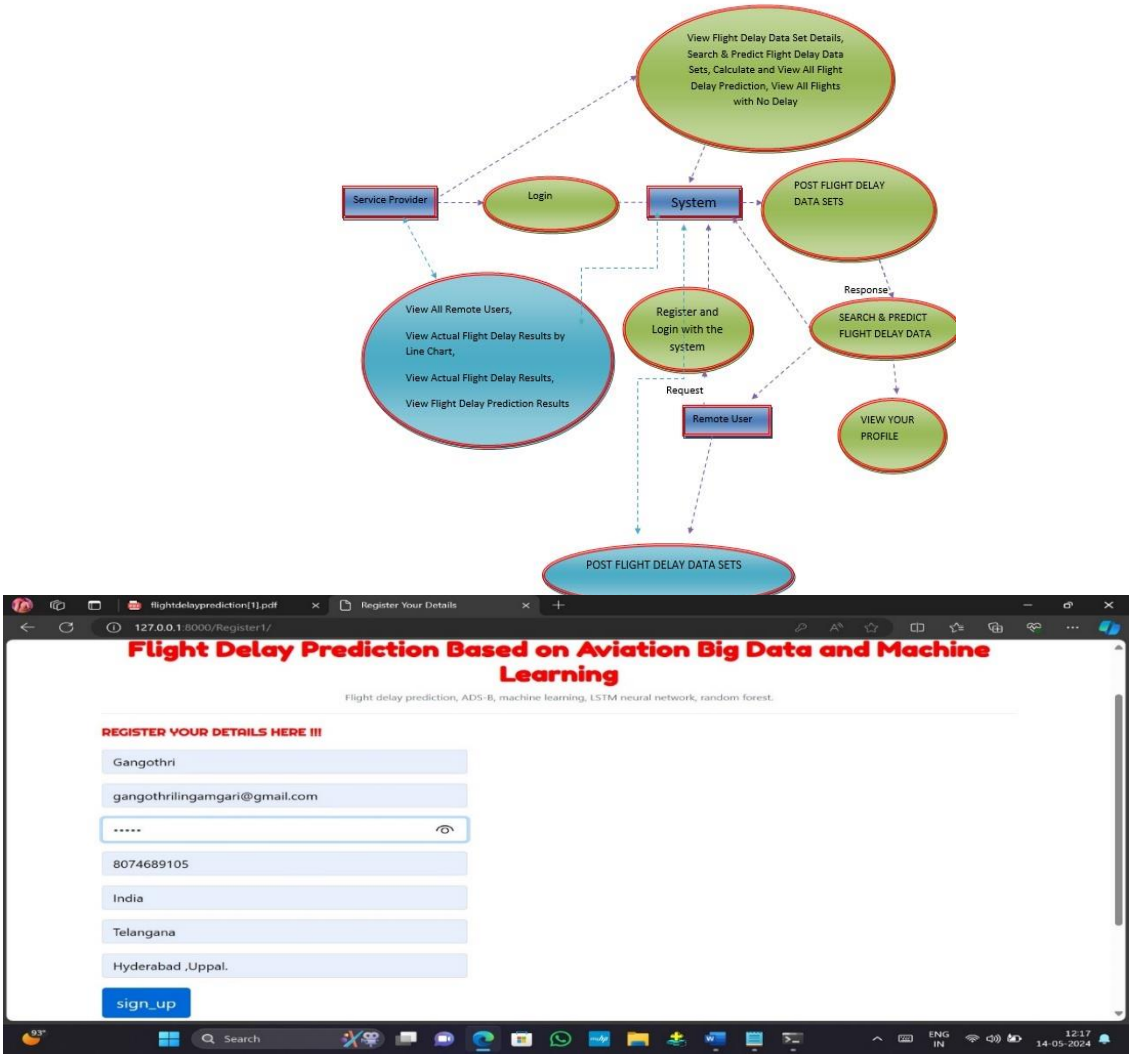
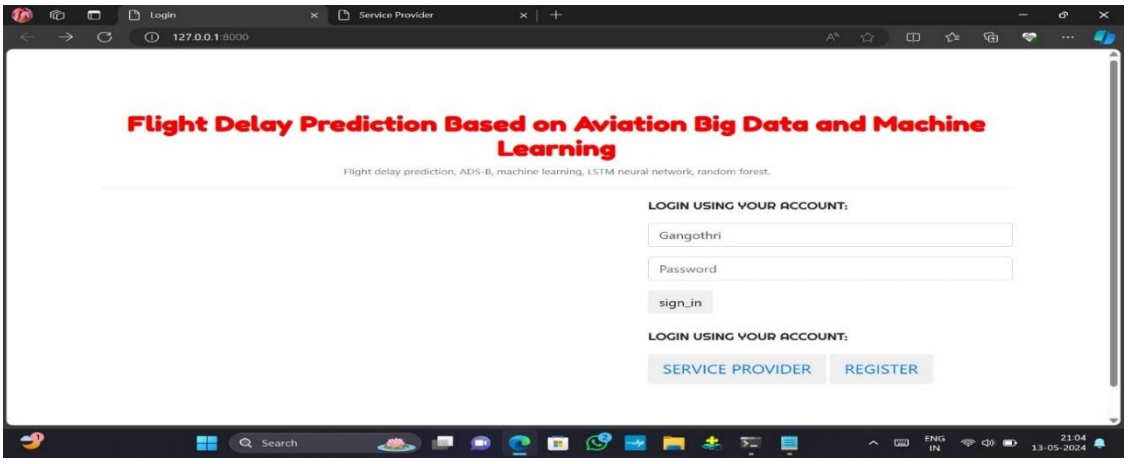
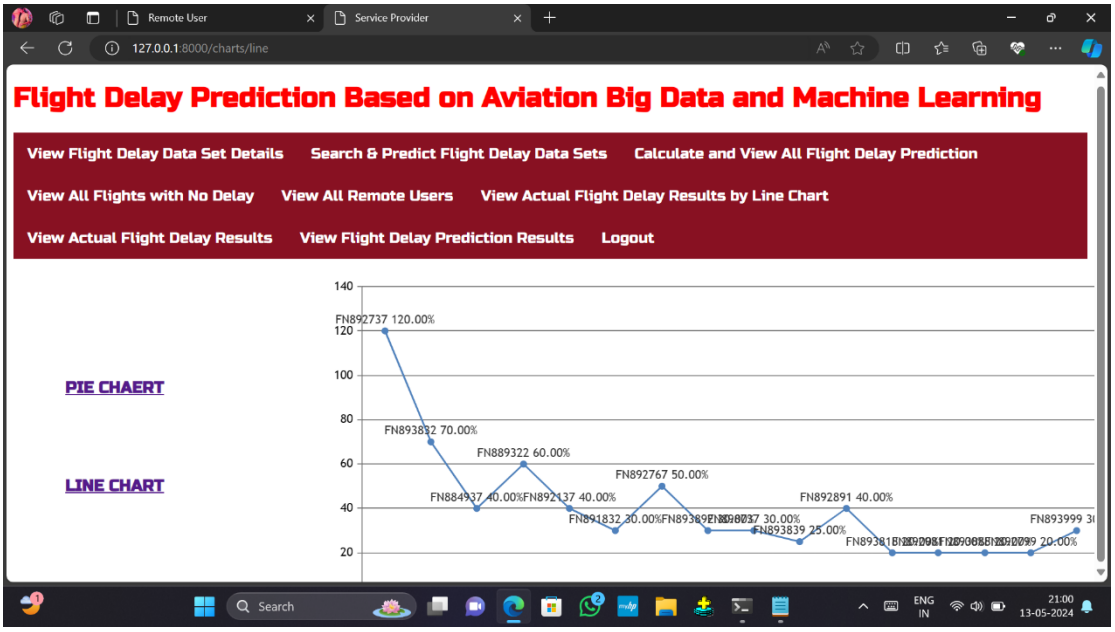
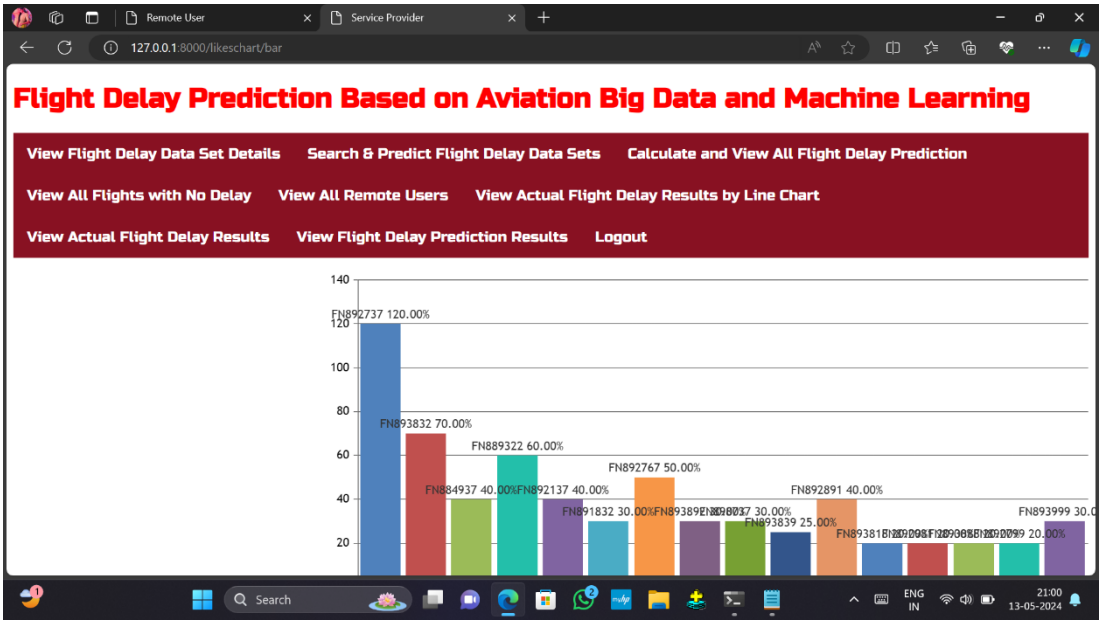


Fig 8 Data Flow Diagram

9. RESULT







10. CONCLUSION

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.

11. REFERENCES

- 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.
- 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 Mult literation,” IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.
- 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. Secure. Technol., pp. 1–5, IEEE, 2018.
- 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.
- 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.
- 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.
- 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.
- 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.
- ANAC, “The Brazilian National Civil Aviation Agency.” anac.gov, 2017. [online] Available:http://www.anac.gov.br/.