Volume 12, No.6, June 2025

Journal of Global Research in Mathematical Archives

ISSN 2320 – 5822

RESEARCH PAPER

Available online at http://www.jgrma.com

A STUDY ON PREDICTIVE MODELING OF FLIGHT DELAYS USING SUPERVISED LEARNING ALGORITHMS

Jyoti Tiwari[1], Rakhi Soni[2] , Sandeep Vishwakarma[3]

[1] Assistant professor, C.S.E DEPT, Shri Rama Krishna College of Engineering Science and Management, Satna, India
[2] Assistant professor, C.S.E DEPT, Shri Rama Krishna College of Engineering Science and Management, Satna, India
[3] Assistant professor, C.S.E DEPT, Shri Rama Krishna College of Engineering Science and Management, Satna, India
jyotitiwari0775@gmail.com[1], rakhisoni017@gmail.com[2], sandeep_ku45@yahoo.co.in[3]

Abstract: Air travel has been very successful since it is the fastest form of transportation, which has given people a lot of confidence in it over the years. Nevertheless, airlines have had to adjust to the issue of aircraft arrival delays, which arise from the fact that runway availability and airspace organization play significant roles in determining the time it takes for a flight to reach its destination. Accurately forecasting flight delays is crucial to the aviation industry's efficiency. In the center of recent research are machine learning strategies for forecasting aircraft delays. Algorithms for supervised machine learning have found widespread use in several areas of ML, including pattern recognition, data mining, and machine translation. Prediction methods in the past have often been limited to a single compass bearing or airport. Unpredictability is a major factor in flight planning, making it one of the most challenging conditions in business. Airports in Lithuania have had their flight time variance examined for this paper. SMOTE is employed to check the statistical reliability of the dataset. Have used the FCMIM method for feature selection. Predicting new delays in combat time using a supervised ML model is now possible. XG-Boost, LightGBM, and AdaBoost, three tree-boosting algorithms, were used in the study.

Keywords: Flight Delays, Supervised Machine learning, Data analysis, Classification algorithms.

1 INTRODUCTION

Any airline has financial, coordination, or technological issues when flight timings deviate from planned times. Travellers may also have trouble or be inconvenienced by timetable changes. Primary and secondary causes account for the majority of battle delays. A complex delay chain spanning all airports may arise from a single incoming aircraft being delayed, which might have a cascading impact on subsequent flights in the line. There has been a lot of recent research that try to predict reactive changes. As an example of a successful strategy, a multi-agent-based approach improves the classification quality to 80.7% [1]. The main type is mostly shaped by the environment. In the previous year, every single flight delay could be traced back to some kind of difficulty at an airport, whether it was bad weather, a lack of available parking, a malfunctioning aircraft, or difficulties with air traffic control, management, or restriction. The departing flight delays have been investigated [2]Using geographical analysis. In light of the fact that weather plays a significant role in various contexts, methods for weather-induced prediction are provided in [3]. However, the airport's size and volume of passengers also matter greatly. Delays in arrival times are often of the reactive kind due to the airport's low daily aircraft capacity. Any delay analysis can make reference to a universal system of delay codes created by the International Air Transport Association [4]. The codes may be used as a class value in different data mining approaches when the effects of these components are explored [5]. There are several causes of flight delays; however, fundamental issues may be avoided by optimizing the airport's physical layout [6]. Finding the best machine learning classification approach for adjusting delay analysis findings for secondary airports is the main objective of the work.

A promising use of ML approaches has been the effective prediction of flight delays by the identification of trends and patterns in large amounts of historical data. These strategies may improve the convenience and comfort of customers while also increasing the productivity and profitability of airlines. Several researchers have looked at the possibility of using ML methods to forecast aircraft delays. However, the majority of these studies have either employed a limited number of approaches or have only taken into account a tiny fraction of the data. By using state-of-the-art ML techniques on a large collection of aircraft data, this study aims to bridge that gap by accurately predicting flight delays. This study's findings could help airlines and customers make better decisions when it comes to using ML methods to predict flight delays. Moreover, this study may set the groundwork for further investigation in this area, leading to the creation of more precise and efficient flight delay prediction models.

1.1 Airports/ Flight Delay Analysis

Air travel is crucial to the functioning of the transportation system and makes important economic contributions. Airports are well-known for their potential to spur economic growth by encouraging more commerce in the areas immediately around them. The demand for air travel grew by 6.3% in 2016 compared to 2015, according to the International Air Transport Association. Weather, maintenance issues, ripple effects from earlier delays, traffic, and other factors are only some of the numerous reasons

why an airplane could be late. When an airplane is late, passengers can't go where they need to be when they need to get there, which may cause a lot of stress. As well as causing financial hardship, this might also aggravate the traveller [7].

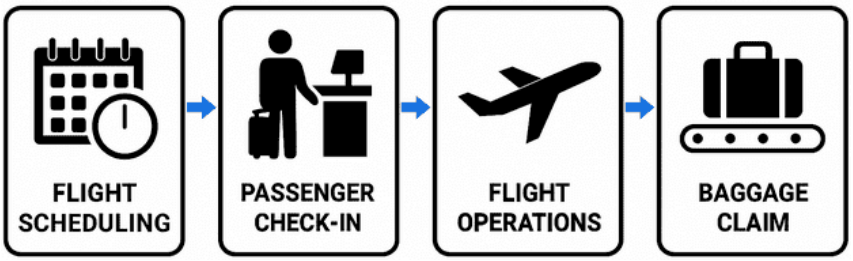

Figure 1: A typical process of Air Transportation System [8]

1.1.1 Problem

In domain taxonomy, the problem statement is the central characteristic. As was discussed in Section 1.2.2, the flight delay prediction issue has three main aspects: delay propagation, root delay, and cancellation. Authors use one of these paths to create their models based on the focus of their study.

- **Root delay and cancellation:** These root delays have an adverse effect on transportation network performance because they increase the likelihood that a new delay will occur. In order to determine the causes of and potential solutions to root delay, researchers use prediction models. This category contains models that aim to effectively predict the duration of a delay, as well as its likelihood and severity, for a given aircraft, airline, and airport.
- **Delay propagation:** The main goal of studying delay propagation is to figure out how a delay in one part of the transportation system affects other parts, such as aircraft and airports. The problem has arisen when the delay of one aircraft affects other planes operated by the same airline. The ability of carriers to recover from delay propagation must be quantified under these conditions. Furthermore, a delay may continue to snowball when essential resources or retentions are scheduled at other airports.

1.2 Machine Learning

Examining the accuracy with which ML algorithms can predict flight delays along with determining the factors that have the greatest effect on these delays is the goal of this study. The emphasis of this study will be on examining large datasets and sophisticated machine-learning methods for forecasting aircraft delays.

The field of machine learning is rapidly expanding, allowing computers to automatically acquire new skills by analysing large amounts of data. The phrase "machine learning" is commonly used to describe the application of techniques to generate mathematical representations and make predictions about the future based on past data and present conditions. Numerous modern-day uses for this technique include email filtering, recommender systems, Facebook auto-tagging, image and speech recognition, etc.

ML is an AI subfield concerned with teaching computers to acquire new skills and improve existing ones by analysing data and gaining insight from their own experiences. The phrase "machine learning" wasn't coined until 1959, when Arthur Samuel first used it. Simply put, ML is the ability of a computer system to acquire new skills and knowledge via exposure to data and examples rather than being explicitly programmed.

To anticipate the outcomes of novel inputs, an ML system constructs prediction models from existing data. More information means a better model can be built, which means more reliable predictions of the final result.

To forecast the outcome of a difficult issue without having to write specific code, we may instead input the problem's data into generic algorithms. These algorithms then construct the logic based on the data and provide a predicted result. The development of machine learning has caused us to reevaluate the situation. The next diagram illustrates the operation of a Machine Learning algorithm [9]:

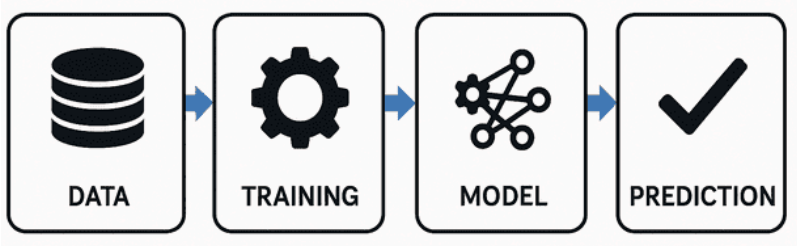

Figure 2: The process of Machine Learning is shown in a block diagram.

1.2.1 Features of Machine Learning

Here are some specifics of ML:

- Data is the currency of ML, and it is used to identify patterns within a particular dataset.
- It can self-improve and learn from its past mistakes.
- It is an information-based system.
- Machine learning, like data mining, is concerned with massive datasets.

1.2.2 Classification of Machine Learning

There are two main categories that may be used to describe machine learning.

1. **Supervised Learning**

It is a kind of ML in which the system is taught to anticipate outcomes by being shown those outcomes in the form of labelled examples. After the model has been trained and processed, we run a test on it to ensure that it can reliably predict future results using sample data. In order to understand the datasets and learn from them, the system builds a model from the labels.

Supervised learning aims to establish a connection between input and output data. In the same way that a student learns under a teacher's watchful eye, supervised learning relies on human oversight. To illustrate supervised learning, consider spam filtering. There are two distinct types of supervised learning:

- **Classification:** An example of a categorization issue would be a test where the possible answers are "red," "blue," "illness," and "no disease."
- **Regression:** For an issue to be classified as regression, the value of the outcome variable must be a real number.

2. **Unsupervised Learning**

With unsupervised learning, a computer is able to learn on its own. The computer is given a dataset that has not been labelled, classed, or categories for training purposes, and the algorithm is expected to make decisions without any human intervention. For unsupervised learning to be successful, the input data must be transformed into either novel features or a collection of objects sharing commonalities in order to be used. There is no expected outcome in uncontrolled learning. The computer analyses the massive dataset in search of actionable insights.

2 TAXONOMY OF THE FLIGHT DELAY PREDICTION

In order to figure out when an aircraft would be late, there are a few different approaches: (i) delay propagation, and (ii) root delay and cancellation. The study of the transmission of delays from one location to another in a transportation system is known as delay propagation. However, because additional issues can subsequently surface, it is equally essential to foresee such delays and comprehend their causes. In this study, we refer to this kind of situation as a root delay issue. Last but not least, there are circumstances in which delays might result in cancellations, resulting in itinerary changes for both airlines and passengers. Cancellation analysis therefore seeks to identify the factors that contribute to cancellations. In addition, it delves into how airlines choose on which flights to cancel [10].

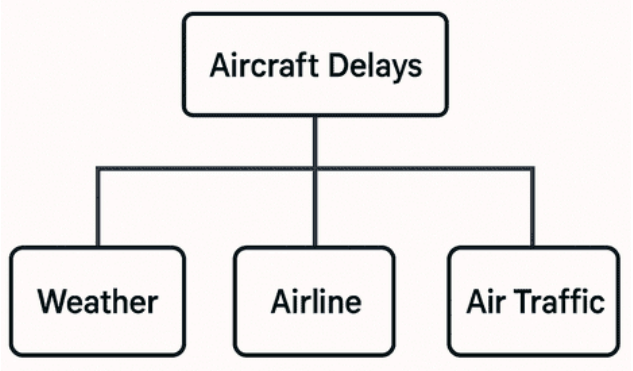

Figure 3: Predicting aircraft delays: a taxonomy[10]

The most pressing issues in predicting flight delays are cataloged in a taxonomy. Scopes, models, and approaches to the flight delay prediction issue are covered. Aspects of the flying domain, such as the issue and the scope, are taken into account, as are Data Science features, such as data and methodologies. The full taxonomy is shown in Figure 3, and the following subsections detail its many parts [11].

2.1 Delay propagation

The main goal of studying delay propagation is to figure out how a delay in one part of the transportation system affects other parts, such as aircraft and airports. One such outcome occurs when one flight's delay has a domino effect on other flights operated by the same airline. In this case, it is crucial to evaluate the carriers' resilience in the face of delay propagation. In addition, if vital resources or retentions need to be scheduled at other airports, the delay might snowball.

2.2 Scope

Airports, airlines, final destination airspace, or all three might be impacted by delays caused by a variety of factors. For the sake of analysis, it is possible to imagine a simplified system in which just one of these actors is examined. It is important to remember that any of the issues listed in Section 3.1 may be paired with any scope of application.

2.3 Data

There are three essential data-related questions, namely: Where can I obtain flight data? Which qualities are most important to look for? Can we improve outcomes by manipulating individual data points? These questions may be addressed by breaking down the data issue into 3 categories: (i) data sources, (ii) dimensions, and (iii) data management.

2.3.1 Data Sources

Airline, airport, and ensemble data are the most common types of air transport datasets. Due to the sensitive nature of their data, airlines and airports often only allow their partners access to their databases. Carriers, airports, and other data from governments, regulators, and service providers may all be part of an ensemble dataset. Dataset types are broken out geographically in Table 1. It lists the total number of works published as well as the three most-cited works in each field. Access to government databases is often available to the public; however, the level of detail may vary. It has been brought to light that the DOT's records show [12]primarily through the FAA [13] And the databases maintained by the Bureau of Transportation Statistics are often queried for flight details [14] The database is provided by a European umbrella government agency. Research on aircraft delays makes heavy use of this dataset as well.[15].

Table 1: Number of verifiable regional air transport data sources

Region	Ensemble	Airline	Airport
Asia	2 [89, 111]	1 [104]	1 [121]
Brazil	2 [110, 5]	0	0
Europe	7 [30, 29, 81]	2 [109, 58]	7 [103, 27, 96]
US	11 [90, 112, 128]	7 [78, 3, 4]	16 [54, 11, 53]

2.3.2 Dimensions

Figure 3 depicts the data model we developed after considering the most important publicly available datasets and the publications we read. They simplify the most important parameters used by delay prediction models. Depending on the goals of the study, researchers may also take into account other factors beyond the departure and arrival timings.

2.3.3 Data Management

As more and more information is being stored in databases, the need for effective data management methods to facilitate quick and easy query processing has grown. Information managers must plan for data integration from several sources, the correction of data discrepancies, and the introduction of new data formats as part of their work. For this reason, it could be helpful to create a data warehouse using OLAP and data management strategies. Multiple sources of information may be combined, as explained in Section 1.2.2.3. That's why it's usual practice to utilise data warehouses in conjunction with Extract, Transform, and Load (ETL) processes to integrate data from several silos [16]. Preprocessing methods from the realm of data management may be used on any number of flight delay prediction datasets. Methods such as "de-noising," "feature selection," "data processing," and "clustering" fall under this category. The elimination of outliers is a crucial part of the data-cleaning process. If you're just interested in how things usually work, extreme situations may produce outliers that are boring [17]. Selecting features involves finding characteristics with low correlation. Model over-fitting or a drop in prediction accuracy might result from the inclusion of

correlated or irrelevant features [15]. These preprocessing steps are crucial because the quality of the prediction models that may be derived from the input data depends on how well it is preprocessed.

3 LITERATURE REVIEW

An improvement in aviation efficiency can only be achieved via better flight delay forecasting. Sreenivasulu et al., (2023). There is now a lot of research going on into the use of ML for the prediction of flight delays. Unfortunately, the majority of historical forecasting methodologies are route or airport specific. Flight planning is plagued with unpredictability, making it one of the most demanding situations in business. Such a circumstance occurs when flights are delayed for a variety of causes, which may be very expensive for the airline, airport, and passengers. Airport communications, luggage handling, mechanical equipment, inclement weather, local and national holidays, airline policies, and the buildup of disruptions from previous flights are all potential causes of flight delays. There is a need to evaluate many ML-focused models for flight delay forecasting since there is a wide variety of potential causes. Data sources like as weather, airline, and airport terminal information, together with ADS-B (Automatic Dependent Surveillance-Broadcast) communications, might be combined and preprocessed to generate a dataset for the suggested method. There will be a regression task and classification difficulties in the next prediction competitions. If airlines want to increase consumer satisfaction and bottom-line profits, they need reliable flight delay prediction systems. They used big data technology, particularly Hadoop, to improve our ability to forecast flight delays. The issue originates from the system's dependence on delayed aircraft data. After that, the data is subjected to regressions. The system takes a number of factors into account. A few of the algorithms used by this system include RF, KNN, LR, logistic regressions, and SVM. The suggested model is built on a random forest, which increases prediction accuracy and prevents over-fitting [19].

In this study, Salam et al., (2023), The flight's tardiness is the primary issue. As the airline industry has grown over the last 20 years, air travel has risen significantly. As a result, the aviation industry suffers significant financial losses and negative environmental effects. Therefore, it is imperative that aircraft not be delayed or cancelled. The present research forecasts aircraft delays using LR, RFR, Logistic Regression, DT, and Sentiment Analysis. The findings will then be used to propose the optimal model. Initial attempts often only cover a single airport or route. This study examines several airlines and possible causes of flight delays [20].

This article compares and contrasts two regression approaches, RIDGE regression and LASSO regression. Evangeline, Joy and Rajan, (2023). Accurate delay forecasting is critical for scheduled airlines' operations and customer satisfaction. Flight delays are unavoidable, yet they may cost or earn airlines a lot of money depending on the conditions. This work aims to better comprehend the spectrum of potential issues associated with aircraft delays by analyzing and comparing two ML approaches in the area of defined extended flight delay time series forecasting. ADS-B signals are collected, deciphered, and combined with other forms of data such as airport details, weather, and flight schedules to provide a dataset for the suggested approach. A regression approach is used to integrate many forecasting jobs in order to meet the given prediction difficulties. After putting the suggested prediction model through its paces, we compared the outcomes to those of more conventional approaches. Results using RIDGE regression had a RMSE of 0.3% while those from LASSO regression were 99.7% accurate. Absolute mean error (0.2%) and squared mean error (0.1%) respectively [21].

The current and the existing, Reddy et al., (2023), flight delays due to traffic congestion pose a number of problems for airline management, including financial losses, negative effects on the environment, worse passenger service quality, higher gasoline and petrol usage, and more. Predictive analytics and ML have the potential to reduce airline flight delays [22].

This research, Rajesh and Srikanth, (2023), emphasizes studies that investigate the potential of ML methods for airline delay prediction. The main goal of this research is to find out which machine learning methods work best for predicting flight delays and which factors cause them. The information utilized for the study includes airport wait times, flight itineraries, weather predictions, and other features. This study will begin with a literature review of previous work on the subject of using ML to predict flight delays. Results from several ML approaches are examined and contrasted here, including DT, RF, SVM, and neural networks. Research shows that ML algorithms (e.g., random forests and decision trees) can reliably predict when a flight could be delayed. The study found that inclement weather, issues unique to certain airlines, and overcrowding at airports were the top three reasons for flight delays. The aviation sector stands to gain a lot from this study's findings, since better flight delay forecasting has the potential to enhance airport and airline management as well as customer pleasure. The study's results point to the potential of ML techniques to improve the precision and practicality of aircraft delay predictions, which might ultimately lead to a more secure and effective aviation system [23].

4 CONCLUSION AND FUTURE WORK

A multitude of forecast methodologies, a deluge of flight data, and a complicated air transportation system all combined to make it difficult to build reliable models for predicting flight delays. An examination of flight delays takes into account both the actual and anticipated arrival and departure times. Using ML methods, this research developed a model to predict airport-wide flight delays. Airline flight cancellations and delays in departure and arrival times are only two examples of the many types of use cases that are considered. Other than passengers, everyone involved in the air transportation system relies on delay prediction to make judgments. This study contributes by providing a Data Science perspective on the analysis of these models. Models of flight

delays have been developed for use in this context throughout the years. Using a taxonomy approach, we classified models based on their characteristics. Using extensive data and state-of-the-art methods, this study proved that machine learning can accurately anticipate flight delays. The results of the study showed that aircraft delays may be correctly predicted by machine learning, giving airlines useful information for optimizing their operations and reducing the negative effects of delays on passengers and the economy. This study makes a significant contribution to the study of airline operations and machine learning by providing empirical evidence of the latter's capacity to enhance the former. In order to boost their operations and customer satisfaction, we suggest that airlines use these cutting-edge methods and concepts.

REFERENCES

[1] Y. Guleria, Q. Cai, S. Alam, and L. Li, "A Multi-Agent Approach for Reactionary Delay Prediction of Flights," *IEEE Access*, vol. 7, pp. 181565–181579, 2019, doi: 10.1109/ACCESS.2019.2957874.

[2] S. Cheng, Y. Zhang, S. Hao, R. Liu, X. Luo, and Q. Luo, "Study of Flight Departure Delay and Causal Factor Using Spatial Analysis," *J. Adv. Transp.*, pp. 1–11, Jun. 2019, doi: 10.1155/2019/3525912.

[3] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, IEEE, Sep. 2016, pp. 1–6. doi: 10.1109/DASC.2016.7777956.

[4] H. Träff, L. Hagander, and M. Salö, "Association of transport time with adverse outcome in paediatric trauma," *BJS Open*, vol. 5, no. 3, pp. 1–7, May 2021, doi: 10.1093/bjsopen/zrab036.

[5] M. Zámková, M. Prokop, and R. Stolín, "Factors Influencing Flight Delays of a European Airline," *Acta Univ. Agric. Silvic. Mendelianae Brun.*, vol. 65, no. 5, pp. 1799–1807, Oct. 2017, doi: 10.11118/actaun201765051799.

[6] E. P. Gilbo, "Airport capacity: representation, estimation, optimization," *IEEE Trans. Control Syst. Technol.*, vol. 1, no. 3, pp. 144–154, 1993, doi: 10.1109/87.251882.

[7] S. A. Mohiddin, G. Komali Priya, D. K. Sai, P. L. Sree, and P. Sriram, "Flight Delay Analysis Using Machine Learning," *Dogo Rangsang Res. J.*, no. 01, pp. 841–844, 2022.

[8] A. HOSSAIN, "Application of Interpretable Machine Learning in Flight Delay Detection," p. 6, 2021.

[9] *et al.*, "Predicting Flight Delays With Error Calculation Using Machine Learning," *YMER Digit.*, vol. 21, no. 05, pp. 364–369, 2022, doi: 10.37896/ymer21.05/40.

[10] L. Carvalho *et al.*, "On the relevance of data science for flight delay research: a systematic review," *Transp. Rev.*, vol. 41, no. 4, pp. 499–528, Jul. 2021, doi: 10.1080/01441647.2020.1861123.

[11] T. Wang, Y. Zheng, and H. Xu, "A Review of Flight Delay Prediction Methods," in *2022 2nd International Conference on Big Data Engineering and Education (BDEE)*, IEEE, Aug. 2022, pp. 135–141. doi: 10.1109/BDEE55929.2022.00029.

[12] K. McDonough, "United States Department of Transportation," *Choice Rev. Online*, vol. 50, no. 06, pp. 50-3233-50–3233, Feb. 2013, doi: 10.5860/CHOICE.50-3233.

[13] CQ Press, "Federal Aviation Administration," in *Federal Regulatory Guide*, 2455 Teller Road, Thousand Oaks California 91320: CQ Press, 2020, pp. 906–912. doi: 10.4135/9781544377230.n127.

[14] J. Reason, E. Hollnagel, and J. Paries, "European Organisation For The Safety of Air Navigation Eurocontrol Experimental Centre Revisiting The Swiss Cheese Model of Accidents," *Eur. Organ. Saf. AIR Navig.*, 2006.

[15] A. J. Reynolds-Feighan and K. J. Button, "An assessment of the capacity and congestion levels at European airports," *J. Air Transp. Manag.*, vol. 5, no. 3, pp. 113–134, Jul. 1999, doi: 10.1016/S0969-6997(99)00006-X.

[16] R. Yao, W. Jiandong, and D. Jianli, "RIA-based visualization platform of flight delay intelligent prediction," in *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, IEEE, Aug. 2009, pp. 94–97. doi: 10.1109/CCCM.2009.5267976.

[17] Y. Tu, M. O. Ball, and W. S. Jank, "Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern," *J. Am. Stat. Assoc.*, vol. 103, no. 481, pp. 112–125, Mar. 2008, doi: 10.1198/016214507000000257.

[18] J.-T. Wong and S.-C. Tsai, "A survival model for flight delay propagation," *J. Air Transp. Manag.*, vol. 23, pp. 5–11, Aug. 2012, doi: 10.1016/j.jairtraman.2012.01.016.

[19] K. Sreenivasulu, B. Sowjanya, V. R. Motupalli, S. H. Yadav, K. K. Baseer, and M. J. Pasha, "Prediction of Flight Delay through Intelligent Algorithms and Big Data Technology," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, May 2023, pp. 1074–1080. doi: 10.1109/ICAAIC56838.2023.10141246.

[20] S. Salam, S. M. Sohail, K. D. Reddy, V. Darvesh, T. Y. Reddy, and O. G. Rao, "Predicting Flight Delay Based on Sentimental Analysis: Machine Learning," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, Jan. 2023, pp. 1–4. doi: 10.1109/ICCCI56745.2023.10128271.

[21] A. Evangeline, R. C. Joy, and A. A. Rajan, "Flight Delay Prediction Using Different Regression Algorithms in Machine Learning," in *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, IEEE, Mar. 2023, pp. 262–266. doi: 10.1109/ICSPC57692.2023.10125675.

[22] R. T. Reddy, P. Basa Pati, K. Deepa, and S. T. Sangeetha, "Flight Delay Prediction Using Machine Learning," in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2023, pp. 1–5. doi: 10.1109/I2CT57861.2023.10126220.

[23] K. Rajesh and V. Srikanth, "Predicting Flight Delays Using Machine Learning : An Analysis of Comprehensive Data and Advanced Techniques," vol. 12, no. 4, pp. 82–85, 2023, doi: 10.17148/IJARCCE.2023.12416.

