Classifying Airport Delays and Recommending Improvements Using Machine Learning and Passenger Feedback Analysis

1st Prasanth Ramaswamy
Dept. of Computer Science and Information Systems
University of Limerick
Limerick, Ireland
24134252@studentmail.ul.ie

2nd Yogesh Sivakumar
Dept. of Accounting and Finance
University of Limerick
Limerick, Ireland
24150029@studentmail.ul.ie

***Abstract*—Aviation is currently a $762.8 billion industry in 2023 and has a critical role in connecting the world of ideas, people, and goods. However, flight delay which occurs between 15-20% of the international flights every year is still a major issue, which results in monetary loss and passengers' dissatisfaction. In the U.S., delays cost airline users approximately $33 billion in 2019 (FAA, 2020). Airports such as JFK (New York) and DFW (Dallas) have a delay rate of more than 40 percent during the busiest travel months and bad weather. The findings of this work use XGBoost, SVM, and Random Forest in categorizing delays as low, moderate, and high. A novel Feature-Interaction Enhanced XGBoost and Deep Learning model enhances performance, by combining both Feed Forward Neural Network (FNN) and XGBoost. Moreover, the NLP approach is applied to more than 6,000 passengers' reviews to uncover various organizational issues, including queue management and cleanliness. Unlike many existing delay studies, this research not only extracts the delay patterns from the structured and unstructured data but also offers specific recommendations relevant to each airport to increase efficiency and passenger satisfaction. This framework is proposed to be generic and applicable to airports globally to provide suggestions for the stakeholders in airport industry.**

***Index Terms*—Machine Learning Techniques, Airport Delay, Supervised Learning, Unsupervised Learning, Delay Prediction, Sentiment Analysis, CRISP-DM Framework**

I. Introduction

Nowadays, air travel is one of the essential ways to connect people, ideas, and goods, and it has a significant impact on a country's economy. In the United States, civil aviation contributed 4% to the GDP for the year 2024 [1]. However, the aviation industry has not escaped challenges, and the most frequent challenge is flight delays that lead to losses and unhappy customers [2]. For the year 2024, in the United States, total flight delays exceeding 15 minutes account for twelve to thirteen percent, and some of the causes include meteorological and operational factors [3]. Delays result in poor aircraft utilization, which ultimately impacts the cost. A 10% reduction in delays would result in gains of $1.50–2.50 per passenger in the US [4]. Hence, predicting delays and working on the suggestions will be beneficial both for the passengers and the airlines.

Existing flight delay prediction models use traditional statistical techniques and structured data formats to predict delays, but they cannot detect complex patterns and they overlook unstructured data formats. Unstructured data sources like text data can provide valuable real-time insights that help generate recommendations for addressing airport-specific delays, which can be used as a solution to improve delay aspects. Moreover, most delay prediction models are of a generalized type; they are less effective in local operational contexts as they fail to consider airport-specific inefficiencies [5] [6] [7].

The objective of this research is to develop a comprehensive model to predict airport delays using delay metrics, classify them into high delay, moderate delay, and low delay, and integrate passenger feedback data to provide suggestions for airport-specific improvements. To overcome the challenges encountered in existing research, this study analyses a combination of structured and unstructured data sources. An extensive time range spanning from 2013 to 2023 is utilized by this research to fully inspect flight delay patterns across multiple years.

The research question primarily focuses on how to classify airports into delay categories and provide recommendations to mitigate operational inefficiencies, particularly for airports experiencing high delays.

To address the research question, this study employs quantitative data, holding the delay metrics and weather data used with hybrid models that combine multiple algorithms. i.e., the developed model merges XGBoost with deep learning methods, where XGBoost handles structured data processing and deep learning for feature extraction and the qualitative data, which is the passenger feedback data that undergoes natural language processing for text analysis and sentiment evaluation.

The key contributions of this work are as follows:

- Developed a hybrid model that merged XGBoost with deep learning algorithms for delay classification purposes.
- Analysis of unstructured airport data through NLP brings forth specific insights at an airport level, providing a

detailed perception of operational improvements for particular airports.

The remainder of this paper is organized as follows: Section 2 provides a detailed review of related work, discussing existing approaches. Sections 3 and 4 discuss data mining methodology and the experimental setup and results of the model. Finally, Section 5 concludes the paper by summarizing the findings.

II. Related Work

Delay on Airlines is an important issue as it not only impacts the travellers and airline company but also has an indirect impact on the economy. Flight delays not only increase fares but also decrease people's willingness to pay for air travel [8]. Consequently, the management of such airlines and airports needs systematization of flight operations through machine learning techniques to perform the flight delay analysis and forecast delays to undertake the necessary precaution. A number of works [2] [5] [9] [10] [6] have been conducted in this regard incorporating various machine learning techniques, yet there are difficulties stemming from restricted dataset range and limited model effectiveness.

The current research lacks broader generalization since it exclusively analyzes delays at particular airports. Research by Tang et al. [2] examined delays at John F. Kennedy Airport alongside Kavitha et al. [10] who studied three major New York airports. Similarly, Stefanovic et al. [11] performed a study of Lithuanian airports. The restricted sample size reduces the ability to transpose findings into various operational settings. Our work resolves this limitation through extensive data collection that contains information from various U.S. domestic airports for a more detailed analysis.

One of the major gap in current studies is the ignorance of unstructured data sources which would supply greater insights. While few studies such as [12] have utilized unstructured passenger feedback data for customer satisfaction analysis, but for airline delay prediction and recommendation, still structured data is widely used. Stefanovic et al. [11] analyzed flight schedule data along with meteorological data without including unstructured passenger feedback information. Sharan et al. [6] alongside Nishant and Vijaylakshmi [13] based their research on structured datasets incorporating weather data and various preprocessing techniques, but still unstructured data hasn't been incorporated. The work by Blessy Trencia Lincy et al. [14] employed cluster methods for discovering flight delay pattern without including texts from passenger complaints. Our research applies both airport-specific parameters alongside weather information to make delay predictions and adds unstructured passenger feedback to create customised airport specific recommendation for a more comprehensive approach.

The performance of models is limited by the duration covered by the available datasets. Lakshmi et al. [15] analyzed flight records from July 2019 through December 2019 which could have excluded seasonal flight delay patterns. The real-time API data used by Kothari et al. [16] creates difficulties when the information sources become unreliable. The accuracy of predictions will improve when analyzing longer time spans along with reliable and dependable sources of data. Our research incorporates data from 2013 to 2023 in order to analyze both long-term patterns as well as seasonal fluctuations which were not studied in previous research.

The selection of features alongside choosing the appropriate complexity level affects the outcome of predictions. The study by Evangeline et al. [5] established that LASSO and Ridge regression did not have sufficient capability to detect intricate relationships between model features. Hatipoğlu and Tosun [17] proved that operational data proved more impactful than meteorological data which underlines the importance of selecting features with precision. Studies by Ashok et al. [18] and Qu et al. [19] demonstrates the effectiveness of a hybrid models which combines multiple deep learning models and they perform better than individual models. Jun Chen and Meng Li [20] shown that choosing relevant features plays an essential role in improving predictive capability. Our approach addresses these gaps by employing a hybrid model that combines deep learning with XGBoost while incorporating a robust feature selection process to refine the most impactful predictors.

To conclude, Current research on flight delay prediction has witnessed major improvements but additional work is still needed to address existing gaps. Predictive models become more effective and applicable because of using unstructured data sources along with longer diverse datasets through advancing feature selection techniques. Our study bridges these gaps by employing a combined approach that integrates structured and unstructured data, a decade-long dataset, and advanced feature selection techniques, contributing to a more comprehensive and accurate prediction framework.

Table 1 shows the techniques employed in existing studies along with the data used and the results. Amoung those techniques, XGBoost and some deep learning techniques has performed well. Selecting important features and using them also has increased the model's prediction capability [20]. But there aren't many research on combining feature selection with deep learning. In our research, we aim to develop a hybrid model combining XGBoost and Deep learning that also uses optimal interaction features to capture combined effects of features, thereby improving the model's predictive power.

Most of the studies discussed here didn't use semi structured textual data for analysis, which could bring in additional insights. Also, some of the studies are scoped to limited number of airports. To bridge this gap, our research incorporates textual data along with the data used by Nishant and Vijaylakshmi [13] as it has a broader and more recent timeframe and it covers all the major airports across the United States.

III. Data Mining Methodology

The study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework to address the research question. This methodology provides a process map and

Article	Dataset	Timeframe of Data	Technique
Tang et al. [2]	Data from JFK airport, USA	Nov 2019 to Dec 2020	Decision Tree SVM Random Forest K-NN
Stefanovic et al. [11]	Data from 3 major Lithuania airports	Oct 2019 to Mar 2020	Gradient boosted trees
Lakshmi et al., [15]	US Domestic flight data + local climatological data from NOAA		XGBoost
Kothari et al. [16]	Bureau of Transportation Statistics (BTS) + US National Oceanic and Atmospheric Administration (NOAA)		Random Forest
Evangeline et al. [5]	Data from Nanjing Lukou Airport, China	Mar 2017 to Feb 2018	LASSO Regression RIDGE Regression
Tijil et al. [9]	Data from Bureau of Transportation Statistics (BTS)	2015	SVM Logistic Regression Random Forest
Sharan et al. [6]	US Flight delay data	2009 to 2018	XGBoost
Swaminathan et al., (2018) [21]	Data from Bureau of Transportation Statistics (BTS)		Decision Tree Logistic Regression
Blessy Trencia Lincy et al., (2022) [14]	Data from Civil Aviation Authority (CAA) of the United Kingdom	2015 to 2020	K-NN
Hatipoğlu and Tosun [17]	Flight delay data + meteorological data	2016 to 2018	XGBoost
Anees and Huang [22]	Data from Bureau of Transportation Statistics (BTS)	2008	Random Forest
Nishant and Vijaylakshmi [13]	Airline Delay cause in USA	2013 to 2023	Dense Neural Network
Ashok et al. [18]	Data from U.S. Ministry of Transport's		Neural Fusion Network (RNN + LSTM)
Qu et al. [19]	Airline On-time Performance Data (AOTP) from the Bureau of Transportation Statistics (RITA) + Quality Controlled Local Climatological Data (QCLCD) provided by the National Climatic Data Center (NCDC)	2016 to 2017	DCNN (Dual-channel Convolutional Neural Network) SE-DenseNet (Squeeze and Excitation-Densely Connected Convolutional Network)
Jun Chen and Meng Li [20]	Data for O'Hare International Airport (ORD), Chicago, USA from Bureau of Transportation Statistics (BTS), Local Climatological Data (LCD) and Aviation System Performance Metrics (ASPM)	2016 July to 2017 June	Multi-label Random Forest and approximated delay propagation

TABLE I
SUMMARY OF TECHNIQUES AND ACCURACY FROM VARIOUS STUDIES ON AIRLINE DELAY PREDICTION.

working plan to accomplish and build findings and solutions in a structured manner using NLP and machine learning.

A. Business Understanding

The main goal of the research is first to categorize flight delays in the context of airports across the United States with respect to different parameters like weather and operations. Furthermore, specific recommendations for each airport are obtained by including passenger feedback to eliminate endemic problems and enhance service delivery.

B. Data Understanding

1) Airline Delay Cause: Following the airline delay cause data outlined by Nishanth and Vijayalakshmi [13], same has been used for this research. Here's a summary of its key elements:

- **Date Information:** year and month
- **Airport Information:** airport_name
- **Flight and Delay Counts:** arr_flights, arr_del15, arr_cancelled, arr_diverted
- **Delay Causes and Duration (in minutes):** carrier_ct, weather_ct, nas_ct, security_ct, late_aircraft_ct, arr_delay, carrier_delay, weather_delay, nas_delay, security_delay, late_aircraft_delay.

2) Weather Data: Key details of the weather data include:

- **Date and Location:** year, month, and airport_name
- **Weather Metrics:** Avg. Temperature, Avg. Dew Point, Avg. Precipitation, Avg. Snow Depth, Avg. Wind Speed, and Avg. Gust Wind.

3) Airport Reviews Data: Following is the summary of key elements of Airport reviews data, which contains airport specific ratings along with the reviews,

- **Date Information:** Date of review, Date of visit.
- **Airport Information:** airport_name.
- **Review Content:** title, content.
- **Trip Details:** trip_verified, experience at the airport, type of traveler.
- **Ratings (on a scale of 1-5):** Queuing Times, Terminal Cleanliness, Terminal Seating, Terminal Signs, Food and Beverages, Airport Shopping, Wifi Connectivity, Airport Staff.

Unlike many other existing studies, where mostly ratings are used for analysis, this data contains both ratings along with reviews which enhances the airport specific recommendations.

C. Data Preparation

1) Flight Delay Dataset Preparation: The flight delay dataset included attributes such as arrival and departure delay

durations, airport names, and operational metrics.

Data Cleaning: Missing values in key delay-related columns were imputed using median values, and irrelevant fields were removed to focus on features relevant to delay analysis.

Normalization: Airport delay-related features were normalized to a [0, 1] range using Min-Max scaling. This approach makes all features uniformly expressive regardless of their original scale, by normalizing values to the range from 0 to 1. Normalization therefore scales down large variations and ensures that those features with higher magnitude do not overwhelm the results.

Feature Engineering: A Weather Delay Score is calculated as the mean of normalized weather-related features, while an Airport Delay Score is calculated as the mean of normalized airport delay-related features. The above two scores found in the first equation is used to derive a Composite Delay Score which is determined by considering both weather and operational contributions to delays in equal measures. The Composite Delay Score is divided into three categories based on percentiles:

- **Low Delay:** Scores less than or equal to 33rd percentile.
- **Moderate Delay:** Scores between the 33rd and 67th percentiles.
- **High Delay:** Scores greater than 67th percentile.

The distribution of the delay categories was normalized using the 33rd and 67th percentiles to avoid extreme values in any single category while maintaining meaningful results. This method provided distinct thresholds that reflected the variability in delay causes. The target variable, *delay_category*, achieved a balanced distribution: Low Delay (33.00%), Moderate Delay (33.99%), and High Delay (33.00%).

2) Weather Data Collection and Transformation: The weather dataset was prepared using web scraping techniques to extract monthly weather metrics from the Wunderground website. This process involved:

Web Scraping Process: The data collection was kept automatic with the help of the Python selenium package which was utilized to scrape the Wunderground website. Selenium controlled a web browser to navigate through the website for each year and month across multiple airports. Key weather attributes were extracted from tables on the webpage, including:

- Average Temperature (°F)
- Dew Point (°F)
- Precipitation (inches)
- Snow Depth (inches)
- Wind Speed (mph)
- Gust Wind Speed (mph)

Unit Conversion: Extracted metrics were standardized to ensure uniformity. Temperatures were converted from Fahrenheit to Celsius. Precipitation and snow depth were converted from inches to millimeters, and wind speeds were converted from miles per hour (mph) to kilometers per hour (km/h).

Handling Missing Data: Missing or incomplete rows were identified during scraping and addressed during preprocessing. For time periods with sparse weather data, mean or median imputation was applied. The final dataset consisted of weather variables adjusted to match flight delay data at the airport and month level.

3) Passenger Reviews Dataset Preparation: **Data Cleaning:** The passenger reviews dataset was a mixture of structured ratings, such as cleanliness, queuing times and free text. Any gaps in the ratings were given a median value instead, while useless reviews with no text or with just random signs were omitted for the sake of data credibility.

Text Preprocessing: Preprocessing of the text data was done by stripping it from punctuations, special characters and eliminating stop words. Reviews were preprocessed by tokenizing them by words, converting all words to lower case, and lemmatizing them to reduce variations in the dataset.

Keyword Extraction and Sentiment Analysis: Terms extracted from negative reviews were extracted using TF-IDF and major areas of concern included staff courtesy, hygiene, and waiting time. The sentiment polarity scores were obtained from TextBlob to identify negative reviews separately to get more insights on the reasons for passengers' dissatisfaction. The processed reviews dataset combined numerical ratings and text analysis with structured evaluation, ensuring a holistic vision of the passengers' experience. This strong data preparation process guaranteed the availability of good quality data for modeling and analysis.

D. Modelling

Different ML algorithms were used to subcategory airport delays: supervised and unsupervised approaches were used to have a complete perspective. Following are the models implemented:

- Support Vector Machines (SVM)
- Random Forest
- Gradient Boosting and XGBoost
- K-Means Clustering
- Feature-Interaction Enhanced XGBoost and Deep Learning

After identifying high-delay airports from all the models, a consolidated list of these airports was prepared. For these high-delay airports, airport-specific recommendations were provided using the airport review dataset. This was done by using Natural Language Processing (NLP) on the passengers' reviews to analyze their findings.

E. Evaluation

The models were assessed based on suitable measures for the supervised and unsupervised methods. For classification models (supervised), accuracy, precision, recall, F1-score were applied. For the clustering model (unsupervised), the Silhouette score was taken into account in the evaluation of the quality of clustering.

F. Deployment

The last stage of the CRISP-DM process was the creation of a synthesis that listed the high-delay airports and specified

the detailed recommendations for each airport improvement. These recommendations were made depending on the findings from the machine learning model and Natural Language Processing on the passengers' reviews. The summary presents tangible measures to tackle delay, based on the operational issues, staff attitude, cleanliness as well as the management of queues as seen from the reviews. The qualitative analysis of these results alongside quantitative analysis of data, makes this step a guide to increasing airport performance and passengers' satisfaction.

IV. Methodology

Under the methodology section, below are the models that were applied along with the performance measures employed to evaluate their effectiveness. It also gives a brief description of the selected models, their roles in the analysis and the evaluation criteria used in the assessment of the results thus getting an overall understanding of the results of the analysis.

A. Models Used

In an endeavour to answer the research question, both classification, also known as supervised learning, and clustering also known as unsupervised learning models was utilized in developing the core of the methodology. SVM, Random Forest, Gradient Boosting, and XG Boost models were used in the identification and classification of delays based on several factors highlighting weather and performance indicators from operations. Secondly, a clustering approach, K-Means Clustering which is an unsupervised model was employed to analyse delay characteristics of airports to categorize them into different groups. Also, a novel approach called Feature-Interaction Enhanced XGBoost and Deep Learning was tried which combines FNN along with XGBoost. Finally, Airport Specific Improvement through NLP Techniques were used. The specifics of data pre-processing for each of these approaches, as well as the methods of feature engineering and the criteria for evaluating their effectiveness, are also considered to better understand their applicability to analyse the issue of airport delays.

Supervised Learning Models:

The classification models exhibited strong predictive performance, with the results summarized below. The research question called for a categorization of flight delays as either Low, Moderate, or High depending on a variety of factors; a list that included weather conditions, among operational parameters which this analysis sought to provide detailed and statistically sound information relevant to the indicated research question.

1) Support Vector Machine (SVM): The accuracy of the Support Vector Machine (SVM) model in the current study was 84.10%. Class-wise performance:

Class	Precision	Recall	F1-Score
High Delay	85%	89%	87%
Moderate Delay	76%	77%	77%
Low Delay	92%	87%	89%

Implications:

- High Delay: It is observed that a high-recall of 89% for High Delay points to SVM's capacity to identify high-delay instances, implying on the ability of this model to accurately reduce delayed instances.
- Moderate Delay: SVM's recall in Moderate Delay(77%) is slightly better than other models, showing that SVM may be better suited for the task of coping with overlapping between two classes.
- Low Delay: The high precision (92%) means that the program is able to give out few false alarm, which makes the results obtained clear.

2) Random Forest: Random Forest model achieved an overall accuracy of 83.86Class-wise performance:

Class	Precision	Recall	F1-Score
High Delay	86%	88%	87%
Moderate Delay	76%	76%	76%
Low Delay	90%	88%	89%

Implications:

- High Delay: The precision and recall show that Random Forest recognizes important factors that contribute to considerable delays and presents useful feature importance to pinpoint causes.
- Moderate Delay: It is indicated that there are some common surrogating elements influencing all the delay types and the feature extraction should be used to explore more details for the separation process.
- Low Delay: High precision increases the confidence of the model for making predictions since high precision and recall is achieved.

3) XGBoost and Gradient Boosting: Using Gradient Boosting, the accuracy was 83.86%, while XGBoost improved the accuracy to 85.08%. Class-wise performance:

Class	Precision	Recall	F1-Score
High Delay	85%	87%	86%
Moderate Delay	76%	76%	76%
Low Delay	91%	88%	89%

Implications:

- High Delay: High recall (87%) means that the chosen model successfully screens most of the high-delay cases, which may create an opportunity for using the findings to confront extensive delays at airports.
- Moderate Delay: Average F-measure of 76% is an expected outcome, which is probably caused by the fact that distinguishing moderate delay from other classes is challenging.
- Low Delay: The high precision (91%) is obtained for Low Delay implying very few false positives in use of the model.

Unsupervised Learning Models

4) K-Means Clustering: In K-Means Clustering, the highest Silhouette Score of 0.343 for k = 2 which indicates the presence of two distinct clusters in the data: one was made of airports with high delay rates and the other was g roup of airports with low delay rates. The score was much lower for higher choice of k hence displaying an implication of lower quality clusters. The outcomes underscore the ability to disaggregate a delay pattern across one or many airports, which forms the preliminary foundation for the identification of airports that are at risk of delays.

Number of Clusters	Average Silhouette Score
2	0.34
3	0.31
4	0.25
5	0.27
6	0.23
7	0.23
8	0.20
9	0.20

TABLE II
SILHOUETTE SCORES FOR DIFFERENT NUMBER OF CLUSTERS

5) Natural Language Processing (NLP) Approach: The specification models were also employed to determine the airport that belong to the High Delay category. These high-delay airports were then ranked with increases delay for further examination to generate recommendation by using different NLP approaches. These high delay airports were identified using the Airport Reviews dataset; the structured feedback included ratings on the aspects of environment, time spent queuing and textual feedback as well.

Sentiment Analysis: TextBlob is used to perform sentiment analysis in order to identify the disposition of passengers especially those with negative experiences. It underlined areas that needed enhancement: staff etiquette, tidiness, and ways customers formed queues.

Keyword Extraction: The used technique of TF-IDF (Term Frequency-Inverse Document Frequency) allows selecting the most salient keywords from negative reviews and place them into specific topics (e.g., customer service, facility maintenance).

Consequently, specific recommendations about specific airports are based on sentiment analysis and keywords extracted. For instance, from the aspects concerning in the reviews such as staff training, facility cleaning, and queue organization, suggestion for enhancement has been made focusing on the issues most frequently raised. This NLP-based approach added further degrees of insightfulness and potentially of impact compared with the purely categorically-based analysis by mapping passengers' comments directly back to potential operational inefficiency and therefore to points for intervention to enhance airport performance.

Deep Learning Model:

6) Feature-Interaction Enhanced XGBoost and Deep Learning: This approach is a combination of XGBoost and Feed Forward Neural Network (FNN). Initially, FNN is trained with the data and the output which is the probability scores for each class is added as an extra feature in the dataset. This updated dataset is used by XGBoost, which makes the final prediction using both the original and neural network generated features.

Along with the features from the dataset, some additional interaction features are also created and used to uncover any hidden patterns. Following are some of the interaction pairs that are used: carrier delay * weather delay

- Average wind speed x Late aircraft count
- Month x NAS Delay
- Average Temperature x Average Percipitation

Out of all these features, only top features which are important are selected and FNN is trained on those features, before using XGBoost.

Implications: Accuracy of the model is around 88%, which means that this is the best model out of all the ones that were tried. When Feed forward Neural Network is employed, initial accuracy was around 83%, and only when it's combined with XGBoost, the accuracy got improved.

B. Performance Measures

For the purpose of performance evaluation of models used for classification as well as clustering, several performance parameters are applied on both the models. The above metrics assist in deterministic of the accuracy, reliability and efficiency of the models.

1) Supervised Learning Models: Following metrics were used for classification models:

Accuracy: Learns how accurately instances are classified. It offers an average idea of the models' performance and shows how technically sound the models are.

Precision, Recall, and F1-Score: These metrics were used to measure performance per class and demonstrated how effectively models can classify delays.

- Precision is the ratios of true positive instances ensuring that false alarm is minimized and that the recommendations are quite clear.
- Recall is the ratios of actual positive instances that are identified thus has a direct influence on identifying major delays.
- F1 score which is the weighted average between precision and recall prevents a sole overemphasis of either of the two ensuring depth in the evaluation.

2) Unsupervised Learning Model: Following metric is used for clustering model:

Silhouette Score: Applied to evaluate the quality of clustering to assess the possible effects of natural clusters in the pattern of delays.

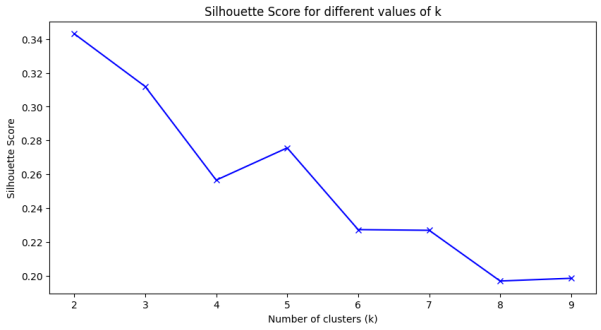

Fig. 1. Silhoutte Score across different Clusters

C. Model Parametrization

Model parameterization is a process of picking and optimization of the relevant parameters that determines the performance of the model.

1) Supervised Learning Models: Algorithms like XGBoost, Random Forest and SVM popular algorithms are learned first with standardized hyper parameters and then learning parameters are tuned using cross validation. Cross-validation ensures that the model isn't over-fit, protecting the validity of the model across various conditions since it gives a more reliable evaluation of the generalization ability of the model than the standard train/test split of the dataset.

2) Unsupervised Learning Models: In K-Means Clustering the number of partitions or clusters that is denoted by k is often set at a range of 2 to 9. The Silhouette Score is determined for each k, and k is set as the number of clusters that yield the highest Silhouette Score of 0.343 for high delay and low delay.

V. Conclusion and Future Work

This study aimed at examining the flight delays and analysing them by using various techniques in machine learning in order to categorize the flights in Low Delay, Moderate Delay, and High Delay categories. For the analysis of the flight delay based on operational metrics and weather conditions, models like XGBoost, Random Forest, SVM, KMeans clustering and Feature-Interaction Enhanced XGBoost and Deep Learning was used by this study. The findings indicate:

- Among all the proposed models, Feature-Interaction Enhanced XGBoost and Deep Learning was the best performing model with accuracy around 88%, followed by XGBoost which had accuracy around 85.08%.
- Random Forest and SVM had given the similar performance, and features importance and recall, which helped in understanding the delay causes.
- K-Means Clustering helped to sort out airports into high and low-delay classes, which also contributed into knowing the airports segmentation according to the delay pattern.

Based on the prediction of high-delay airports from each model, a consolidated list of high delay airports is taken and airport specific recommendations were given for those airports.

A. Limitations

While the study achieved its primary objectives, several limitations were identified:

- **Moderate Delay Classification:** Moderate Delay category was hard to go with as its characteristics were very close to the other two classes for the models such as XGBoost, Random Forest. The F1-scores are slightly lower for the Moderate Delay videos, suggesting that potentially, the features adopted do not differentiate enough.
- **Feature Representation:** The weather and operational data used, although extensive, were not detailed as those at an even finer level including hourly or per flight level. Also, the addition of more specific parameters like air traffic control delay or crew schedule problem could increase the efficiency of the model even more.

B. Future Work

To extend the study and address its limitations, the following future directions are proposed:

1) Enhanced Feature Engineering:: Other factors, including live flight data and characteristics of the aircraft and crews, and the levels of congestion of airports could enhance the models capacities regarding moderate delay classification. Additional temporal characteristics might also be useful, for example, data for specific hours or even day of week or monthly flight data might be beneficial.

2) Deep Learning Approaches:: Analyse the way in which the event of delay repeats itself over time, and use Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to approach it. These models could offer more detail, particularly in relation to time series factors, used in predicting delays.

3) Natural Language Processing (NLP) Extensions:: In future research, a valuable addition to the analysis could be the trends of sentiment dynamics to reveal shifts in passenger attitudes and the relationship between sentiment and the operational results of airport work.

4) Operational Integration:: Some measures could cover the utilization of model outputs to real-time recommendation to airport management systems. The application of a dashboard that displays the predictions and recommendations of a built model could present useful information for airport operators.

C. Key Implications

The research question is answered by the findings of this study as it accomplishes the task of categorizing delays and offering relevant airport improvements. The findings obtained from the classification models are beneficial for the airports which want to decrease the delay and increase operational efficiency. The airport specific improvements obtained from the analysis are conducive to impacting the operations at airports that experience high levels of delay occurrences.
These models and NLP techniques are reproducible thus can be used in other airports in other regions. Due to the further broadening of the range of features and the use of more sophisticated modeling, the benefits of this work could be

further enhanced, and the efficiency of airport work and passenger experience could be increased as a result.

REFERENCES

[1] Federal Aviation Administration. The economic impact of u.s. civil aviation: 2024. Technical report, Federal Aviation Administration, 2024. Accessed: 2025-03-03.

[2] Yuemin Tang. Airline flight delay prediction using machine learning models. *ACM International Conference Proceeding Series*, pages 151–154, 10 2021.

[3] Bureau of Transportation Statistics. Flight delay causes, 2025. Accessed: 2025-03-03.

[4] Rodrigo Britto, Martin Dresner, and Augusto Voltes. The impact of flight delays on passenger demand and societal welfare. *Transportation Research Part E: Logistics and Transportation Review*, 48:460–469, 3 2012.

[5] A. Evangeline, R. Catherine Joy, and A. Albert Rajan. Flight delay prediction using different regression algorithms in machine learning. *ICSPC 2023 - 4th International Conference on Signal Processing and Communication*, pages 262–266, 2023.

[6] Sai Sharan, M. Sriniketh, Harsha Vardhan, and Dannana Jayanth. State-of-art machine learning techniques to predict airlines delay. *2021 International Conference on Forensics, Analytics, Big Data, Security, FABS 2021*, 2021.

[7] Yanjun Wang, Yakun Cao, Chenping Zhu, Fan Wu, Minghua Hu, Vu Duong, Michael Watkins, Baruch Barzel, and H. Eugene Stanley. Universal patterns in passenger flight departure delays. *Scientific Reports 2020 10:1*, 10:1–10, 4 2020.

[8] Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the united states. 10 2010.

[9] Yash Tijil, Nripendra Dwivedi, Satyam Kumar Srivastava, and Anmol Ranjan. Flight delay prediction using machine learning techniques. *Proceedings - International Conference on Computing, Power, and Communication Technologies, IC2PCT 2024*, pages 1909–1913, 2024.

[10] P. V. Kavitha, Ln Manoranjani, V. Mithra, and P. Monal. Flight delay prediction using machine learning model. *2022 International Conference on Futuristic Technologies, INCOFT 2022*, 2022.

[11] Pavel Stefanovič, Rokas Štrimaitis, and Olga Kurasova. Prediction of flight time deviation for lithuanian airports using supervised machine learning model. *Computational Intelligence and Neuroscience*, 2020, 2020.

[12] Aileen Chun Yueng Hong, KHAI WAH KHAW, XINYING CHEW, and WAI CHUNG YEONG. Prediction of us airline passenger satisfaction using machine learning algorithms. *Data Analytics and Applied Mathematics (DAAM)*, 4:7–22, 4 2023.

[13] Nishant Sharma and S. Vijayalakshmi. Flight arrival delay prediction using deep learning. pages 1–6, 9 2024.

[14] Blessy Trencia Lincy S. Trencia, Hannah Al Ali, Ahmad Abdulla Abdulaziz Mohd Majid, Omeer Arif Abdelbaqi Abdalla Alhammadi, Aysha Momen Yousuf Mohammed Aljassmy, and Zindoga Mukandavire. Analysis of flight delay data using different machine learning algorithms. *New Trends in Civil Aviation*, 2022-October:57–62, 2022.

[15] N. Lakshmi Kalyani, G. Jeshmitha, Bindu Sri U. Sai, M. Samanvitha, J. Mahesh, and B. V. Kiranmayee. Machine learning model - based prediction of flight delay. *Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020*, pages 577–581, 10 2020.

[16] Ravi Kothari, Riya Kakkar, Smita Agrawal, Parita Oza, Sudeep Tanwar, Bharat Jayaswal, Ravi Sharma, Gulshan Sharma, and Pitshou N. Bokoro. Selection of best machine learning model to predict delay in passenger airlines. *IEEE Access*, 11:79673–79683, 2023.

[17] Irmak Hatıpoğlu and Ömür Tosun. Predictive modeling of flight delays at an airport using machine learning methods. *Applied Sciences 2024, Vol. 14, Page 5472*, 14:5472, 6 2024.

[18] Ashok Reddy Kandula, Potru Sandhya Priya, Mohammad Sameena Simmin, Vesangi Naga Arathi, and Mohammad Afrin. Flight delay prediction using neural fusion network. *3rd International Conference on Automation, Computing and Renewable Systems, ICACRS 2024 - Proceedings*, pages 920–926, 2024.

[19] Jingyi Qu, Ting Zhao, Meng Ye, Jiayi Li, and Chao Liu. Flight delay prediction using deep convolutional neural network based on fusion of meteorological data. *Neural Processing Letters*, 52:1461–1484, 10 2020.

[20] Jun Chen and Meng Li. Chained predictions of flight delay using machine learning. *AIAA Scitech 2019 Forum*, 2019.

[21] Vijayarangan Natarajan, Swaminathan Meenakshisundaram, Gautham Balasubramanian, and Shubham Sinha. A novel approach: Airline delay prediction using machine learning. *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, pages 1081–1086, 12 2018.

[22] Azib Anees and Wei Huang. Flight delay prediction: Data analysis and model development. *2021 26th International Conference on Automation and Computing: System Intelligence through Automation and Computing, ICAC 2021*, 2021.