Enhancing Flight Delay Prediction Using Machine Learning Techniques

By

Kilonzo, Kelvin Ndambu

168209

Submitted in Partial Fulfilment of the Requirements for the Degree of

Master of Science in Data Science and Analytics of Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

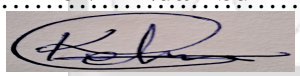
Nairobi, Kenya

June, 2025

Declaration and Approval

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name: **Kelvin Ndambu Kilonzo**

Signature: 

Date: May 23, 2025

Approval

The dissertation of Kilonzo Kelvin Ndambu was reviewed and approved by the following:

Prof. Bernard Omolo

Visiting Professor, Oguna-Omolo

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Flight delays remain a persistent challenge within the U.S. aviation sector, impacting both operational efficiency and passenger satisfaction. This study presents a comprehensive machine learning framework to accurately predict flight arrival delays in minutes using an enriched dataset comprising historical flight records and weather conditions.Multiple models were developed and evaluated, with particular emphasis on deep learning approaches. A fine-tuned Long Short-Term Memory (LSTM) network outperformed other models, achieving a high coefficient of determination ($R^2 = 0.9385$), a low Mean Absolute Error (MAE) of 10.7 minutes, and a Root Mean Squared Error (RMSE) of 15.3 minutes. Comparative models such as XGBoost and Artificial Neural Networks also demonstrated strong performance, validating the robustness of the feature set. To ensure interpretability and model transparency, Local Interpretable Model-Agnostic Explanations (LIME) were applied, revealing that departure delay, time-of-day, and specific carriers were among the most influential features. The final LSTM model was deployed through a user centric application, enabling real time delay prediction based on user flight details.

KEY WORDS: Scheduled Time of Arrival, Actual Time of Arrival, Delays, Airport Artificial Neural Network, Machine Learning, LSTM.

Table of contents

List of figures

List of tables

List of Abbreviations

ANN	Artificial Neural Network	AUC	Area Under the Curve
BTS	Bureau of Transportation Statistics	CRISP-DM	Cross Industry Standard Process for Data Mining
GBM	Gradient Boosting Machine	LSTM	Long Short-Term Memory
MAE	Mean Absolute Error	MAPE	Mean Absolute Percentage Error
ML	Machine Learning	NOAA	National Oceanic and Atmospheric Administration
RMSE	Root Mean Square Error	ROC	Receiver Operating Characteristic
STA	Scheduled Time of Arrival	SVM	Support Vector Machine
TLP	Two-Layer Prediction	UNSDG	United Nations Sustainable Development Goals
USD	United States Dollar		

Acknowledgement

First and above all, I am grateful to the Almighty God for His provision, grace, and for granting me good health throughout the study period.

I am grateful to Mr. Nzomo for his tremendous support in my journey and for making this possible, my parents, brothers, L.N and work colleagues for their immeasurable support throughout the study period.

I am indebted to my supervisor, Prof. Bernard Omolo for his availability, kind guidance, and support without which this thesis would not have been a success.

I am grateful to my classmates for their support and encouragement.

Dedication

This dissertation is dedicated to God Almighty for giving me wisdom and good health.

Chapter 1

Introduction

1.1 Background to the Study

The aviation industry, the fastest and safest means of transport, has experienced tremendous growth in the past decade and it is predicted that this will double by 2035. This surge can be attributed to the increase in living standards, people getting to understand the importance of time, flight ticket offers by airlines to promote their services (Dhanawade et al., 2019; Sreenivasulu et al., 2023). In Kenya, a number of airlines have ventured into the aviation industry with Jambojet celebrating 10 years in the market as at April 2024 and already having 52% of the market share locally (Jambojet, 2024). This is attributed to its ability to offer its clients competitive fares and loyalty programmes, being Kenya's first low cost airline.

Globally, airlines are ranked based on how they perform in different categories like customer service, crew, on-time performance as well as safety. According to Skytrax, a company that runs worldwide airline and airport reviews and awards rankings, Ethiopian Airlines was ranked as the best airline in Africa for the seventh consecutive time (Skytrax, 2024).

Qatar Airways was ranked as the overall best airline in the world in the 2024 awards based on customer preferences (Skytrax, 2024). Qatar Airways' success can be attributed to its investment in data analytics, employing realtime monitoring of flights by the analytics team. In May 2023, Qatar Airways and Google cloud collaborated on Data and AI to advance the Airline's Passenger experience hence providing personalized offerings to its millions of passengers (Airways, 2023). In the U.S, the biggest consumer of the aviation industry with a total of 366 airports and with thousands of flights operating per day, Delta Air lines has been ranked as the best airline for the fourth year in a row.

As the aviation industry continues to grow, the consumers of the services in the industry heavily depend on such analytics and insights as they choose the airlines to use. With this growth, there comes a number of challenges and one that stands out is one of flight delays (Gui et al., 2020). According to the Bureau of Transport Statistics, a flight is considered delayed if it arrives at its destination gates 15 minutes later than its scheduled time of arrival or departs 15 minutes later than its scheduled time of departure, (Lykou et al., 2020). Delayed flights can have adverse effects on the passengers, airlines as well as the airports. For passengers, a delayed flight can result in missed connecting flights, anger, and reduced brand loyalty due to frustrations (Sreenivasulu et al., 2023; Wang et al., 2022).

Flight delays are attributed to a number of factors that are unavoidable like weather conditions, which is a dominant factor especially is the U.S.; operational inefficiencies such as scheduling and maintenance; and traffic congestion and security issues (Nivitha et al., 2023; Sreenivasulu et al., 2023). Further, these delays have an adverse effect on the environmental pollution, by causing more wait time and longer engine running times, which is against what Sustainable Aviation Fuels(SAF) advocates for in its goal to achieve net-zero emissions by 2050. Flights in the U.S. consume an additional 740 million gallons of jet fuel which amount to $ 1.6 billion in extra costs (Anees and Huang, 2021).

In 2023, Avianca airline was ranked as the most punctual airline in the world. Avianca achieved an on-time arrival (OTA) of 85. 73% with a flight completion factor of 99.08% out of 213,039 flights. Azul, a Brazilian airline came second achieving OTA of 85.51%, with a completion factor of 97.12% out of 310,972 flights. Qatar Airways was ranked third, after achieving an OTA of 85.11% with a completion factor of 98.32% out of 183,090 flights (Skytrax, 2023). This highlights one of the reasons it was ranked as the best airline in the world by Skytrax.

Airports also play a major role when it comes to a flight's on time performance based on how they handle congestion, air traffic as well as the resources such as the number of runways an airport has. For instance, according to (Skytrax, 2023), the airport that recorded the best on time performance globally was Cape Town Airport with 86% of the flights being on time. Brasilia-Presidente Juscelino Kubitschek Aiport, Brazil, came in second while Panama City Tocumen Airport was ranked third. Flight delays can create significant management challenges by causing

congestion on airport surfaces and within terminal areas, primarily due to the accumulation of aircraft. This congestion can lead to a decline in the overall performance and efficiency of the airport network (Feng et al., 2021).

1.2 Statement of the Problem

Air transport has become one of the major means of transport globally. However, flight delays have become a persistent issue leading to dissatisfaction among the users of this service. More than 20% of fights in the U.S. experience delays which lead to losses amounting to billions of dollars. Flight disruptions cost passengers more than $20 billion annually, while airlines incur an average of $26 billion outdated technologies used in flight delay predictions annually (Nivitha et al., 2023).

Current models used to predict delays focus on classification techniques, which limit decision making processes by only predicting if a flight is delayed or not (Meel et al., 2020; Purushothaman et al., 2024). Furthermore, linear regression models used cannot handle complex and large scale data. Most researchers have incorporated deep learning techniques which have shown dominance in-terms of predicting delay minutes (Divya et al., 2023). However, their work does not incorporate factors that lead to these delays like weather conditions.

The inability to have models that can incorporate flight schedules and factors that lead to flight delays leads to inefficiencies, increased fuel consumption and environmental harm due to emissions (Anees and Huang, 2021). There is a need for advanced predictive models that can capture the complex nature of aviation big data and the factors that cause flight delays. This research aims to address this gap by developing machine learning and deep learning models that incorporate historical flight data and weather patterns to provide more accurate delay duration. Such robust predictive models could lead to optimization of airline operations, airport operations, giving insights to passengers on time and minimizing environmental impacts.

1.3 Research Objectives

1.3.1 General Objective

The main objective of this research is to utilize machine learning and deep learning techniques incorporating weather patterns to accurately predict the delay duration of a flight.

1.3.2 Specific Objectives

1. To analyze historical flight data integrated with weather data to identify the factors that majorly contribute to flight delays?
2. To compare different machine learning models and deep learning models performances based on evaluation metrics?
3. To deploy the best performing model to a user friendly web application?

1.4 Research Questions

1. Which factors have the highest impact on flight delays?
2. Which model performed best based on the evaluation metrics compared?
3. Which deployment platform is the best for deploying the model and allow continuous monitoring of the model's performance?

1.5 Justification of the Study

Disruptions in airlines are an inevitable occurrence that every airline and airport has to deal with. This is due to the nature of the factors that cause delays. In order to reduce the delays, the associated costs, carbon emissions, and passenger inconveniences caused, an accurate model

that can predict the delay minutes of an aircraft is desirable. This model will also enable the airlines to reduce costs associated with cancelling flights and ensure that schedules are returned to normalcy, and ensure customers are advised on changed schedules on time. In the end, the airlines in focus will record an improvement in on-time performance, which has a direct implication on its profits.

1.6 Scope and Limitations

1.6.1 Scope of the Research

This research focused on the prediction of flight delay durations in the U.S. airline industry using machine learning models. The study incorporated both historical flight data and weather conditions from January 2022 to December 2024 to develop a model capable of accurately predicting delay times in minutes. The research targeted the U.S. domestic aviation system, which has a large and complex structure. The output of the best machine learning model was deployed to a user-friendly interface designed for airlines and passengers, enabling them to input flight details and receive real-time delay predictions.

1.6.2 Limitations of the Research

The primary limitation of this research was its dependence on the availability and accuracy of real time data, particularly weather data, which can vary in quality and coverage. Additionally, the study was confined to U.S. domestic flights, limiting the generalizability of the model to international flights or other regions with different operational dynamics. Moreover, the model did not fully account for rare, unpredictable events, such as sudden technical failures or security issues, which can cause delays but are difficult to predict using historical data.

Chapter 2

Literature Review

2.1 Introduction

Here we provide an overview of the models developed so far and understanding the causes of delays. Predicting the duration of flight delays have become important areas of research, especially in the United States, where over 20% of flights experience delays annually (Wang et al., 2022). Delay time is the difference between the scheduled time of departure and the actual time of departure (Evangeline et al., 2023). A flight is considered delayed if it arrives at its destination gates 15 minutes later than its scheduled time of arrival or departs 15 minutes later than its scheduled time of departure (Lykou et al., 2020).

2.2 Models

2.2.1 Classification Techniques in Flight Delay Prediction

Much of the research on flight delays has focused on classification models, which predict whether a flight will be delayed or not. These classification models often rely on machine learning algorithms such as Decision Trees, Random Forest, and Logistic Regression. These models have proven to be effective in identifying delays based on historical data. For example, Meel et al. (2020) utilized a Random Forest classifier to predict flight delays based on flight schedule data combined with weather, reporting that Random Forest produced superior accuracy compared to other algorithms like Decision Trees and Naive Bayes. However, while these models are useful in binary classification (delayed or not delayed), they do not provide insights into the actual delay duration, which is a critical factor for airlines and passengers.

Anand (2023) in his work leverages PySpark, a data processing framework to work with aviation big data considering the complexities of data collected and required for forecasting delays. Anand (2023) leverages four most commonly used machine learning techniques to get the one that predicts accurately: Random Forest Classifier, Logistic Regression, Gradient Boosted Tree Classifier and Decision Tree. His study highlighted the use of big data in research and how PySpark can be used to allow faster and scalable data analysis. From the analysis, Random Forest Classifier and Logistic Regression are the models that stood out in terms of accuracy compared to the rest.

Wang et al. (2022) used a two-layer model, where in the first level they classified if an aircraft was delayed or not. They partitioned the delay level into 4 parts. 0 if the delay duration was 15 minutes or less, 1 if the duration was between 15 - 60 minutes, 2 if the delay was between 60-120 minutes, 3 if the duration of delay was 120 - 180 minutes and 4 if the delay was beyond 180 minutes. They then compared different classification machine learning models based on accuracy: Support Vector Machine(SVM), Stochastic Gradient Descent(SGD), KNN, Decision Tree and Random Forest. Random Forest is chosen as it has the highest prediction accuracy of 88.1%. The output for this level was then used as the input in a second layer that was a regression-based layer to predict the duration of flight delays.

For the second level, Wang et al. (2022) compared linear regression model, random forest regressor, decision tree regressor and LightGBM regressor. Random forest regressor emerged as the best model with the lowest MSE. The joint model first predicts if a flight is delayed or not, and if it is delayed, it proceeds to level two to predict the delay duration. However, if the flight is not delayed, the training stops and outputs that the flight is on time. This two layer model had a higher accuracy in predicting the delay duration compared to directly predicting the delay duration using random forest without first classifying the delay. This is a good model, however, it does not take into consideration external factors that cause delays like weather and air traffic congestion.

2.2.2 Regression Approaches in Delay Duration Prediction

To overcome the limitation of classification models, several researchers have turned to regression-based models, which predict the actual duration of a delay. Regression models are critical in forecasting how long a flight will be delayed, providing airlines and passengers with more useful insights for managing disruptions.

Wang and Pan (2022) used ARIMA time series model to predict flight delay with an aim of improving efficiency in operations and the quality of services. Wang and Pan (2022) makes the series data stationary by difference operation and during peak seasons is predicted using determination and residual test. Training data used for this case is from May 1 to May 3 2021 and the variables used are flight number, flight origin, flight destination, scheduled departure, and actual departure

Sreenivasulu et al. (2023) advanced their research in predicting flight delay duration through intelligent algorithms based on United States utilizing Hadoop which helped them scale the aviation big data consisting millions of records with accuracy. Sreenivasulu et al. (2023) studied the interconnectedness of the airports, and incorporated the Automatic Dependent Surveillance–Broadcast (ADS-B) system, which is an integrated communication and surveillance system that helps in air traffic management(ATM). Integration of air traffic data is one of the strengths in his work as most research has not incorporated this. The ADS-B offers the chance to increase accuracy of the prediction model offering realtime data inputs into the model giving real-time forecasting capabilities. On comparing Random forest, K-Nearest Neighbor, Linear Regression , logistic regressions and Support Vector Machine, random forest emerged the best with the lowest MSE.

Evangeline et al. (2023) also incorporated ADS-B data with flight schedule data and weather data. In this study. Evangeline et al. (2023) is presented with massive amount of data that cannot be delt by the standard methods of analysis and opts for a comparison of the ridge and lasso regression techniques. Their aim was to get the delay minutes in the most accurate form possible and the ridge regression out perfoms the lasso regression with an accuracy score of 99.89%. Evangeline et al. (2023) highlights on the importance of accurately predicting flight

arrival delays as they have a huge impact on the next flight delays not forgetting they impact they have on passengers who have connecting flights.

2.2.3 Deep Learning Techniques for Delay Prediction

Jiang et al. (2020) in his article studied flight delay arrival by comparing several machine learning models using Airline On time perfomance data and Quality Controlled Local climatological Data. Jiang et al. (2020) considered weather as a factor of study that affects delays and used a five-class delay classification together with historical flight delay patterns. Unlike previous methods of forecasting, Jiang et al. (2020) used convolutional neural network model to help in pattern recognition to study the trends that lead to flight delays in the US. Jiang et al. (2020) compared Support Vector Machine, Decision Tree, Random Forest and Multilayer Parceptron models and the Multilayer Parceptron emerges as the best model for forecasting with an accuracy of 89.07%. However, the focus of this to study was to determine if a flight was delayed or not rather than predicting the delay duration which would be more informative.

Cai et al. (2022) used a deep learning approach on various airports unlike most researchers who have put their focus on specific airports of study. Cai et al. (2022) studied the time varying spacial interactions that is hidden in interconnected airports through time evolving graphs. Cai et al. (2022) explored 74 civil airports in China in 2018 and by comparing his models to other models that do not consider the interconnection of various airports, his deep learning grahical model performs better than classical models where interactions among airports are ignored. However, this study only focused on the airport delays as a whole without considering the airline specific operations as specific airline schedules have an impact on delays.

Dou (2020) focused on flight arrival delay using ensemble learning. In his paper, he focused on US domestic flights on-time perfomance using a Cat-boost model which is a novel gradient boosting technology that takes into consideration weather which is one of the factors that affects flight delays combined with Logistic Regression Model. To improve operational efficiency, and to put his work to significance, predicting whether a flight is delayed or not is not enough, for decision making, he went ahead and predicted the duration of the delay in minutes. Arrival

delay time is the dependent variable while 12 influencing factors are the independent factors in his work such as quarter, distance and carrier. The research employed grid search and cross-checking methods in determination of optimal parameters and fit the data to selected models. From this study, Cat-boost deals better with categorical variable encoding problems and imbalanced datasets which are dominant in flight delay prediction considering that on time flights are about three to four times more compared to delayed flights.

Gui et al. (2020) explored flight delay with a number of factors that cause delay unlike most papers that only explore a single route or airport: weather, airport traffic flow, leave port speed, arrive port speed, season, holiday, unexpected event, historical traffic flow, peak traffic flow, date, flight number, scheduled time, pre-flight, historical flight delay.The dataset used is built from an Automatic Dependent Surveillance-Broadcast (ADS-B) where messages are received, pre-processed, and integrated with the weather data, airport information and schedules. LSTM and random forest- based architectures are employed to predict individual flight delays. LSTM is good in handling time structures while random forest is good when it comes to classification on whether a flight is delayed or not. Comparing the two models, the random forest obtained accuracy of 70% while the LSTM obtained an accuracy of 99% on the training dataset. However, on the testing set, the random forest was more superior than the LSTM. This can be attributed to the size of the dataset used which was collected from December 2018 to May 2019, and had 5761 items.

In response to the limitations of using either classification or regression models in isolation, hybrid models have emerged as a more effective approach. These models combine machine learning techniques with deep learning architectures to leverage the strengths of both approaches. For instance, Divya et al. (2023) integrated a Genetic Algorithm with an Artificial Neural Network (ANN) to predict flight delays. The genetic algorithm optimized the parameters of the ANN, significantly improving the model's ability to predict both whether a flight would be delayed and by how long.

2.3 Gaps in Existing Research

Based on the literature, several studies have been done to predict flight delays using machine learning techniques. Wang et al. (2022) and Meel et al. (2020) used classification techniques to predict if a flight is delayed or not. However, these models might have been effective for binary classification, they did not provide insights on the actual duration of delay. These models did not adequately explore how the different factors that cause the delays influenced the model, such as weather conditions and air traffic conditions.

Some researchers like Wang et al. (2022) tried to overcome this shortcoming of binary classification, by employing a hybrid model to predict delay duration. While this model was more accurate, it did not incorporate factors like weather which is a major factor that cause flight delay (Dhanawade et al., 2019). Furthermore, while deep learning techniques have also been introduced, such as convolution neural networks (CNN) to forecast delays, (Cai et al., 2022; Jiang et al., 2020), these factors have eliminated the concept of airport chain effect. This is where delays in one airport could lead to delays in the next airport as shown by Cai et al. (2022) in a study based in China airports.

What remained unaddressed was the use of models that can capture the complex nature of aviation big data that is in millions of records, integrated with factors that lead to delays such as weather conditions. There is a gap in using such models and data specifically focusing on the U.S. aviation industry. (Divya et al., 2023) emphasized the importance of transitioning predictive models into production for practical application within the aviation industry. However, the majority of studies remained theoretical, with few researchers taking the additional step to deploy their models in real-world settings.

This research aimed at bridging that gap by deploying the best model using streamlit which to enabled users to input details such as flight schedules and weather data using APIs to ensure real time weather forecasting. This model was trained using data that spans to two years to avoid instances of over-fitting. Gui et al. (2020) used LSTM, but as a result of using data that spans to 6 months, their model ended up over-fitting the two-year data span ensures a wide range of seasonal and operational variables, making the model more robust.

Chapter 3

Methodology

3.1 Introduction

This section details the methodology used to achieve the research objectives of predicting flight delay durations using machine learning techniques. The study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, encompassing data collection, pre-processing, modeling, evaluation, and deployment. The methodology aligns with the specific objectives of the study, detailing how data was analyzed, models were built and trained, their performance was compared, and the best model was deployed.

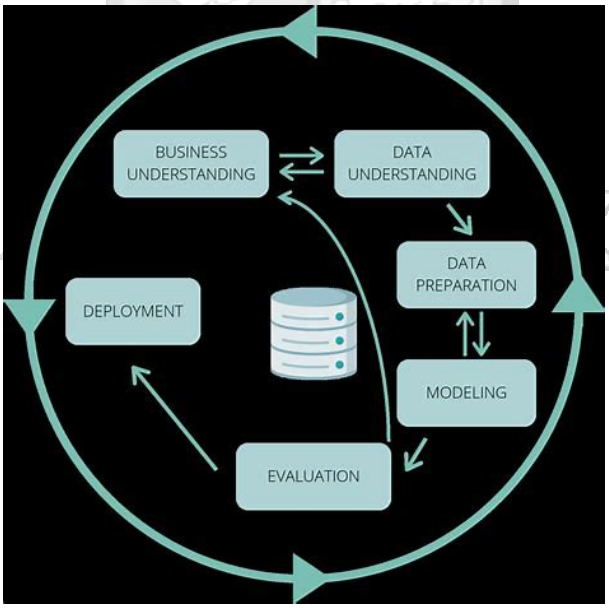


Figure 3.1: Crisp D-M Cycle

3.2 Data

3.2.1 Data Sources

The data used in this study was collected from the Bureau of Transportation Statistics (BTS) and the National Oceanic and Atmospheric Administration (NOAA). The dataset from BTS consists of detailed records of flight schedules, including departure and arrival times, airline-specific delays, flight numbers, and airport operations. This dataset spans from January 2022 to December 2024, covering over 366 airports across the United States, with approximately 1 million flight records. The NOAA dataset provides corresponding weather conditions such as precipitation, humidity, wind speed, and temperature, for the same period, ensuring meaningful correlations between flight delays and atmospheric conditions.

The BTS dataset was obtained in CSV format and preprocessed for integration, while the NOAA weather data was accessed and downloaded via an API that allowed for dynamic retrieval, and stored as CSV. The raw datasets contained structured tabular data, where each record represented an individual flight with details on scheduled and actual departure and arrival times, and meteorological conditions.

3.2.2 Data Integration

The integration of flight and weather data was performed to create a comprehensive dataset suitable for predictive modeling. The BTS flight data was merged with NOAA weather data using a left join operation, ensuring that all flight records retained their corresponding weather conditions. The merge was conducted on Date,time stamp, and Origin Airport as these were the common identifiers in both datasets. By using these attributes, each flight record was accurately matched to the prevailing weather conditions at the departure and arrival locations.

Given that flight schedules are recorded with precise timestamps while weather data is reported at hourly intervals, linear interpolation was applied to assign the most relevant weather readings to each flight's scheduled departure and arrival times. This ensured a more accurate repre-

sentation of the atmospheric conditions influencing flight operations. Additionally, time-zone normalization was performed to standardize timestamps across different airports, preventing temporal mismatches in the dataset.

To further refine the dataset, validation checks were conducted to ensure consistency and completeness. Duplicate records were removed to prevent redundancy, while missing values in weather features were handled using appropriate imputation techniques. Anomalous data points, such as negative delay durations or unrealistic temperature values, were flagged and corrected where possible. The final integrated dataset contained over 70 million flight records, each enriched with weather-related variables, making it a robust dataset for predictive modeling.

3.2.3 Key Features

The key features used in this study are detailed in Table 3.1

Table 3.1: Dataset attributes integrated to train the proposed model

Characteristic	**Input Variables**
Flight Information	Departure airport
	Arrival airport
	Flight number / Tail number
	Airline carrier
	Actual departure and arrival time
Scheduled Times	Scheduled departure time
	Scheduled arrival time
	Flight date
	Day, month, and weekday
Delay Indicators	Departure delay (Minutes)
	Arrival delay (Minutes)
Weather Conditions	Temperature
	Humidity
	Wind direction
	Wind speed
	Atmospheric pressure

To select the features that have a high impact on delay minutes, we used a correlation matrix. This showed how the different features impact a flight's delay. This improved the model's efficiency and reduced the chances of overfitting. By doing that, we achieved our first objective of determining the factors that have the highest impact on flight delays.

3.2.4 Data Preprocessing

Once the dataset was collected and integrated, we proceeded with pre-processing to ensure the data was clean, consistent, and suitable for training predictive models. This step involved handling missing values, encoding categorical variables, scaling numerical features, and reducing dimensionality where necessary.

Handling Missing Data

Missing values are common in large datasets, and addressing them appropriately is crucial to maintaining the integrity of the model. In our dataset, missing values were primarily present in weather-related attributes such as wind speed, precipitation, and humidity. We handled these by applying mean imputation for continuous numerical variables, ensuring that each missing value was replaced by the average of its respective column. For categorical variables, including airline name and airport identifiers, mode imputation was used, filling in missing entries with the most frequently occurring value in that category. Any rows where more than 50% of the data was missing were removed to prevent noise from affecting model performance.

Categorical Encoding

Machine learning models work best with numerical data, so categorical variables needed to be transformed into a numerical format. We applied one-hot encoding to categorical features such as airline, origin airport and destination airport. This method prevents numerical misrepresentation and ensures that models do not assign unintended weight to categorical attributes. For ordinal features, where values followed a specific ranking order, label encoding was applied to preserve their relative importance.

Feature Scaling

Given the varying scales of different features, normalization was required to bring all numerical variables into a comparable range. Flight delays are measured in minutes, whereas wind speed is

recorded in kilometers per hour, which could introduce scale bias in machine learning algorithms. To mitigate this, we used a standard scaler, transforming all numerical values to a range between -1 and 1. This method maintains the relationships between values while ensuring that no feature disproportionately influences the model's learning process.

Feature Selection and Dimensionality Reduction

To improve computational efficiency and avoid overfitting, we performed feature selection by examining correlations among variables. A correlation matrix was used to identify features that were strongly correlated with flight delays while filtering out redundant attributes.

Final Dataset Preparation

Following these preprocessing steps, we obtained a structured and refined dataset ready for model training. The cleaned dataset was then split into training and testing sets, ensuring that the model could be validated effectively on unseen data. By carefully preprocessing the data, we aimed to enhance the accuracy and reliability of our flight delay prediction models.

3.3 Modeling

In this section, we explored the machine learning and deep learning models chosen to predict flight delays in minutes, and how they will process the available data. This section helped us achieve our second objective. Machine learning is the ability of a machine to learn from data and be able to make decisions on its own. The data in this case includes historical flight information such as flight number, origin, destination, scheduled departure time and weather conditions such as precipitation, wind speed, and temperature. We will use five models: Random Forest, XGBoost, Long Short-Term Memory (LSTM),and Artificial Neural Networks (ANN). Each model was trained to capture the non-linear relationships between these features and flight delays, with an emphasis on minimizing error in the delay duration prediction.

3.3.1 Random Forest Regressor

Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions. Each tree is trained on a bootstrapped subset of the data, making Random Forest robust to overfitting. In the context of flight delay prediction, Random Forest will learn from features like flight number, origin, destination, scheduled departure time, and weather conditions. The model will capture non-linear relationships between these features and the delay time. The final prediction, $\hat{y}$, is the average prediction from all trees:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} f_m(x), \tag{3.1}$$

where:

- $\hat{y}$ is the predicted delay in minutes.
- $f_m(x)$ is the prediction from the m^{th} decision tree.
- x represents the input features (flight details, weather data).

Baseline Modeling with Random Forest Regressor

As a foundational step in model development, a Random Forest Regressor was employed to establish a benchmark for arrival delay prediction. The Random Forest model, being an ensemble of decision trees, is capable of capturing non-linear relationships without requiring extensive preprocessing or feature transformation, making it ideal for initial experimentation.

Model Configuration and Training

The Random Forest model was trained using the `RandomForestRegressor` class from the `scikit-learn` library. Key hyperparameters included:

- `n_estimators = 100`: Number of decision trees in the ensemble

- `max_depth = None`: Trees were allowed to grow until pure leaves or all leaves contained fewer than the minimum number of samples
- `random_state = 42`: Ensured reproducibility
- `n_jobs = -1`: Utilized all CPU cores for parallel training

Significance of the Baseline

The Random Forest Regressor served two key purposes:

- Provided an interpretable benchmark to gauge feature importance and relative predictive power of different variables.
- Helped guide hyperparameter tuning and feature engineering for subsequent deep learning models.

By establishing a solid baseline with Random Forest, the study ensured that performance improvements from deep learning were meaningful and not simply due to overfitting or unnecessary model complexity.

3.3.2 XGBoost Regressor

XGBoost is a powerful gradient boosting technique that builds decision trees sequentially, where each tree corrects the errors made by the previous one. XGBoost is particularly well-suited for large, structured datasets like flight records and can handle the complex interactions between weather and flight information. Extreme Gradient Boosting was implemented in this study as a powerful and efficient ensemble learning method tailored for regression tasks. It operates by building an additive sequence of decision trees, where each new tree attempts to correct the residual errors made by the previous ones through a process known as gradient boosting. This iterative correction is guided by minimizing a regularized loss function, which combines a convex loss and a penalty term to control model complexity and prevent overfitting. In this implementation, XGBoost was trained on features such as encoded airport and carrier

identifiers, weather variables, cyclical temporal encodings, and holiday indicators. The model's hyperparameters such as learning rate, maximum tree depth, and subsample ratio were optimized through grid search.

The model minimizes the following objective function:

$$\mathscr{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{3.2}$$

where:

- $l(y_i, \hat{y}_i)$ is the loss function (typically Mean Squared Error) between actual and predicted delays.
- $\Omega(f_k)$ is a regularization term to prevent overfitting.

Enhanced Modeling with XGBoost Regressor

Following the baseline Random Forest model, an Extreme Gradient Boosting (XGBoost) Regressor was implemented to improve predictive accuracy and model efficiency. XGBoost, a powerful ensemble learning algorithm based on gradient-boosted decision trees, has proven effective for structured data regression tasks due to its ability to handle missing values, regularize complexity, and prevent overfitting.

Hyperparameter Tuning

To optimize model performance, RandomizedSearchCV was used to tune hyperparameters over multiple folds. The search space included:

- `learning_rate`: $\{0.01, 0.05, 0.1\}$
- `max_depth`: $\{3, 5, 7, 9\}$
- `n_estimators`: $\{100, 200, 400\}$

- `subsample`: $\{0.6, 0.8, 1.0\}$
- `colsample_bytree`: $\{0.6, 0.8, 1.0\}$
- `gamma`, `reg_alpha`, and `reg_lambda` for regularization

The best performing configuration was determined using cross-validated performance on the training set.

3.3.3 Long Short-Term Memory (LSTM)

LSTM is a type of recurrent neural network designed to handle sequential and time-series data, making it ideal for predicting flight delays based on historical flight information and time-varying features like weather conditions. The LSTM will use historical sequences of flight delays and real-time weather data to predict delays for upcoming flights. The core of LSTM is the gating mechanism, which controls the flow of information through the network:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{3.3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{3.4}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \tag{3.5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{3.6}$$

$$h_t = o_t * \tanh(C_t), \tag{3.7}$$

where:

- f_t, i_t, o_t are the forget, input, and output gates, respectively.
- h_t is the hidden state at time step t.
- x_t is the input at time step t, which includes weather data and past delays.
- C_t is the cell state.

LSTM Model Development and Optimization

To harness the temporal nature of flight delay patterns, a Long Short-Term Memory (LSTM) neural network was developed and trained on structured flight and weather data. The LSTM architecture was selected due to its capacity to model long-term dependencies and sequences in time-series data, making it well-suited for capturing patterns in flight delay influenced by historical and temporal factors.

Data Preparation

Prior to feeding the data into the LSTM network, a series of preprocessing and feature engineering steps were performed:

- **Reshaping:** The data was reshaped into a 3D format `(samples, timesteps, features)` required by LSTM models.

Model Architecture and Training

The final LSTM model was designed using Keras' `Sequential` API. The architecture consisted of the following layers:

- **Input Layer:** Accepting sequences with shape $(1, 21)$ — one timestep with 21 normalized features.
- **LSTM Layer:** A single LSTM layer with 64 units and ReLU activation to capture temporal dynamics.
- **Dropout Layer:** Dropout rate of 0.2 to prevent overfitting.
- **Dense Layers:** Two fully connected layers (one with 32 units and ReLU activation; final output layer with 1 unit and linear activation)

The model was compiled using the `mean squared error (MSE)` loss function and the `Adam`

optimizer. The performance was tracked using the `mean absolute error (MAE)` metric during training and validation.

Optimization and Early Stopping

To enhance generalization and reduce overfitting, the following training techniques were employed:

- **Early Stopping:** The training process was monitored using a validation set, and training was automatically halted when the validation loss failed to improve for 5 consecutive epochs. This ensured efficient training while avoiding unnecessary epochs.
- **Epochs:** The model was trained for up to 100 epochs with batch sizes of 128, although early stopping typically concluded training earlier.

3.3.4 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) model consists of multiple layers of neurons that process the input features and predict the delay in minutes. The ANN was used to model the non-linear relationships between input features such as flight details (number, origin, destination), weather data (precipitation, wind speed), and operational data (congestion levels). The prediction is given by:

$$\hat{y} = f\left(\sum_{i=1}^{n} w_i x_i + b\right), \tag{3.8}$$

where:

- $\hat{y}$ is the predicted delay in minutes.
- x_i are the input features (e.g., flight number, origin, destination, weather data).
- w_i are the weights of the connections.

- b is the bias term.
- f is the activation function (e.g., ReLU).

Development of Artificial Neural Network (ANN)

To establish a strong baseline for regression performance, an Artificial Neural Network (ANN) model was developed alongside other traditional and deep learning models. The ANN was designed to capture nonlinear relationships between flight delays and various influencing features, such as departure timing, airport identifiers, weather conditions, and calendar-based attributes.

ANN Architecture and Training Procedure

The ANN was implemented using the `Sequential` API from Keras and consisted of multiple densely connected layers. The initial model was configured with:

- An input layer of 21 neurons corresponding to the number of features.
- Two to three hidden layers with varying neuron sizes (64, 32, and 16 units).
- ReLU activation functions in hidden layers to model non-linearities.
- Dropout layers to mitigate overfitting.
- A final output layer with one neuron and linear activation to predict the continuous arrival delay.

Compilation was done using the `Adam` optimizer with a learning rate of 0.001. The model used `mean squared error (MSE)` as the loss function and tracked `mean absolute error (MAE)` during training.

Training Strategy

To improve model generalization and convergence, the following training techniques were applied:

- **Batch Size:** A batch size of 128 was used to balance memory efficiency and learning stability.
- **Early Stopping:** Monitored validation loss with patience of 5 epochs to avoid overfitting.
- **Epochs:** The model was allowed to train for up to 100 epochs, with actual stopping dependent on early stopping criteria.
- **Validation Strategy:** A separate 20% validation split was used to evaluate training progress and halt training appropriately.

Model Tuning

After initial training, the model underwent iterative refinement:

- Adjusting the number of neurons and layers to better capture the complexity of the data.
- Tuning dropout rates and batch sizes to enhance regularization.
- Experimenting with deeper architectures and additional hidden layers.

These refinements aimed to improve the model's capacity to generalize while avoiding excessive complexity and overfitting.

3.4 Evaluation Metrics

The evaluation metrics for our models are regression based and therefore we will take into consideration the metrics below as suggested by (Meel et al., 2020):

Mean Absolute Error (MAE)

The MAE is the absolute difference between the actual true value and the value that the model gets to predict.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{3.9}$$

where:

- y_i is the actual delay (in minutes),
- $\hat{y}_i$ is the predicted delay, and
- n is the number of predictions.

Interpretation: The lower the MAE, the better the model's predictions. Since the units of MAE are in minutes, it provides an intuitive understanding of the average error.

Mean Squared Error (MSE)

The MSE measures the average of the square of the errors. It gives more weight to large values or errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.10}$$

Interpretation: A lower MSE indicates better model performance, but because it squares the error, large errors are penalized more severely compared to MAE.

Root Mean Squared Error (RMSE)

RMSE is the square root of the MSE, and it is a better metric compared to the MSE as it brings the error back to the same units as the target variable in minutes).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.11}$$

Interpretation: RMSE gives an evaluation metric that is in the same scale as the data points, in minutes. The best model will have the smallest RMSE.

R-Squared (Coefficient of Determination)

R-squared measures the proportion of the variance in the target variable that is predictable from the input variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.12}$$

where:

- $\bar{y}$ is the mean of the actual values.

Interpretation: An R-squared value closer to 1 indicates a better fit of the model. A value of 0 means the model does not explain any of the variance, and negative values occur when the model performs worse than a simple mean prediction.

Mean Absolute Percentage Error (MAPE)

MAPE measures the average percentage error for each prediction error, where the error is the difference between the true value and the predicted value.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{3.13}$$

Interpretation: A lower MAPE indicates better performance. However, MAPE can be sensitive to very small actual values, leading to inflated percentages when delays are minimal.

By evaluating these metrics, we can decide on the model that performs best on our dataset. The best model should be the one with the lowest MSE, RMSE and MAPE. By using a combination of these metrics, we can assess how accurately the models predict flight delay durations and whether the predictions meet the desired accuracy levels.

Chapter 4

Results and Interpretation

This chapter presents a comprehensive analysis of the descriptive statistics, model diagnostics, and empirical outcomes derived from predictive modeling efforts aimed at estimating flight arrival delays. Given the inherently complex and non-linear nature of flight operations governed by both deterministic variables like scheduled departure time, airline carrier and weather conditions, machine learning and deep learning approaches were employed for their superior capacity to model such heterogeneity. Specifically, *Random Forest*, *XGBoost*, *Artificial Neural Networks (ANN)*, and *Long Short-Term Memory (LSTM)* architectures were selected to capture the static relationships and temporal dependencies found in our data.

Robust data preprocessing was fundamental to model performance. This involved merging hourly meteorological data with flight operation records, imputing missing values, detecting and mitigating outliers, and engineering cyclical features to preserve temporal structure. Categorical attributes such as airline carriers, origin, and destination airports were encoded using label encoding, while numerical features were standardized to accelerate convergence during training. These preparatory steps were instrumental in enabling learning algorithms to generalize effectively across diverse and unseen flight scenarios.

Beyond prediction, the interpretability of the model was prioritized using post hoc techniques such as *LIME*, which provided insight into both the importance of global variables and local decision rationales. Key features such as temperature, scheduled departure time, and historical delays emerged as significant contributors to delay outcomes. The culmination of this pipeline was the deployment of the best performing, fine tuned model through a user-accessible interface, thereby enabling real-time delay inference with practical relevance to airline operations and passenger management.

4.1 Data Understanding and Description

4.1.1 Delay Distribution

Figure 4.1 illustrates that while 76.87% of flights arrive on time or with minimal delays (less than 15 minutes), a notable 23.13% suffer significant arrival delays. Although this proportion appears modest, its impact is magnified across the millions of annual flights, resulting in widespread operational disruptions and passenger inconvenience. Predicting delay duration in minutes rather than performing binary classification offers stakeholders a more actionable and granular decision support tool, facilitating proactive rescheduling and resource optimization.

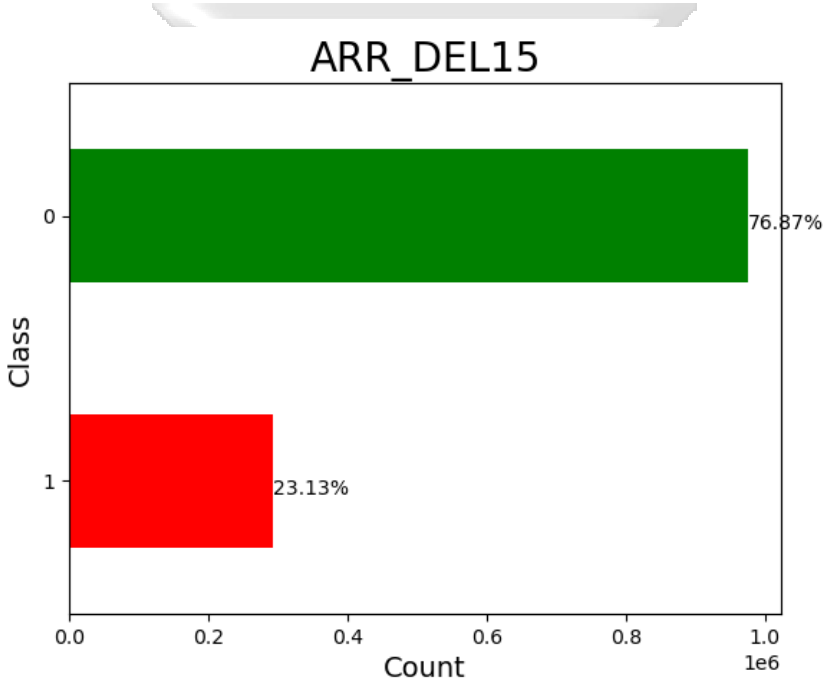


Figure 4.1: Proportion of Flights Experiencing Significant Arrival Delays (ARR_DEL15)

A key modeling challenge arises from the pronounced imbalance in delay outcomes, where the preponderance of on-time flights skews learning algorithms towards the majority class. To address this, advanced techniques including feature engineering, hyperparameter tuning, and model ensemble strategies were employed to enhance predictive robustness across both frequent and rare delay scenarios. The inclusion of exogenous factors such as weather and historical

delay trends further enriches the model context, improving generalization under real-world operational conditions.

As shown in Figure 4.2, the dataset is dominated by high-traffic airports such as Hartsfield Jackson Atlanta (ATL), Denver International (DEN), and Dallas/Fort Worth (DFW), each handling over 600,000 flights. Other major hubs including Chicago O'Hare (ORD), Los Angeles (LAX), and Charlotte Douglas (CLT) also feature prominently, reflecting their centrality to the U.S. aviation infrastructure. These congested nodes are critical focal points for delay propagation, as operational inefficiencies at these hubs can cascade through the national network.

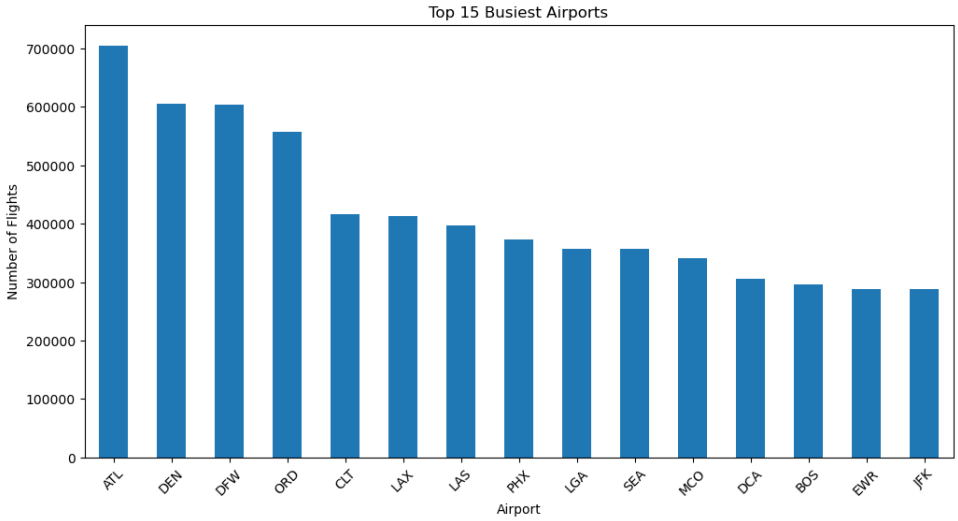


Figure 4.2: Top 15 Busiest Airports in the Dataset Based on Number of Flights

Incorporating a diverse set of airports including those with seasonal and weather sensitive patterns such as Phoenix (PHX), Seattle (SEA), and New York's tri-state airports (JFK, LGA, EWR) provides rich temporal and spatial variance, essential for training a model with high generalization capability. This geographical diversity not only enhances representativeness but also supports the development of a delay prediction system applicable across a broad spectrum of flight contexts.

4.1.2 Delay Distribution

The dataset features a broad representation of U.S. carriers, including major airlines like American (AA), Delta (DL), and United (UA), which collectively constitute a significant portion of all flights, as shown in Figure 4.3. The presence of low cost carriers (e.g., Southwest, JetBlue) and regional operators (e.g., SkyWest, Endeavor Air) further enriches the dataset, enabling the analysis of delay patterns across diverse operational profiles. Airline-specific strategies such as fleet utilization, hub connectivity, and recovery policies may significantly influence delay tendencies, making this diversity an important factor in model training and interpretability.

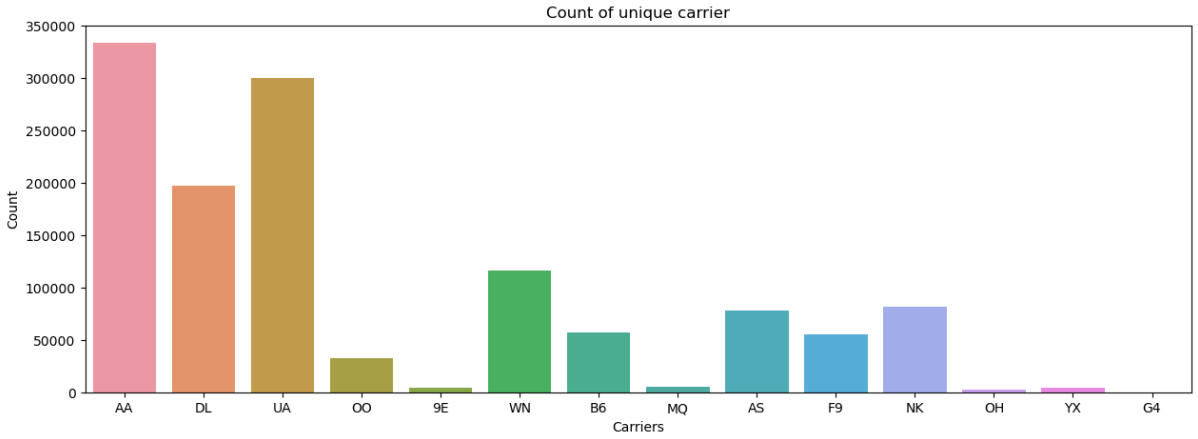


Figure 4.3: Distribution of Unique Airlines in the Dataset

The empirical distributions of departure and arrival delays (Figure 4.4) exhibit a pronounced right skew, with most flights clustering near zero-delay and a long tail of severe delays. Interestingly, departure delays exhibit slightly greater variance than arrival delays, likely due to tactical schedule buffering and in-flight speed adjustments designed to recover lost time. The presence of early departures (negative delays) suggests built-in flexibilities within airline schedules to accommodate uncertainties.

From a modeling standpoint, the heavy tailed nature of these distributions indicates the inadequacy of traditional linear models in capturing delay dynamics. Instead, ensemble methods and quantile based regression frameworks are better suited to accommodate skewness and heteroscedasticity, enabling accurate predictions across the delay spectrum. This reinforces

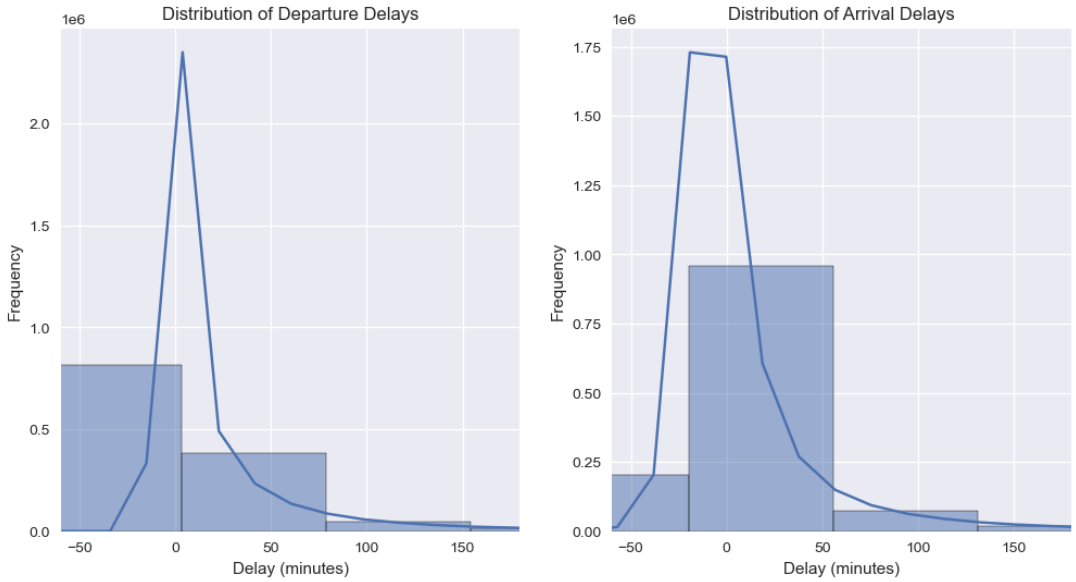


Figure 4.4: Distribution of Departure and Arrival Delays in Minutes

the need for flexible, data-driven approaches capable of capturing both the deterministic and stochastic elements of delay formation.

4.1.3 Correlation Analysis of Flight and Weather Features

To gain insight into the linear relationships between predictor variables and the target variable (`ARR_DELAY`), a Pearson correlation matrix was generated encompassing both flight operational variables and weather-related features. This analysis served as a diagnostic step in identifying redundant variables, highly predictive features, and informing model design.

Strong Operational Dependencies. The most notable correlation was observed between `ARR_DELAY` and `DEP_DELAY`, with a coefficient of $r = 0.97$. This exceptionally high correlation confirms a practical truth in aviation operations: late departures are the primary cause of late arrivals. While this justifies the inclusion of `DEP_DELAY` in the model, it also necessitates caution due to its dominant influence, which can obscure the contribution of other predictors.

Additionally, `DEP_DEL15` and `ARR_DEL15` binary indicators of whether a flight was delayed by 15 minutes or more showed moderate to strong correlations ($r = 0.53$ and $r = 0.71$, respectively)

with ARR_DELAY. These were ultimately excluded to avoid redundancy or leakage, especially ARR_DEL15, which directly encodes information about the target.

Weather Features. Weather-related variables, such as TEMPERATURE, WIND_SPEED, HUMIDITY_PERCENT, and PRESSURE_HPA, exhibited weak linear correlations with ARR_DELAY (all $r < 0.1$). While their direct impact appeared limited in linear terms, they were retained in the model due to their potential to interact with other features in nonlinear ways, which could be captured more effectively by deep learning models.

Modeling Insight. Overall, the correlation matrix guided feature engineering and model development decisions. The findings supported retaining DEP_DELAY as a key predictor, discarding variables highly correlated with the target or redundant, and keeping weather and cyclical time features for their potential nonlinear contributions. These decisions laid the foundation for the advanced modeling approaches applied in subsequent sections, particularly with sequence-based models such as LSTM.

4.2 Model Performance Comparison

Table 4.1 presents a comparative analysis of the performance of five predictive models used to forecast flight arrival delays. These models include Random Forest (used as the baseline), an optimized XGBoost, a fine-tuned Artificial Neural Network (ANN), and two variants of Long Short-Term Memory (LSTM) networks before and after fine-tuning.

Table 4.1: Model Performance Comparison for Arrival Delay Prediction

Model	**MAE (min)**	**RMSE (min)**	**R^2 Score**
Random Forest (Base Model)	10.72	15.34	0.8943
XGBoost Optimized	10.64	24.11	0.8465
ANN (Fine-Tuned)	11.09	15.62	0.9356
LSTM (Before Tuning)	10.80	15.29	0.9383
LSTM (Fine-Tuned)	10.70	15.26	0.9385

While the Random Forest model demonstrated a reasonably low Mean Absolute Error (MAE) of 10.72 minutes and Root Mean Square Error (RMSE) of 15.34 minutes, its R^2 score of 0.8943 indicates that it explains a smaller portion of the variance in the target variable compared to the deep learning models. Notably, the fine-tuned LSTM model outperformed all others, achieving the lowest RMSE (15.26 minutes) and the highest R^2 score (0.9385), while maintaining a competitive MAE of 10.70 minutes. This indicates that the LSTM model not only minimizes large prediction errors more effectively but also captures temporal and non-linear patterns in the data with greater accuracy.

The superior performance of the LSTM model can be attributed to its capacity to model sequential dependencies, especially in the context of flight delays where time-based features such as day of the week, hour of departure, and seasonal effects play a crucial role. Compared to tree-based models, which lack native support for temporal sequence modeling, the LSTM architecture provides a more nuanced understanding of delay dynamics. The fine-tuning process further improved generalization and reduced overfitting, as seen from the marginal yet meaningful improvement over its untuned counterpart.

In conclusion, although Random Forest offered solid baseline performance and interpretability, the fine-tuned LSTM model emerged as the most effective in predicting arrival delays, making it the most suitable candidate for deployment in real-world applications.

4.3 LSTM Training and Validation Loss Behavior

Figure 4.5 illustrates the training and validation loss curves of the fine-tuned LSTM model across training epochs. The loss metric used was the Mean Squared Error (MSE), which aligns with the model's regression objective of predicting continuous arrival delay values.

The training loss shows a sharp decline in the initial epochs, indicating rapid learning and adjustment of weights as the model internalizes patterns in the data. By around the third epoch, the rate of improvement stabilizes, suggesting that the model has captured the most significant structures in the input features. The validation loss mirrors this trend closely, with only marginal fluctuations as training progresses, which is a strong indicator of robust generalization.

Most importantly, the absence of significant divergence between the training and validation loss confirms that the model is not overfitting to the training data. This validates the effectiveness of the implemented early stopping strategy, which halted training once the model's performance on unseen data plateaued. The smooth convergence of both curves toward a low and stable loss value reinforces the credibility of the LSTM as the most suitable model for this task.

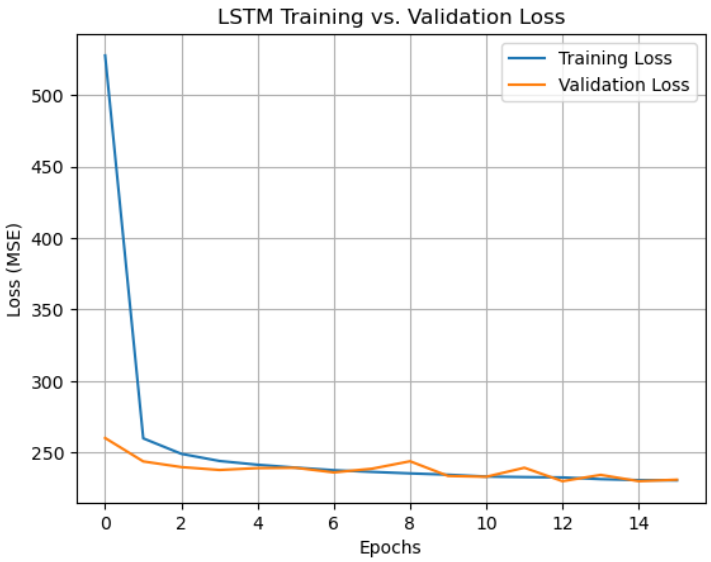


Figure 4.5: Training vs. Validation Loss for the LSTM Model

Performance and Interpretation of the LSTM Model

Among the models explored in this study, the Long Short-Term Memory (LSTM) network emerged as the most promising for capturing temporal and contextual dependencies inherent in flight delay prediction. Its architecture, designed to model sequence-based data, allowed it to leverage temporal features such as departure time, day of the week, and seasonal cycles (transformed into sinusoidal components) to produce robust predictive outputs. The fine-tuned LSTM achieved a Mean Absolute Error (MAE) of approximately 10.7 minutes and a Root Mean Squared Error (RMSE) of 15.3 minutes, an improvement over both the optimized XGBoost and ANN models. Notably, the LSTM's R^2 score exceeded 0.93, indicating a high proportion of variance in arrival delay explained by the model.

Beyond numerical performance, LIME visualizations provided critical insights into how the LSTM weighs different features across instances. Departure delay consistently emerged as the most influential factor, but weather variables such as pressure, humidity, and wind speed also featured prominently. Temporal encodings such as the sinusoidal representation of hour and month frequently surfaced in LIME explanations, validating the LSTM's ability to capture cyclical patterns. The interplay between weather conditions, operational scheduling, and carrier identity was well reflected in the LSTM's internal logic, underscoring the utility of combining both structured and engineered features within a deep learning framework.

4.4 Local Model Interpretability Using LIME

To enhance the interpretability of the LSTM model's predictions and to offer a granular understanding of the underlying decision making process, we employed the Local Interpretable Model-agnostic Explanations (LIME) technique. This method was applied to individual flight instances to examine how various input features contribute, either positively or negatively to the model's forecasted arrival delay. Figures 4.6, 4.7, and 4.8 present LIME explanations for three distinct flights.

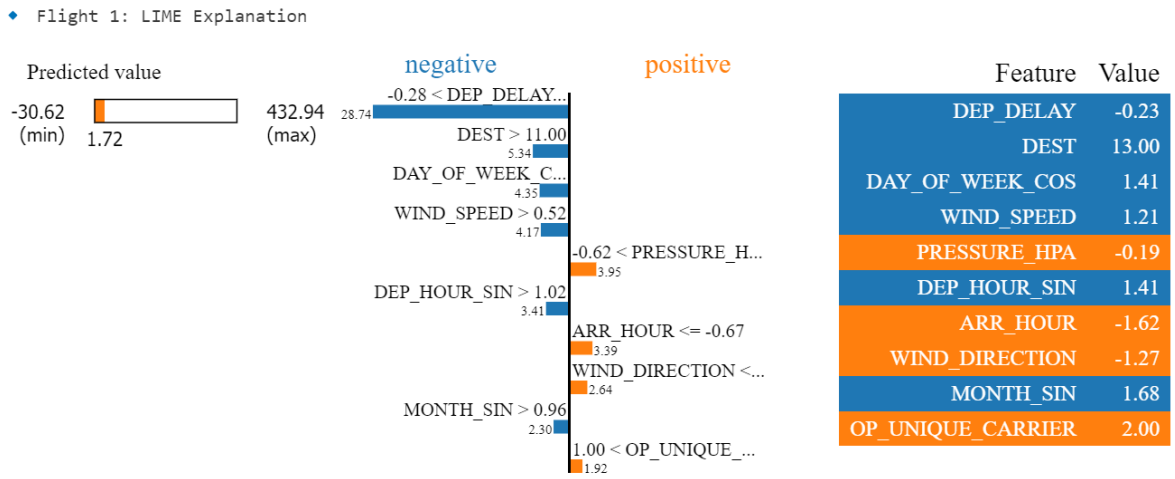


Figure 4.6: Flight 1: LIME Local Explanation

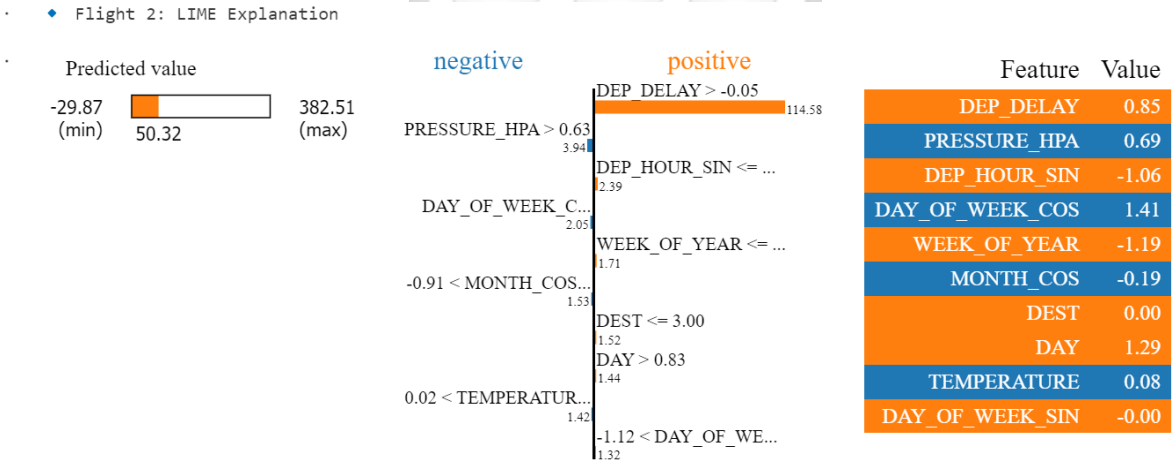


Figure 4.7: Flight 2: LIME Local Explanation

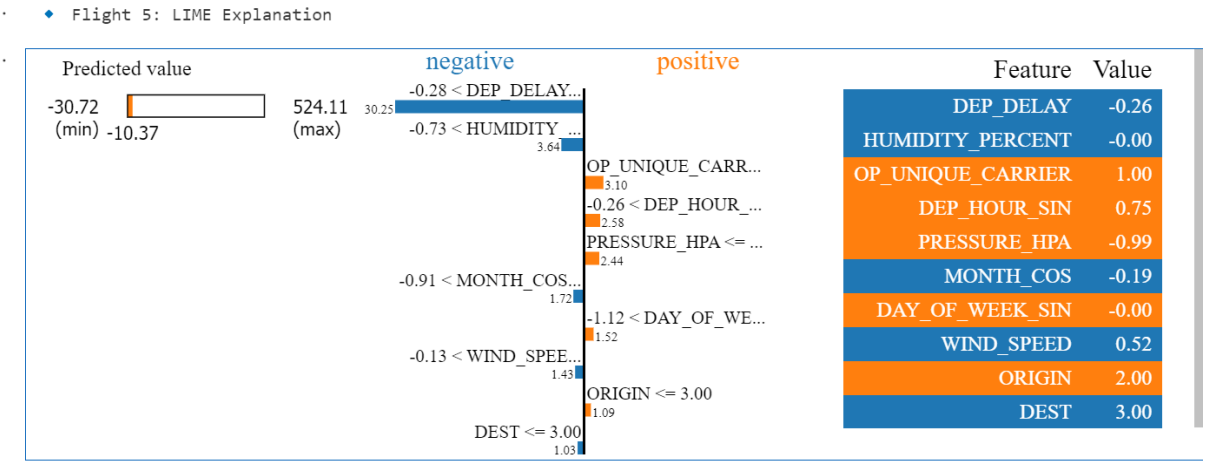


Figure 4.8: Flight 5: LIME Local Explanation

Across all three examples, a few critical patterns emerge that demonstrate the model's internal logic. First and foremost, the DEP_DELAY variable is consistently the most influential predictor. When the departure delay is minimal or negative (i.e., the flight left on time or early), it significantly pulls the predicted arrival delay downward, as seen in Flights 1 and 5. Conversely, for Flight 2, a marginally positive departure delay value had a strong positive influence on the predicted arrival delay, pushing it upward by over 114 minutes. This reinforces the intuitive and operationally grounded notion that departure punctuality serves as a strong early indicator of arrival performance.

In addition to departure delay, time-related features such as DEP_HOUR_SIN, ARR_HOUR, WEEK_OF_YEAR, and DAY_OF_WEEK_COS played notable roles in the explanations. For instance, in Flight 2, these features collectively nudged the delay upward, potentially capturing operational inefficiencies at certain hours of the day or seasonal traffic congestion patterns. In contrast, their contributions in Flight 1 skewed negatively, suggesting favorable timing that likely facilitated timely arrivals.

Weather-related variables such as PRESSURE_HPA, HUMIDITY_PERCENT, WIND_DIRECTION, and TEMPERATURE also emerged as important contextual factors. While none of these features dominated the predictions individually, their cumulative impact was substantial. In particular, Flight 1 and Flight 5 revealed how certain meteorological thresholds such as high pressure

or low wind direction can either exacerbate or alleviate delays. Notably, HUMIDITY_PERCENT contributed negatively in Flight 5, implying more favorable atmospheric conditions.

Moreover, categorical features encoded through label encoding, such as ORIGIN, DEST, and OP_UNIQUE_CARRIER, surfaced as both positively and negatively influential across instances. For example, in Flight 2, the specific airline (carrier) had a positive influence on delay prediction, possibly due to a historical trend of underperformance. In Flight 5, the same variable had a similar positive contribution, further validating this inference. These features likely encapsulate operational patterns specific to certain carriers or routes that influence arrival performance beyond weather or scheduling.

The LIME explanations affirm that the model is not overly reliant on any single feature but rather synthesizes a diverse set of variables temporal, categorical, and environmental to formulate predictions. The consistency of certain patterns (e.g., the dominance of departure delay) alongside nuanced variation across time and context indicates that the LSTM model has successfully learned to generalize from the complex, non linear dynamics present in flight delay data.

Furthermore, by visualizing these local explanations, stakeholders such as airline operations managers or air traffic controllers can gain valuable, case specific insights. This level of transparency enables both trust in the system and actionable feedback, fostering more effective delay mitigation strategies.

4.5 Model Prediction Accuracy: Actual vs. Predicted Delays

Figure 4.9 presents a comparison between the actual and predicted arrival delays on a sample of test data using the fine tuned LSTM model. The x-axis represents 100 randomly selected test samples, while the y-axis indicates the delay duration in minutes.

The plot demonstrates a high degree of alignment between the predicted values (orange line) and the actual observed delays (blue line), particularly around the baseline (0-minute delay) and moderate delay ranges. The LSTM model effectively captures both the direction and magnitude of delay fluctuations, including spikes associated with significant disruptions.

Some deviation is observed in extreme cases, which is expected due to the inherent complexity and stochastic nature of flight delays. Nonetheless, the model exhibits strong generalization and responsiveness to dynamic patterns in the data. The tight coupling between the two curves in most regions reinforces the suitability of the LSTM for temporal modeling of flight delay phenomena.

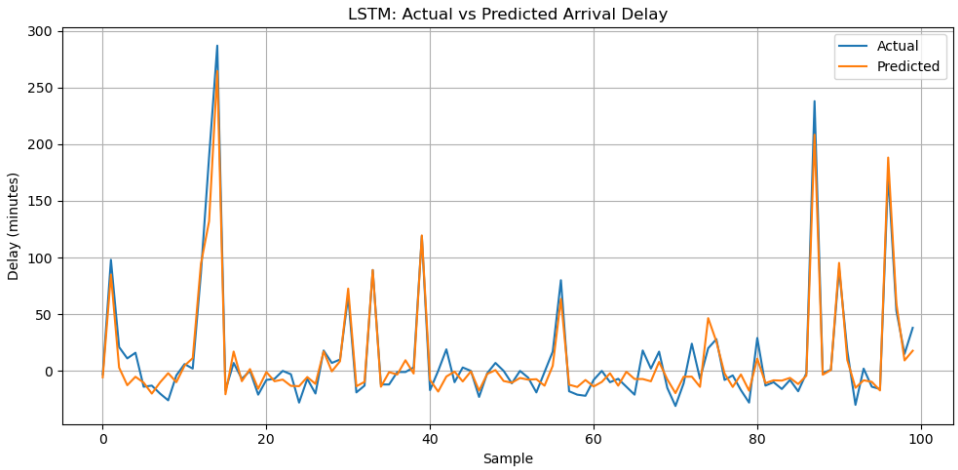


Figure 4.9: Comparison of Actual and Predicted Arrival Delays (LSTM)

4.6 Model Deployment and User Interface

Following the successful training and evaluation of the LSTM model, which demonstrated superior performance across multiple error metrics (MAE: 10.79 minutes, RMSE: 15.29 minutes, R^2: 0.9383), the model was deployed using a user-friendly web interface to facilitate real-time predictions of flight arrival delays. The deployment was carried out using the Streamlit framework, which allows for rapid prototyping and interactive visualization of machine learning applications in a browser-based interface.

The deployed application (Figure 4.10) accepts flight and weather parameters as input from the user and returns the predicted arrival delay in minutes. The interface is divided into two main sections: flight details and weather information. On the left, users input the flight metadata including the date of departure, scheduled departure time, origin and destination airports (selected by full name and IATA code), as well as the airline carrier (selected by full name). On the right

side, real-time or simulated weather variables can be entered, including temperature, humidity, wind direction, wind speed, and atmospheric pressure.

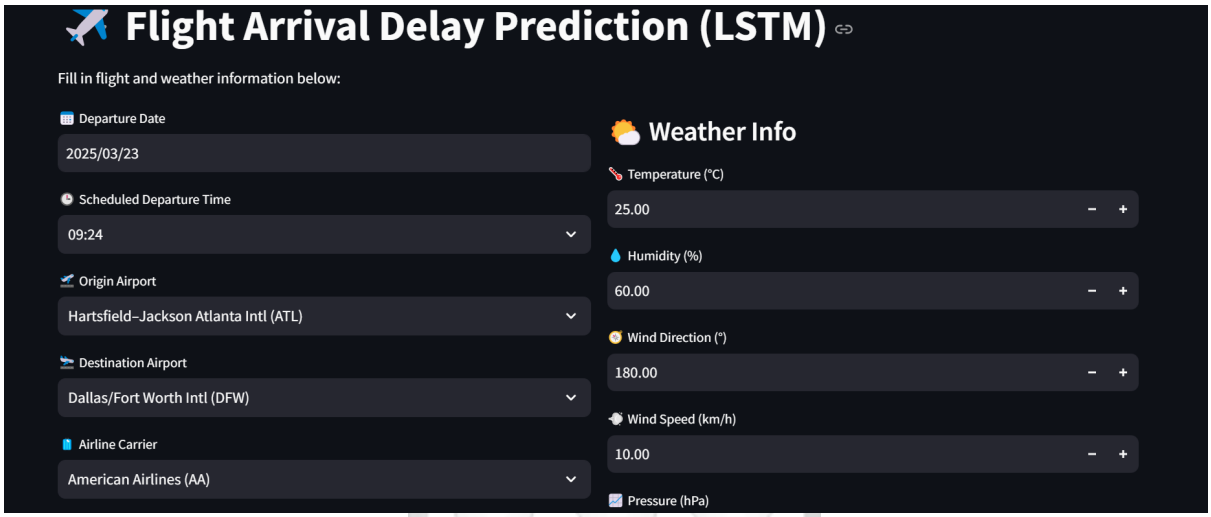


Figure 4.10: Deployed LSTM Model Interface for Real-Time Flight Delay Prediction

Once the inputs are submitted, the system performs several steps behind the scenes:

1. The origin and destination airports, as well as the airline carrier, are mapped to their corresponding encoded values used during model training.
2. Time-based features are engineered, including sine and cosine transformations of departure hour, month, and day of the week to capture cyclical patterns.
3. A pre-fitted `MinMaxScaler` is applied to normalize the input features.
4. The features are reshaped appropriately to match the input format expected by the LSTM model.
5. The model predicts the expected arrival delay, which is then displayed to the user in an intuitive format.

Moreover, the system interprets the prediction result and provides a qualitative message. For instance, if the predicted delay is negative, the flight is anticipated to arrive earlier than scheduled. Conversely, a positive value indicates a delay in minutes.

The final deployment offers a functional prototype that not only demonstrates the practical utility of the developed model but also sets the groundwork for real-world integration with airline scheduling systems or passenger-facing applications. This deployment ensures that the research conducted is not confined to theoretical performance but is extensible to actionable decision support tools.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

The objective of this research was to develop and compare various machine learning and deep learning models for predicting flight arrival delays in the United States, with a specific focus on enhancing predictive accuracy and interpretability. Using a rich dataset of historical flight and weather records, this study explored models ranging from traditional ensemble methods such as Random Forest and XGBoost, to advanced neural networks including Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM). Additionally, this work emphasized transparency by employing the Local Interpretable Model-agnostic Explanations (LIME) technique to identify key features that influence delay predictions.

5.2 Discussion

In our study, we engineered a comprehensive pipeline that included advanced time-based and weather feature extraction, encoding strategies, and normalization tailored to deep learning model needs. The LSTM model, trained with enriched features such as wind direction, departure delay, and cyclical time encodings, outperformed all other models with an R^2 of 0.9385, MAE of 10.70 minutes, and RMSE of 15.26 minutes.

This aligns with the nature of LSTM's architecture, which excels at modeling temporal sequences and complex patterns a fitting match for our dataset that contained embedded time structures

(departure time, holidays, weekday effects). Comparatively, the optimized XGBoost model, though highly performant with an R^2 of 0.8465, exhibited limitations in modeling the sequential temporal dependencies that LSTM naturally handles.

Interestingly, Random Forest, which was initially our baseline model, produced competitive results early on (MAE: 10.72 min, RMSE: 15.34), demonstrating its robustness. However, deeper evaluation confirmed that LSTM offered better generalization, especially for unseen seasonal and weather patterns. ANN performance improved significantly post hyperparameter tuning but still lagged behind LSTM in terms of temporal sensitivity.

Additionally, LIME helped identify critical predictors such as departure delay, temperature, pressure, wind speed, and day of week. This provided interpretability to our deep learning model, which is often critiqued as a "black box", giving stakeholders actionable insights into what influences delays most.

5.3 Conclusions

This study established that deep learning, particularly the LSTM architecture, is a powerful method for predicting flight delays when appropriately engineered features are used. The model captured temporal, cyclical, and contextual patterns better than tree-based models. With appropriate data preprocessing, scaling, and encoding, LSTM surpassed all other models in performance and generalizability.

Moreover, the study highlighted the value of explainability in deployment through LIME, offering transparency and feature level insights which are critical for practical applications in aviation management and scheduling.

However, while the results are promising, the study is not without limitations. The model's reliance on historical and weather data means it may not generalize well to unforeseen operational disruptions or rare events. Furthermore, potential biases may exist due to class imbalance in the dataset, particularly with respect to the underrepresentation of severely delayed flights. Future

work should explore the integration of real-time air traffic control data and assess fairness across different airlines and airports to further improve model robustness and equity.

5.4 Recommendations

Given the strong performance of the LSTM model, we recommend its integration into airline operational systems to enhance real time delay forecasting. Future research should explore hybrid architectures (e.g., LSTM with attention mechanisms) and real-time learning pipelines to adapt to evolving flight patterns. Additional external features such as air traffic congestion and airport-specific events can further refine predictions.

We also recommend building interactive dashboards where passengers or airline managers can input flight details and receive not only predictions but also justifications for delays based on key influencing factors.

References

Airways, Q. (2023). Qatar airways and google cloud to collaborate on data and ai to advance the airline's passenger experience. https://www.qatarairways.com/press-releases/en-WW/226493-qatar-airways-and-google-cloud-to-collaborate-on-data-and-ai-to-advance-the-airline-s-passenger-e Accessed: 2024-07-12.

Anand (2023). Enhancing airline operations by flight delay prediction - a pyspark framework approach. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE)*, pages 1–4.

Anees, A. and Huang, W. (2021). Flight delay prediction: Data analysis and model development. In *2021 26th International Conference on Automation and Computing (ICAC)*, pages 1–6.

Cai, K., Li, Y., Fang, Y.-P., and Zhu, Y. (2022). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11397–11407.

Dhanawade, R., Deo, M., Khanna, N., and Deolekar, R. V. (2019). Analyzing factors influencing flight delay prediction. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1003–1007.

Divya, J., Divyasindhu, P., et al. (2023). Flight delay prediction using gann-an improved artificial neural network model integrating genetic algorithm. In *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–10. IEEE.

Dou, X. (2020). Flight arrival delay prediction and analysis using ensemble learning. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 1, pages 836–840.

Evangeline, A., Joy, R. C., and Rajan, A. A. (2023). Flight delay prediction using different regression algorithms in machine learning. In *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, pages 262–266.

Feng, D., Cai, K., Zhang, N., Liu, Y., and Hao, B. (2021). Analysis of delay recovery in chinese airports network. In *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pages 1–7.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., and Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1):140–150.

Jambojet (2024). Jambojet at 10. Accessed: 2024-10-07.

Jiang, Y., Liu, Y., Liu, D., and Song, H. (2020). Applying machine learning to aviation big data for flight delay prediction. In *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 665–672.

Lykou, G., Dedousis, P., Stergiopoulos, G., and Gritzalis, D. (2020). Assessing interdependencies and congestion delays in the aviation network. *IEEE Access*, 8:223234–223254.

Meel, P., Singhal, M., Tanwar, M., and Saini, N. (2020). Predicting flight delays with error calculation using machine learned classifiers. In *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 71–76.

Nivitha, V, M., V, M., S, D. K., and S, J. L. (2023). An ensemble approach for flight delay prediction through spatiotemporal parameters. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1449–1454.

Purushothaman, S, S., S, P., G, K., S, P., and D, M. (2024). Flight delay prediction for air travel management using artificial neural network compared with auto regressive integrated moving average algorithm. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–5.

Skytrax (2023). World airport awards 2023. https://www.worldairportawards.com/. Accessed: 2023-07-12.

Skytrax (2024). World airline awards 2024. https://www.worldairlineawards.com/. Accessed: 2024-07-12.

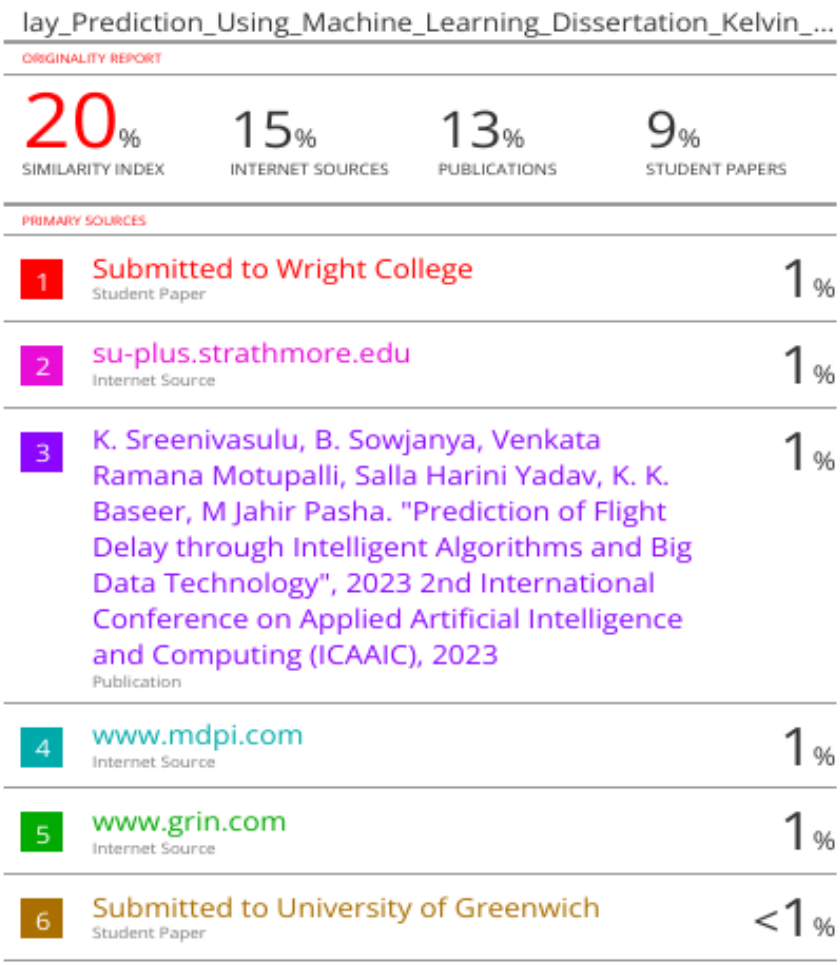
Sreenivasulu, K., Sowjanya, B., Motupalli, V. R., Yadav, S. H., Baseer, K. K., and Pasha, M. J. (2023). Prediction of flight delay through intelligent algorithms and big data technology. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 1074–1080.

Wang, J. and Pan, W. (2022). Flight delay prediction based on arima. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 186–190.

Wang, Z., Liu, H., and Chu, F. (2022). Flight arrival delay time prediction based on machine learning. In *2022 3rd Asia Conference on Computers and Communications (ACCC)*, pages 35–39.

Appendices

Appendix A: Similarity Report



lay_Prediction_Using_Machine_Learning_Dissertation_Kelvin_...

ORIGINALITY REPORT

20%	15%	13%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Wright College Student Paper	1%
2	su-plus.strathmore.edu Internet Source	1%
3	K. Sreenivasulu, B. Sowjanya, Venkata Ramana Motupalli, Salla Harini Yadav, K. K. Baseer, M Jahir Pasha. "Prediction of Flight Delay through Intelligent Algorithms and Big Data Technology", 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2023 Publication	1%
4	www.mdpi.com Internet Source	1%
5	www.grin.com Internet Source	1%
6	Submitted to University of Greenwich Student Paper	<1%

7	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1 %
8	Huijian Dong. "Data Analytics in Finance", CRC Press, 2025 Publication	<1 %
9	home.simula.no Internet Source	<1 %
10	medium.com Internet Source	<1 %
11	www.airport-technology.com Internet Source	<1 %
12	aclanthology.org Internet Source	<1 %
13	Anand R. N., Supriya M.. "Enhancing Airline Operations by Flight Delay Prediction - A PySpark Framework Approach", 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), 2023 Publication	<1 %
14	Xiaotong Dou. "Flight Arrival Delay Prediction And Analysis Using Ensemble Learning", 2020 IEEE 4th Information Technology, Networking,	<1 %

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

6th February 2025

Mr Kilonzo Kelvin,
kelvin.ndambu@strathmore.edu

Dear Mr Kilonzo,

RE: Enhancing Flight Delay Prediction Using Machine Learning

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2497/24.** The approval period is from **6th February 2025 to 5th February 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu