TU Delft

Delft University of Technology

Forecasting of Airline En Route Delay for Individual Flights with Supervised Learning

Dolman, C.; Ribeiro, M.J.; Sun, Junzi; Lothaller, P.R.J.R.; de Wilde, Jasper; Piva, Alexander; Vossen, F.A.K.

Publication date
2025

Document Version
Final published version

Published in
First US-Europe Air Transportation Research and Development Symposium (ATRDS2025)

Citation (APA)
Dolman, C., Ribeiro, M. J., Sun, J., Lothaller, P. R. J. R., de Wilde, J., Piva, A., & Vossen, F. A. K. (2025). Forecasting of Airline En Route Delay for Individual Flights with Supervised Learning. In *First US-Europe Air Transportation Research and Development Symposium (ATRDS2025)*

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

First US-Europe Air Transportation Research and Development Symposium (ATRDS2025)

Forecasting of Airline En Route Delay for Individual Flights with Supervised Learning

Casper Dolman, Marta Ribeiro, Junzi Sun, Phillipe Lothaller
Delft University Technology
Delft, Netherlands

Jasper de Wilde, Alexander Piva, Frans Vossen
KLM Royal Dutch Airlines
Netherlands

***Abstract*—Air traffic delays have a major impact on the aviation industry, affecting airlines, passengers, and the broader ecosystem. With increasing regulatory and sustainability pressures, accurate delay predictions are critical as they allow for precise determination of the contingency and discretionary fuel required for flights. This research aims to develop an explainable supervised learning model to improve existing en route delay predictions, focusing on intercontinental flights from North America to Amsterdam Schiphol Airport. While prior studies have explored flight delay prediction, they have not addressed two critical research gaps identified in this research: the inclusion of day-of-operations features, such as passenger information, aircraft weights, and cost index, and the use of transatlantic flight data for predictions 90 minutes before departure. To address these gaps, two Gradient-Boosted models, CatBoost and LightGBM, were trained using internal airline, airport, and METAR data. Both models outperformed the airline's current in-use statistical model, with CatBoost achieving an MAE of 3.44 minutes and RMSE of 4.61 minutes and LightGBM achieving an MAE of 3.43 minutes and RMSE of 4.56 minutes. The most significant performance increase over the current model was observed under adverse weather conditions. This research advances en route delay prediction by providing more accurate delay forecasts, particularly in critical weather conditions, and proposes practical improvements to support future studies focused on enhancing model adaptability across diverse operational contexts.**

***Keywords*—En Route Delay; Airline Operations; Fuel Planning; Gradient-Boosting; Supervised Learning**

I. Introduction

Air traffic delays, particularly en route delays, represent a persistent and costly challenge for the global aviation industry, affecting airlines, passengers, and the broader ecosystem daily. According to Eurocontrol, arrival and departure punctuality have decreased by almost 7% since 2019, with arrival punctuality at 65% and departure punctuality at 58% [1]. In July 2024, air traffic flow management delays in Europe almost reached 7M minutes, averaging 6.5 minutes per flight, an increase of 64% compared to the same period in 2023. These elevated delay levels were driven by a combination of factors: limited network capacity, adverse weather, constrained rerouting options due to the Ukraine conflict, increased military activity, and major technical disruptions [2].

Numerous factors can influence the actual flight time of a flight, ranging from air traffic control instructions and weather conditions to airspace congestion and safety-related rerouting [2]. Accurate forecasts of en route delays are essential for improving operational efficiency and planning accuracy. Better predictions of expected en route delay can often reduce the contingency and discretionary fuel required on board [3, 4]. This lower fuel load reduces the aircraft's weight, leading to a more optimised flight and lower fuel consumption throughout the entire journey. Thus, better predictions can lead to optimised fuel management and improved flight planning, which are critical for cost reduction and achieving sustainability, safety, and customer satisfaction objectives.

In regular airline operations, flight dispatchers generate the final flight plan approximately 90 minutes before departure. Currently, dispatchers rely on simple statistical models that average delays from previous flights, overlooking flight-to-flight variations in weather, congestion, and other critical parameters. Having a more accurate estimate of en route flight delay at this stage would allow dispatchers and pilots to make more informed decisions about fuel loading, ensuring that sufficient fuel is carried to accommodate possible time recovery during the flight or to manage holding patterns, rerouting, and other in-flight adjustments due to unforeseen circumstances. This precision could directly translate into cost savings through optimised fuel usage while reducing the airline's environmental footprint.

This study advances prior research on en route flight delay prediction by leveraging proprietary operational data from a European Airline to develop a supervised machine-learning model explicitly tailored to flights originating from North America and landing at Amsterdam Schiphol Airport. The model incorporates a novel set of features, including detailed flight data such as Flight Planning Software (FPS) flight times, expected congestion, and en route weather conditions. These additional features aim to reduce prediction errors and improve the accuracy of en route delay forecasts.

The remainder of this research paper is structured as follows. The existing literature on flight delay prediction and research gaps is discussed in section II. Next, in section III, the problem at hand is discussed in more detail, highlighting the importance of model explainability. After this, the methodology used during this study is presented in section IV. The results are given in section V. In section VI, the model validation process is presented. The findings and resulting recommendations are discussed in more detail in section VII. Finally, the conclusion is given in section VIII.

II. Literature Review

Early studies on flight delay prediction used statistical and basic regression-based models to analyse delay patterns. Mueller and Chatterji [5] found that Poisson distributions effectively modelled departure delays, while Normal distributions suited en route and arrival delays. Tu et al. [6] improved predictions with a non-parametric model incorporating seasonal trends and daily patterns. However these approaches often lacked the adaptability to handle complex, real-time conditions in aviation delay forecasting. As the complexity of delay factors grew and data volumes increased, the aviation industry began transitioning toward machine learning methods, which offer the flexibility and adaptability needed to capture dynamic, real-time conditions.

A. Machine Learning in Aviation

The data-rich nature of aviation makes it ideal for machine learning applications, with recent developments in algorithms, the availability of vast datasets, and powerful computational resources driving the rapid adoption of these models. Advanced machine learning algorithms like Gradient Boosting and Neural Networks are now commonly applied by researchers such as Dalmau et al. [7] and Zhu and Li [3], showing significantly improved performance in delay prediction over traditional statistical methods. This transition is further accelerated by increased funding and interest from both public and private sectors [8, 9].

1) Flight Delay Prediction using Regression Approaches: Multiple regression approaches have been compared to each other, showing that Random Forest models outperform simple Linear models and Decision Trees, as they leverage multiple decision trees to capture complex patterns and interactions within the data [10, 11]. Initial regression approaches for flight delay prediction include Rebollo [12], who achieved their results using aggregate pre-flight variables to select appropriate forecast horizons. Kalliguddi [11] improved prediction accuracy by employing Random Forest, which outperformed other models due to its ability to handle complex relationships in the data, achieving a Root Mean Squared Error (RMSE) of 12.5 min. Manna et al. [13] achieved lower RMSEs by using Gradient-Boosted Decision Trees, which enhanced prediction accuracy by effectively handling data variations and Boosting techniques. They achieved an RMSE of 10.7 min for arrival delay and 8.2 min for departure delay.

Thiagarajan [14] and Ayhan [15] conducted comparative studies that delved deeper into the impact of model selection and data completeness on prediction outcomes. Thiagarajan et al. [14] used a two-stage model with Extra-Trees and Random Forest to predict on-time performance, demonstrating that these models outperformed Boosting techniques, particularly when the dataset was limited to flight and weather data. The model achieved an RMSE of 8.3 min with global training and 5.1 min for selective training on single origin-destination pairs for domestic flights in the USA. In contrast, Ayhan et al. [15] focused on predicting the Estimated Time of Arrival (ETA) using a more comprehensive dataset with trajectory information. Their work showed that Boosting methods like AdaBoost and Gradient Boosting delivered superior accuracy, highlighting the critical role of data richness in enhancing model performance, achieving an RMSE between 2.8 and 4 min for different domestic flight routes in Spain with a forecast moment before departure.

Achenbach [16] and Birolini [10] explored hybrid and ensemble models to enhance delay prediction. Achenbach [16] combined Gradient Boosting and Linear Regression to optimize flight arrival predictions and fuel consumption. Their Gradient-Boosted Linear Regression model achieved an RMSE of 5.9 min at the block-off moment. Similarly, Birolini and Jacquillat [10] used segmented Random Forest and XGBoost models for day-ahead routing and delay mitigation. Their XGboost model achieved an RMSE of 7.2 min one day before the flight for flights in Europe.

In recent years, Neural Networks have gained attention for their ability to model complex, nonlinear relationships in flight delay prediction. Silvestre [17] followed this trend by applying a Long Short-Term Memory network to predict in-flight estimated arrival times, using 4D trajectory and weather data, achieving an RMSE of 3.6 min with prediction moment at 100 NM from the arrival airport. Similarly, Yu [18] employed a Deep Belief Network combined with Support Vector Regression to predict flight delays, outperforming several other traditional models. These studies underscore the effectiveness of Neural Networks in handling complex, dynamic data to improve prediction accuracy. Their model achieved a Mean Absolute Error (MAE) of 8.4 min with a forecast moment 2 hours before the flight.

2) Flight Time and Estimated Time of Arrival Prediction: Zhu [19] were among the first to investigate en route flight time prediction. They used machine learning techniques to model traffic volume and convective weather effects and improve flight time predictions. Despite the limited dataset, the study provided valuable insights into the importance of considering en route variables and suggested that incorporating additional data, such as aircraft-specific information and detailed flight plans, could further enhance prediction accuracy. LightGBM, XGBoost, and Random Forest achieved an RMSE of 7.1, 7.2, and 7.2 min, respectively, with the forecast moment being before the flight.

Zhu [3] introduced a Spatial Weighted Recurrent Neural Network to predict actual flight times to optimize fuel consumption. Their model utilised data from Automatic Dependent Surveillance-Broadcast (ADS-B) systems, Meteorological Aerodrome Reports (METAR), and airline operational records. Their model achieved an RMSE of 7.55 min, with the forecast moment before the flight. Moreover, their study demonstrated significant operational benefits, showing that optimised flight time predictions could save fuel by 0.016%-1.915% without compromising safety.

Wang [20] had an automated data-driven framework for predicting ETA on the runway at the entry point of the Terminal Manoeuvring Area (TMA) using ADS-B data from Beijing Capital International Airport. The framework achieved improved accuracy by clustering flights by runway-in-use and applying a stacked ensemble model. Their Gradient Boosted Machine performed best at an RMSE of 87.3 seconds.

B. Research Gaps and Contribution of this Paper

There is a critical gap in considering detailed day-of-operations data, such as flight plan data, for flight time prediction: these features include flight data, cost index, and operational and business requirements. These day-of-operations attributes directly impact flight time, and their inclusion in the feature set is anticipated to enhance the performance of forecasting models.

Additionally, typical prediction horizons for ETA or flight time are just before or during the flight, where uncertainty is relatively low due to real-time data inputs [15, 17, 20]. This shorter prediction horizon works well for domestic or short international flights but becomes more challenging for longer intercontinental flights. For these flights, predictions must account for conditions that will be encountered along the route, where data uncertainties significantly increase as the pre-departure time extends. Notably, no prior research has addressed en route delay prediction for transatlantic flights 90 minutes before departure. This 90-minute pre-departure window is critical, as it is typically when the final flight plan, including the final fuel allocation, is released. Accurate predictions of potential departure or en route delays at this stage would allow for timely adjustments.

III. Problem Statement

This research is conducted in collaboration with an internal airline, which operates flights to over 170 destinations worldwide. We focus on transatlantic flights from North America to Amsterdam Schiphol Airport. The choice to focus on North America was made after evaluating other routes: flights to Asia were excluded due to disruptions in routes caused by unrest in the Middle East and Russia, flights to Africa were limited by the few available destinations, and European flights were too short to capture the en route delay patterns of interest. Between North and South America, North America was chosen due to a broader range of destinations and the absence of triangular flights. The dataset comprises flight data from 2018 to 2024, excluding years affected by the COVID-19 pandemic, resulting in almost 45,000 flights.

The trip time given by Flight Planning Software (FPS) is used to calculate the en route delay. Flight Planning Software is a comprehensive flight planning tool airlines use to optimize flight routes, manage fuel efficiency, and ensure regulatory compliance. En route delay is considered to be a deviation from the trip time provided by FPS and the actual flight time. This differs from previous studies on flight delays, where the scheduled trip times are often the comparison or baseline. These standardised schedules do not account for daily variations in route or weather and instead reflect average or typical times for a route.

Note that the en route delay value has been corrected for the actual cost index that has been flown. This means the en route delay caused by the difference between the planned and actual cost indexes has been corrected. This is a correction for the difference in speed as this was not considered en route delay. This is achieved using a standardised table that gives the difference in flight time in minutes for different cost indices for a specific route. A higher cost index than planned would mean the flight is going faster than planned, and time should be added to find the corrected FPS flight time. By correcting this, the model does not have to identify patterns in expected speed differences.

The airline currently uses a statistical model that outputs an expected value of en route delay. This estimate is then shared with the dispatcher, who can use it to make more informed flight and fuel planning decisions. Together, this can be used to anticipate any expected en route delays. The performance improvement of the new models with respect to the currently in-use statistical model will be shown in more detail in subsection V-C. The current in-use model by the airline will be referred to as the Current model in the remainder of this paper.

IV. Methodology

This section will discuss the methodology used during this research. First, the data preprocessing steps are presented in subsection IV-A. After this, the model development procedure is discussed in subsection IV-B. Finally, the feature engineering steps are laid out in subsection IV-C.

A. Data Preprocessing

Three data sources are used to build the features for the en route delay prediction model as described below. Through the collaboration with a European airline, detailed airline data, including flight plan data, is available. Historical actual weather measurements are available from the Iowa State University Environmental Mesonet database. Finally, the Opensky Network is used as a database to gather ADS-B messages from flights all over the world.

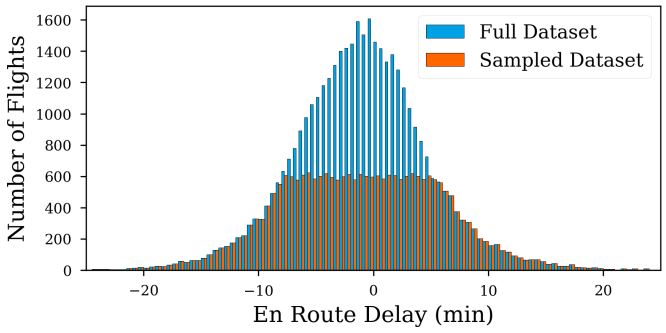
- Airline Data: Data with respect to flight plans, flight information, passenger information, en route weather. After removing flights with missing information, almost 45,000 flights are considered for training and testing.
- Meteorological Aerodrome Report (METAR) Data: METAR weather data available from Iowa State University Environmental Mesonet. The weather reports are used for weather features at arrival airports [21].
- The OpenSky Network ADS-B Messages Data: Data used to construct congestion features at outstation airports (not used in final use case as only Schiphol Amsterdam Airport was analysed) [22].

1) Missing Data Handling: Two steps are taken to deal with missing data in the dataset. Flights that lack essential data used to build the features, like scheduled or actual times or other data inputs, are removed from the dataset. The weather data from the METAR dataset occasionally has empty cells for rows used in the model. For these missing weather data points, linear interpolation was used to estimate the weather situation at the recorded time. If a gap of more than two hours in the measurements existed, the flight was removed from the dataset to prevent the chance of a significant difference between the interpolated and actual values.

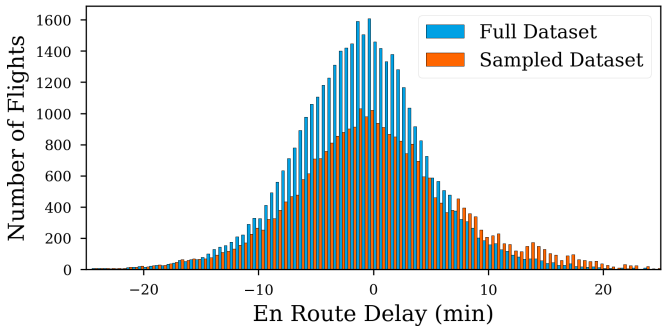
2) Data Sampling and Outlier Removal: The distribution is centred just below zero and over 90% of the flights experience within -10 and 10 minutes of en route delay. By training the model on the entire dataset, the model gets biased towards

this majority class. As a result, flights from the majority class (-5, 5) are removed so that the model is not overtrained on these instances. The final distribution can be seen in Figure 1a. This sampling method effectively increases the 'weight' of flights outside this range so the model can better recognise patterns in these cases. This suggests that the model's overall performance may decline as the performance for the majority class reduces. However, it will likely improve performance for flights outside this range. This primary method of removing part of the majority class of the data distribution is used in the results from subsection V-A to subsection V-B.

However, from an operational standpoint, there is particular interest in flights experiencing over 5 minutes of en route delay, as better predictions in this area are crucial for safety and planning, enabling more proactive responses to potential disruptions. A second sampling technique was thus proposed to achieve this, focusing more on >5 minutes en route delayed flights. For the middle part of the distribution, a total number of flights is specified instead of a maximum number of flights per bin; this causes the original distribution of more flights around 0 to stay intact. In Figure 1b, the resulting distribution is visualised. As expected by sampling in this way, the performance on the right-hand side of the distribution is further increased compared to the first sampling method previously described. In subsection V-C, the second sampling method will be used to show how the model performance can be further increased for the right-hand side of the distribution.



(a) En route delay distribution after removing samples from the majority class



(b) En route delay distribution after removing samples from the majority class and resampling in >5 minutes region

Figure 1: Visual of the resulting distributions of en route delay for both sampling methods

For training of the model, the dataset will be split into a training dataset (90% of the data) and a test dataset (10% of the data). Finally, the model performance improved when all flights with more than four standard deviations from the average were removed. This resulted in 151 flights being removed from the dataset. Such led to some erroneous entries in the flight data were removed as some flights had delay values of -40 and 40+, which are unlikely to be achievable.

B. Model Selection and Development

We use two variations of the Gradient-Boosted Decision Tree algorithm, Catboost and LightGBM, to predict flight delays. Numerous studies show that Boosted algorithms perform better than more simplistic models like Linear Regression, Decision Trees, and Random Forest [14, 15, 17]. There are two prominent reasons for choosing Boosted methods over Neural Networks. First, the explainability and interpretability of a model are very important. Since this study was conducted in collaboration with an airline, Boosted algorithms were chosen to increase explainability and understandability for end users like dispatchers and pilots. Boosted methods, such as CatBoost and LightGBM, are generally considered more explainable than Neural Networks, as their decision-making processes are more straightforward to interpret through feature importance and Decision Trees [23, 24]. The second reason is the dataset size - Neural Networks need extensive datasets to achieve high performance, which is hard to achieve by only considering the intercontinental flights of one airline.

CatBoost is a type of Gradient-Boosting Decision Tree specifically designed to work well with categorical input features. Instead of one-hot encoding or label encoding, it uses a technique called order target encoding. In this way, the dimensionality of the dataset is reduced without losing critical information about the categories, increasing efficiency and performance. It uses ordered boosting to prevent data leakage, a problem that can occur for traditional Gradient-Boosting methods [25]. The hyperparameter values used for the final Catboost model can be found in Table Ia.

LightGBM, just like Catboost, is a specific variation of the Gradient-Boosting Decision Tree algorithm. It is widely known for its fast and efficient implementation, particularly with large datasets. By using several optimisations, it becomes faster and more memory efficient compared to a standard Gradient-Boosted Decision Tree. The leaf-wise growth approach that LightGBM uses is the prominent driver of this. Instead of growing each tree level by level, splitting all nodes at each depth, LightGBM expands the tree by choosing the largest loss reduction leaf [26]. LightGBM provides more adjustable hyperparameters than CatBoost, allowing for greater flexibility and precision in model optimization. The final hyperparameter values can be found in Table Ib.

C. Feature Engineering

1) Selected Features for Final Model: Table II presents the features used in the final model. These features were selected using the RFE algorithm and correlation analysis. The features encompass a range of categories, including

temporal factors, flight plan details, weather conditions, and congestion indicators.

2) Feature Correlation of Numerical Features: In Figure 2, the correlation matrix for all numeric variables in the dataset can be found. None of the features have a really high correlation. A moderately strong correlation exists between the Wind Gust and Wind Speed - higher wind speeds result in high gusting. However, the Wind Gust feature was found to be especially critical in adverse weather situations, and information was added for the model compared to using only the Wind Speed feature. Therefore, it was decided that both features should be kept in the model. Additionally, a moderate correlation can be observed between the Arrival Hour and the Number of Arrivals features. Amsterdam Schiphol Airport works with arrival and departure banks, causing this relationship to appear. Finally, it can also be seen that there is a moderate negative correlation between the Average Wind Component and the FPS Flight Time Standard Deviations - a higher Average Wind Component (tailwind) will result in shorter flight times.

3) Feature Uncertainty due to Data Limitations: During the training of the model, some limitations in data availability were encountered, which affect the predictions and performance of the model. The first limitation encountered during the model training is that not all flight plan versions are saved for a long time. This means that during the model's training, only the latest issued flight plan is available, which is not necessarily the flight plan that was available 90 minutes before departure. In total, 85% of the flight plans used were 90 minutes before, and 96% of the flight plans were created at least 60 minutes before departure. The fact that flight plans later than 90 minutes are also used introduces a bias in the model. During the actual use of the model, only the flight plan 90 minutes before departure is available. This means that in some cases, this flight plan does not have the latest update on the route or weather forecast, causing a worse representation of the actual flight. This means that the model's performance, when tested on new data, will be slightly worse as the data will not be as accurate as the training data.

The second limitation concerns the weather data used for Schiphol. For training, actual measurements from the Iowa State University Meteorological Aerodrome Report database are used. However, 90 minutes before departure, only Terminal Aerodrome Forecasts (TAFs) are available. Given the long flight durations from North America (up to 10 hours), TAFs for Schiphol may be issued well in advance, leading to discrepancies between forecasted and actual conditions. This is especially problematic in adverse weather, with wind speeds over 15 knots or visibility below 3 kilometres.

TABLE I. Hyperparameters of the selected methods.

(a) CatBoost

Hyperparameter	Value
iterations	1300
learning_rate	0.03
depth	9
subsample	0.7
L2_leaf_reg	5
objective	RMSE

(b) LightGBM

Hyperparameter	Value
n_estimators	1000
learning_rate	0.02
num_leaves	50
max_depth	20
min_child_samples	10
subsample	0.8
colsample_bytree	0.6
reg_alpha	0.5
reg_lambda	0.0
objective	MSE

TABLE II. Selected features after elimination

Feature Name	Unit	Type	Example
Hour of Day	[-]	Numeric	12
Week of Year	[-]	Numeric	24
Season	[-]	Categorical	S22
Departure Airport	[-]	Categorical	JFK
Aircraft Type	[-]	Categorical	789
Average Wind Component	[kts]	Numeric	65
Cost Index	[-]	Numeric	200
Planned Arrival Runway	[-]	Categorical	18R
Wind Speed	[kts]	Numeric	18
Wind Direction (cos)	[-]	Numeric	0.68
Wind Direction (sin)	[-]	Numeric	-0.69
Wind Gust	[kts]	Numeric	40
Visibility	[km]	Numeric	6
Number of Arrivals	[-]	Numeric	18
FPS Flight Time Standard Deviations	[-]	Numeric	2

V. Results

In this section, the performance will be analysed globally in subsection V-A and for the binned distribution in subsection V-B. Finally, in subsection V-C, the developed models will be compared to the currently in-use model.

A. Global Model Performance

Below in Table III, performance metrics are given for four models. The table includes a Random Forest model to show the superior performance of Boosted methods. But also the Current model introduced in section III, to show the performance increase over the currently in use model. The Catboost model and the LightGBM model outperform the Current model and the Random Forest model. Although the fit of all models can still be considered low, the Catboost and LightGBM models perform better than the Current model. The Boosted methods also perform slightly better regarding the MAE and RMSE. The higher difference in RMSE for the Current model compared to the other models is explained by the fact that the model does not understand outliers well, as it is a simple statistical model that uses medians. However, this means that the Current model predicts flights closer to the middle of the distribution more accurately.

The LightGBM model required less than 20 seconds to train and produce results, while the CatBoost model took around 2.5 minutes. This efficiency makes both models suitable for operational use. Notably, these results were achieved on a 13th Gen Intel(R) Core(TM) i5-1345U CPU with integrated Intel(R) UHD Graphics.

TABLE III. Global performance of the models

Model	MAE [min]	RMSE [min]	R2
Random Forest	3.68	4.91	0.17
Catboost	3.44	4.61	0.23
LightGBM	3.43	4.56	0.22
Current	3.69	5.23	0.11

Figure 2: Feature correlation for the numerical features used in the models

B. Binned Model Performance

Figure 3 shows that the MAE increases when moving to the sides of the distribution. It is also visible that the number of flights drops significantly in a similar fashion to the training dataset. In the plot, the performance of the four models is visible again. It can be seen that the Current model has the best performance in the range of -5 to 2.5 minutes of en route delay. For most of these flights, nothing major happens, and the median is a good prediction. However, outside of this region, the machine-learning models perform better.

Catboost has the best performance of the machine learning models in the middle of the distribution but loses performance compared to the Random Forest and LightGBM models when moving to the right side of the distribution. Still, it performs better than the Random Forest for most of the left side of the distribution. The LightGBM model performs slightly worse than Catboost in the range of -5 to 7.5 but does outperform the Catboost and Random Forest models outside of this region. It significantly outperforms the current model when moving more to the outside.

C. Model Improvement Compared to the Current Model

The Current model is a simple statistical model that uses averages to predict en route delay, which means that it does not understand changes in day-to-day weather, congestion, and flight plans. This causes significant errors for more extreme cases of en route delay, both for negative and positive values. This means it is exciting to see the increase in performance of the Catboost and LightGBM models for those flights. A new model, Right Side LightGBM, is introduced to illustrate the potential for further improving performance on the right side of the distribution. The model uses the second way of sampling training data as discussed in subsubsection IV-A2. This is only shown for the LightGBM model as it had superior performance over the Catboost model on the right-hand side of the distribution as shown in subsection V-B.

In Figure 4a, three models are compared to the Current model: Right Side LightGBM, LightGBM, and Catboost. Only flights with an actual en route delay higher than 0 are considered in the figures. The left y-axis shows the percentage of flights correctly predicted above a specified en route delay threshold (in minutes). A flight is considered correctly predicted above the threshold when both the predicted and actual en route delay are higher than the specified delay value on the x-axis. For instance, 14.1% of flights in the test set (consisting of 1,655 flights from North America) experience en route delays above 6 minutes. The Right Side LightGBM model correctly predicts delays above 6 minutes for 40% of these flights. The relaxed bounds provide further context, showing that for the same 14.1% of flights with delays over 6 minutes, the model predicts delays above 5 minutes in 47% of cases. While the regular LightGBM and CatBoost models perform slightly less accurately, they still significantly outperform the Current model.

The model understands certain conditions better than other conditions. If a flight has experienced an en route delay of more than 0 minutes and one or more of the following conditions below is true: (1) Wind Speed: Wind speed over 15 [kts], (2) Wind Gust: Wind gust over 25 [kts], and (3) Visibility: Visibility under 3 [km], the flights are included in Figure 4b. In the training data, 16% of the flights meet one or more of these conditions. The increase in performance is visible in Figure 4b as now 67% of flights are correctly identified as above 6 minutes (48% for LightGBM and 38% for Catboost Models), and for the relaxed bound -1, it is almost 80%. This shows that the increase in model performance is significant under certain specific conditions. From an operational standpoint, this also gives confidence in the model prediction for these conditions.

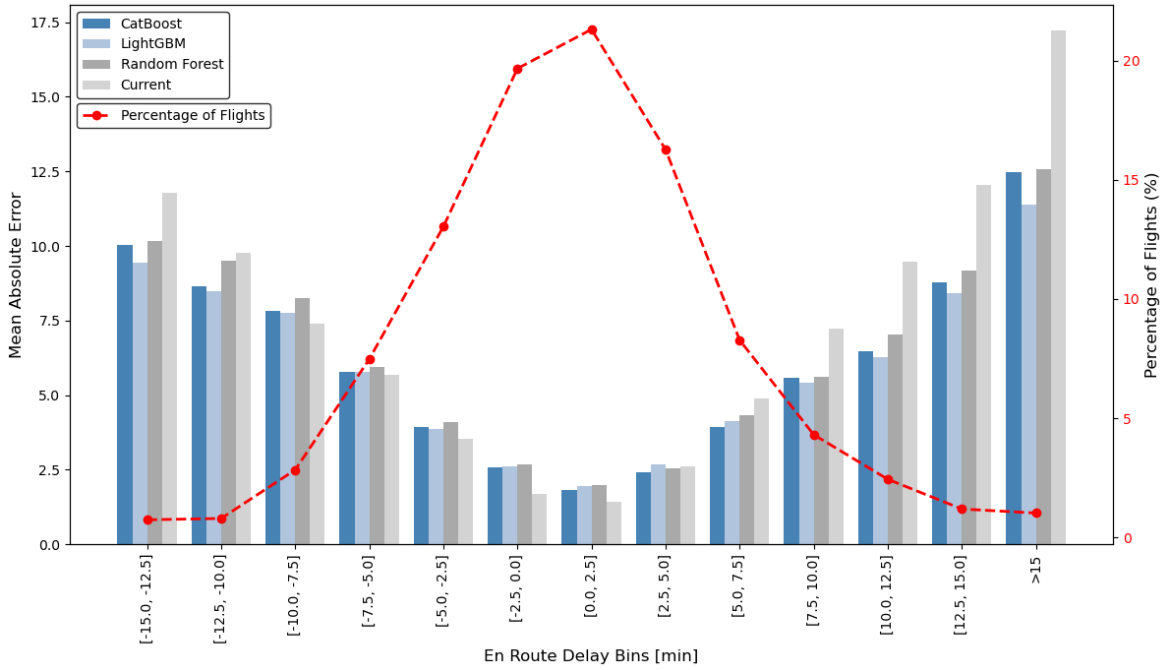
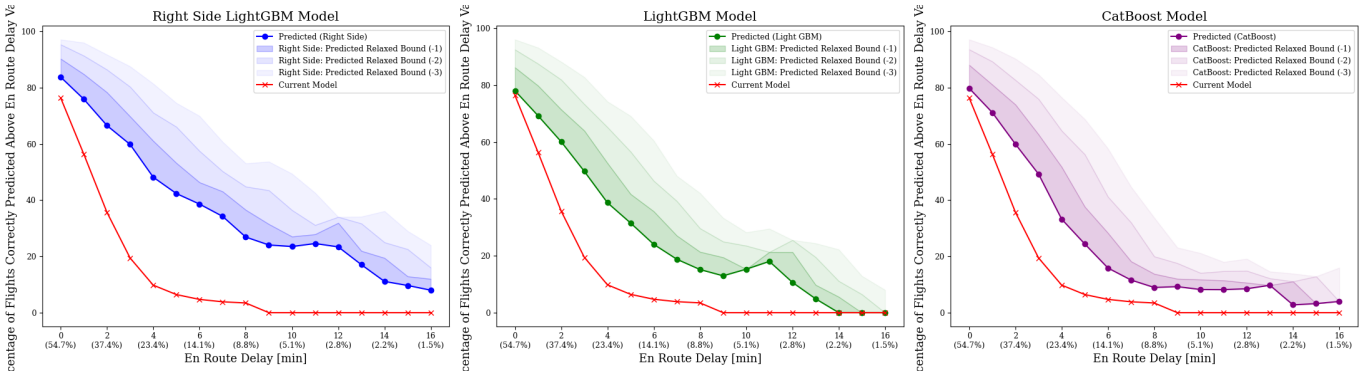
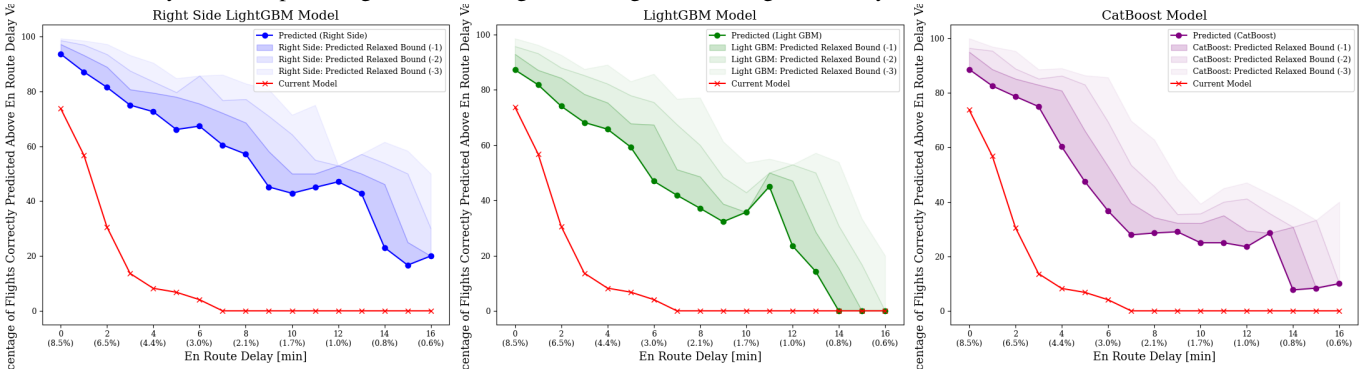


Figure 3: Binned model performance of the Catboost and LightGBM models compared to the Current model and Random Forest model



(a) Comparison of three models against the Current model in predicting flights exceeding specified en route delay values. The x-axis shows the en route delay, with the percentage of test set flights meeting or exceeding each delay value indicated in brackets



(b) Comparison of three models compared to the Current model in predicting flights exceeding specified en route delay values under specific weather conditions. The x-axis shows the en route delay, with the percentage of test set flights meeting or exceeding each delay value indicated in brackets

Figure 4: Comparison of models in predicting flights above specific bounds. (a) Full test data set, and (b) Under specific weather conditions

VI. Model Validation

In this section, the models that have been built are validated. In section VI, two sensitivity analyses are discussed. The test run performance is presented in subsection VI-B.

A. General Errors

For all ranges of delays often there is a difference between the actual wind experienced by the aircraft and the planned wind over the full range of the flight. This wind affects the total travel time as the ground speed can change significantly depending on the wind speed. This also means that this error can be significant for longer flights, like the flights analysed during this research from North America to Schiphol. A 1-knot difference in wind speed can translate into a 1-minute difference in flight time for flights of 3000-4000 nautical miles; the flights analysed in this research are all over 3000 nautical miles and go up to almost 6000 nautical miles.

Another error that has been observed frequently is the error caused by the difference in time spent from the Standard Terminal Arrival Route (STAR) entry point to the landing runway. All flights from North America enter via 1 of 4 entry points: LAMSO, REDFA, TOPPA, and MOLIX. An average time from each of these points towards every runway can be calculated from historical data. From this, it was clear that flights can have a difference of 3-4 minutes within the Terminal Manoeuvring Area (TMA) without having significant deviations from the route, such as a holding. This means that a large part of the error can be explained by a difference in time spent in this area. Flights with a positive prediction error (prediction longer than actual) often had short, straight paths to the runway. Whereas flights with a negative prediction error often had longer, curved paths to the runway. Another primary reason for the difference in time spent in the TMA is a change in the landing runway. The arrival runway differed from the planned arrival runway in about 35% of the flights.

Following, different parts of the en route delay distribution are discussed in more detail, and commonly found reasons for errors in the model prediction are given:

- Flights that experience an en route delay of -5 minutes or less typically encounter one or more of the following factors. First, they encountered a significantly better average wind factor along the route. Second, they have had a significant shortcut somewhere along the route. Or they landed on a runway different from the one planned in the flight plan.
- Flights that experienced en route delay of [-5, 5]: The error in this range appears random. Most of the flights in this delay range do not experience anything significant. No extensive shortcuts are given, and no significant events happen in the TMA. It is challenging for the model to explain precisely what happens. A minute and even a 5-minute difference on such long flights can stem from many reasons, including differences in en route wind, shortcuts, vectoring, and runway changes.
- Flights that experienced en route delay of [5, 15]: Most flights in this delay range experience several things that cause the en route delay. These flights spend, on average, a longer time in the TMA due to minor vectoring procedures or runway changes. The flights experience a worse tailwind than planned due to unplanned altitude changes. From 10 minutes en route delay and onwards, flights often experience at least one holding pattern.
- Almost every flight with an en route delay of >15 minutes has a holding pattern. The model has difficulty predicting whether or not a flight will experience a holding and the time spent in holding. For some very extreme conditions, like extreme gusting or very strong winds, the average delay is over 15 minutes. For these conditions, there is a big range of what delays are experienced. It is challenging for the model to predict what will happen on individual flights. Flights arriving within 20 minutes of each other on the same conditions might experience en route delay differences of up to 20 minutes, even in adverse weather circumstances.

B. Test Run Performance

A test run was performed to test performance on flights outside the training period. Flight data over 1 month was gathered, including weather forecast. For this test run, weather forecasts were used instead of actual measurements used during training. Flight plans created at least 90 minutes before departure were used - the median creation time of flight plans in the test set was 153 minutes before departure. During training, flight plans of a later stage were sometimes used due to the limitations in saved flight plan data. Additionally, since the test run considers a new operational period, potential seasonal variations could impact model performance. Limited familiarity with this new season may result in reduced model accuracy. The resulting performance is shown in Table IV.

The performance of the Catboost and LightGBM models using TAFs is similar to the performance discussed in subsection V-A. However, the R2 value is significantly lower when the TAF and 90-minute flight plan are used. As shown in subsubsection IV-C3, the actual weather can differ significantly from the TAFs. Furthermore, flight plans created within 90 minutes of departure were used during training with new routing and en route weather information. However, this effect is smaller than that of the actual weather.

As shown in subsection V-C, the LightGBM and Catboost models understand adverse weather situations better. This could also explain the additional drop in R2 performance compared to what was given in Table III, as the training dataset contains 16% flights that comply with the conditions given in subsection V-C. Whereas the test run set only has 3% of flights that comply with those conditions. As a result,

TABLE IV. Results of the shadow run

Model	**MAE [min]**	**RMSE [min]**	**R2**
Catboost TAF	3.43	4.65	0.05
Catboost Actual Weather	3.31	4.47	0.13
Catboost Actual Weather & Latest Flight Plan	3.24	4.37	0.14
LightGBM TAF	3.50	4.70	0.04
LightGBM Actual Weather	3.42	4.54	0.10
LightGBM Actual Weather & Latest Flight Plan	3.36	4.47	0.10

the test set is primarily composed of flights under non-critical conditions, where delays tend to be more variable and lack a consistent directional trend. By contrast, critical conditions provide a more evident directional tendency (towards delayed), enabling the model to generate more reliable forecasts.

The binned model performance of the test run is given in Figure 5. A similar response is visible as in Figure 3. The Catboost model outperforms the LightGBM model in the centre part of the distribution, and the LightGBM model performs better to the sides of the distribution except for flights in the range of -12.5 to -7.5 minutes of en route delay. Furthermore, it is visible that for almost every bin, the performance of the models improves as more actual data (actual weather and latest flight plan) is used.

Finally, the test set has six flights that experience an en route delay of more than 15 minutes. All these experienced arrival wind speeds of less than 13 knots and maximal visibility of 10+ kilometres. Also, no critical values of the Number of Arrivals or any other feature were visible. This again shows that it is often impossible to correctly predict all possible flight scenarios during the en route phase of a flight.

VII. Discussion and Recommendations

This section discusses a final, high-level overview of all the research outcomes in subsection VII-A. From this discussion, a set of recommendations follows that will be presented in more detail in subsection VII-B.

A. Discussion of Research Outcomes and Implications

While the MAE and RMSE indicate good performance across all models, the R2 score shows that there is still room for improvement. The Current model demonstrates a superior performance for flights that experience low en route delays as it uses averages to predict en route delay, and most flights experience low values of en route delay. However, the Catboost and LightGBM models, while outperforming the Current model for high-delay flights (more than 5 min) and faster flights (less than 10 min), struggle to achieve the same accuracy in more 'normal' cases (-5 to 5 min). The models aim to capture the broader variability in delays, introducing a range of delay values to better account for unexpected delays in typically 'non-critical' flights. This added variability can shift predictions away from the average, providing the flexibility to predict higher delays when necessary but potentially reducing accuracy for flights with minimal delays.

For many of the features, there is a range of values that can be considered 'non-critical'. Examples include lower wind speeds, good visibility, and fewer incoming flights at the arrival airport. In these conditions, delays tend to fall in the -5 to 5-minute range, though with outliers in both directions. Flights under such 'non-critical' conditions often display more random variations in en route delay, making precise predictions challenging. Variations in en route delay arise due to unmodelled factors, such as differences between planned and actual en route winds, vectoring and shortcuts in the TMA, or unexpected runway changes. This unpredictability under non-critical conditions complicates clear delay expectations and partially explains the relatively low R2 values, given that the majority of flights operate under these conditions.

The most significant performance increase compared to the Current model could be seen in specific adverse weather conditions. This understanding is absent in the Current model, which only adjusts for adverse weather over extended periods rather than on a flight-to-flight basis. This shows a possible initial use case for the newly developed models. The model can help improve the prediction of en route delay under the conditions identified in subsection V-C. Additionally, training the models to focus more on these conditions, as demonstrated with the second sampling technique used for the Right Side LightGBM model, offers the potential for further performance gains. By increasing the number of flights that experience these conditions, this can be improved even further. Currently, the models frequently underestimate the full extent of en route delay under critical conditions, as many non-critical flights tend to 'pull' predictions toward the average en route delay.

B. Recommendations for Model Improvement

The analyses performed during this research have highlighted several key areas that could further improve the models. First, while the sensitivity plots of both models for the Number of Arrivals feature demonstrated the expected increase in en route delay with higher arrival numbers, it only increases the prediction over a range of 3-3.5 minutes maximum. It is recommended to link it to an actual capacity factor that can be expected at arrival time. By incorporating the dynamics of landing and take-off procedures based on expected runway configurations, the model would better understand when arrival numbers become critical. This enhancement would improve the model's ability to understand varying capacities throughout the day, allowing it to predict en route delays under different conditions more effectively.

Additionally, it is recommended that a more refined feature or model be developed to improve the prediction of events within the TMA. Unpredictable delays within the TMA could be attributed to a notable portion of the model's prediction error per flight. By including data from the local Air Traffic Control authority, it might be possible to better predict runway changes and shortcut and vectoring procedures under certain circumstances from the Standard Terminal Arrival Route (STAR) entry point to the landing runway.

VIII. Conclusion

In an increasingly complex aviation environment, accurate flight time prediction is becoming essential for airlines. En route delays not only have a significant financial impact on the aviation industry and the broader ecosystem as a whole, but they also affect airspace congestion, safety, passenger satisfaction, and the achievement of regulatory and sustainability goals. This research aims to proposed Catboost and LightGBM models outperformed the currently in-use statistical model (Current model) by the European airline and a Random Forest model. The LightGBM model achieved a MAE of 3.43 minutes and an RMSE of 4.56 minutes, while the CatBoost model reached an MAE of 3.44 minutes and an RMSE of 4.61 minutes. The most significant performance improvement

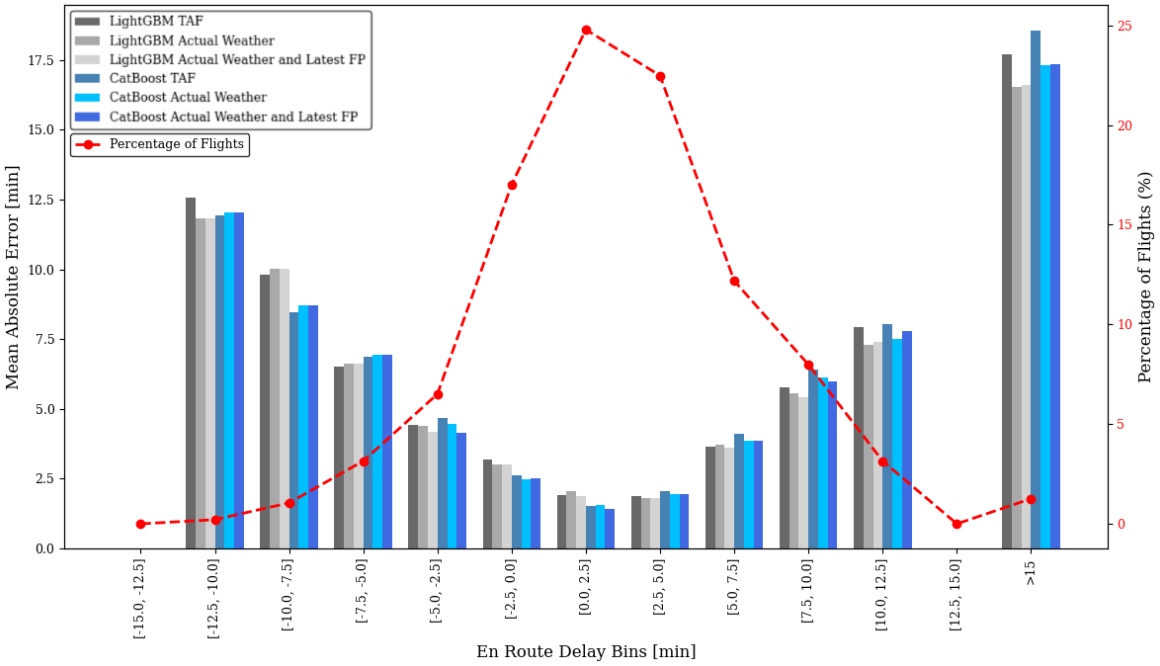


Figure 5: Binned model performance of the Catboost and LightGBM models during the test run

was observed in predicting delays during adverse weather conditions. The models understand how these adverse weather conditions reliably increase the average expected en route delay. Such enables the developed models to deliver more accurate predictions in critical weather conditions.

Future research will focus on better understanding the arrival airport's capacity at a given time. For this, a new congestion feature is proposed that considers capacity constraints connected to runway maintenance, adverse weather and other factors that influence the expected time in the TMA.

References

[1] EUROCONTROL. *European Aviation Overview, June 2024*. Accessed: 2024-10-29. 2024.

[2] EUROCONTROL. *July 2024 Overview of Network Performance*. Accessed: 2024-10-29. 2024.

[3] Xinting Zhu and Lishuai Li. "Flight time prediction for fuel loading decisions with a deep learning approach". In: *Transportation Research Part C* 128 (2021).

[4] Lei Kang and Mark Hansen. "Improving airline fuel efficiency via fuel burn prediction and uncertainty estimation". In: *Transportation Research Part C* 97 (2018).

[5] Eric Mueller and Gano Chatterji. "A nalysis of aircraft arrival and departure delay characteristics". In: *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*. AIAA, 2002.

[6] Yufeng Tu, Michael O Ball, and Wolfgang S Jank. "Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern". In: *Journal of the American Statistical Association* 103 (2008), pp. 112–125.

[7] Ramon Dalmau et al. "An explainable machine learning approach to improve take-off time predictions". In: *Journal of Air Transport Management* 95 (2021).

[8] Yirui Jiang, Trung Hieu Tran, and Leon Williams. "Machine learning and mixed reality for smart aviation: Applications and challenges". In: *Journal of Air Transport Management* 111 (2023).

[9] Michael Jordan and Tom Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349 (2015).

[10] Sebastian Birolini and Alexandre Jacquillat. "Day-ahead aircraft routing with data-driven primary delay predictions". In: *European Journal of Operational Research* 310 (2023).

[11] Anish M Kalliguddi and Aera K Leboulluec. "Predictive Modeling of Aircraft Flight Delay". In: *Universal Journal of Management* (2017), pp. 485–491.

[12] Juan Jose Rebollo and Hamsa Balakrishnan. "Characterization and prediction of air traffic delays". In: *Transportation Research Part C* 44 (2014), pp. 231–241.

[13] Suvojit Manna et al. "A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree". In: *International Conference on Computational Intelligence in Data Science*. IEEE, 2017.

[14] Balasubramanian Thiagarajan et al. "A machine learning approach for prediction of on-time performance of flights". In: *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. Institute of Electrical and Electronics Engineers Inc., Sept. 2017.

[15] Samet Ayhan, Pablo Costas, and Hanan Smet. "Predicting Estimated Time of Arrival for Commercial Flights". In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2018, pp. 33–42.

[16] Anna Achenbach and Stefan Spinler. "Prescriptive analytics in airline operations: Arrival time prediction and cost index optimization for short-haul flights". In: *Operations Research Perspectives* 5 (2018), pp. 265–279.

[17] Jorge Silvestre et al. "A deep learning-based approach for predicting in-flight estimated time of arrival". In: *The Journal of Supercomputing* (2024).

[18] Bin Yu et al. "Flight delay prediction for commercial air transport: A deep learning approach". In: *Transportation Research Part E* 125 (2019), pp. 203–221.

[19] Guodong Zhu et al. "En Route Flight Time Prediction Under Convective Weather Events". In: *Aviation Technology, Integration, and Operations Conference*. AIAA, 2018.

[20] Zhengyi Wang, Man Liang, and Daniel Delahaye. "Automated data-driven prediction on aircraft Estimated Time of Arrival". In: *Journal of Air Transport Management* 88 (2020).

[21] Iowa State University. *Iowa Environmental Mesonet: ASOS/AWOS Network METAR Data*. Accessed: 2024-10-03. 2024.

[22] OpenSky Network. *OpenSky Network: Free ADS-B and Mode S data for Research*. Accessed: 2024-07-03. 2024.

[23] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras & Tensorflow*. O'Reilly Media, 2017.

[24] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2nd ed. 2022.

[25] CatBoost. "Catboost Documentation ". In: (). Accessed: 2024-10-15.

[26] LightGBM. LightGBM Documentation. Accessed: 2024-10-15.