Science, Engineering and Technology Vol. 7, No. 1, pp. 61-75
www.setjournal.com https://doi.org/10.54327/set2027/v7.i1.489

A Machine Learning Framework for Predicting Airport Airside Key Performance Indicators and Analysing Statistical Relationships Across European Airports

Edvin Šimić[1], Muharem Šabić[1]

[1]*University of Sarajevo, Faculty of Traffic and Communications, Zmaja od Bosne 8, 71000 Sarajevo, Bosnia and Herzegovina*

Abstract

The objective of this study is to predict nine EUROCONTROL airport airside KPIs (key performance indicators) using historical monthly airport-level data and to identify statistical relationships between airport infrastructure, operational characteristics, and KPI performance using supervised machine learning algorithms with feature importance analysis across multiple European airports. Nine EUROCONTROL airside KPIs were selected as prediction targets, using monthly data from 76 European airports (2018–2019), comprising 1,824 observations per KPI after preprocessing. The input variables were airport airside features. Several supervised machine learning regression algorithms, including Multiple Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), were tested and evaluated for KPI prediction. Model performance was assessed using standard regression metrics, including the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). Model validation was performed using cross-validation and an independent test set to assess the generalisation performance of the models. To improve the interpretability of the predictive framework, feature importance analysis was performed to identify the relative contribution of individual airport characteristics to each predicted KPI (illustrated for the ATOT indicator). The results showed that the machine learning models achieved different levels of predictive performance across the nine KPIs, and that nonlinear ensemble methods generally provided stronger predictive performance than linear regression models. The proposed model highlights the potential of machine learning for predicting airport airside performance and analysing relationships between infrastructure and operational characteristics.

Keywords: Airport airside, Key Performance Indicators, Machine Learning, Feature importance, Airside Operations

1. Introduction

With the continuous growth in air traffic demand, many airports face increasing pressure on existing capacity, leading to potential congestion and operational inefficiencies. At the same time, infrastructure expansion at many European airports is often constrained due to spatial limitations and surrounding urban development, making large-scale capacity increases difficult to implement. As a result, understanding and evaluating airport performance based on existing operational data has become increasingly important for supporting data-driven analysis and planning. In this context, key performance indicators (KPIs) provide a structured way to monitor and compare airport airside performance across different operational conditions. Recent advances in machine learning and data analytics enable the development of models that can capture complex relationships between airport infrastructure characteristics and operational performance indicators. This study

Corresponding author: Edvin Šimić (edvin.simic@fsk.unsa.ba)
Received: 29 July 2026; Revised: 5 November 2026; Accepted: 12 November 2026; Published: 20 November 2026
© 2026 The Author(s). This work is licensed under a Creative Commons Attribution 4.0 International License

proposes a machine learning-based framework for predicting and analysing airport airside KPIs using infrastructure and operational features. The model is intended as a data-driven predictive tool for scenario analysis and performance estimation based on observed historical relationships. In addition, feature importance analysis is used to improve interpretability by identifying the relative contribution of input variables to each predicted KPI. The proposed approach supports exploratory analysis of airport performance and comparative assessment of airside characteristics without implying causal effects of specific interventions.

2. Literature review

A substantial proportion of research and analytical studies in this area is produced by international aviation organisations such as the International Civil Aviation Organisation (ICAO), European Union Aviation Safety Agency (EASA), and EUROCONTROL. However, in recent years, the academic community has become increasingly engaged in aviation research, primarily due to improved access to comprehensive aviation databases and operational datasets, enabling more extensive scientific analysis and innovation within the sector. This development is expected to further accelerate scientific research and innovation within the aviation sector. EUROCONTROL has established the Performance Review Unit (PRU), which is responsible for monitoring, analysing, and reporting performance [1]. In carrying out its responsibilities, the PRU collects data from various aviation stakeholders, including air navigation service providers, airspace user organisations, and the air transport industry. Through its web portal, the PRU also provides public access to network and airport performance data, including the key performance indicators (KPIs) that are used in this study to train machine learning algorithms.

The following section presents a brief overview of the existing literature in this field. One of the distinctive characteristics of the air transport system is the limited scientific foundation available for researching these complex problems, as well as the insufficient coordination between the academic community and the operational sector. The increased availability of large-scale aviation datasets in recent years has enabled greater involvement of the academic community in aviation research, facilitating more comprehensive analyses and the development of advanced data-driven methodologies. The availability of these datasets is expected to facilitate significant advances in the study of complex air transport systems in the near future. Since the model proposed in this study employs machine learning algorithms, the following section reviews the existing literature in which machine learning techniques have been applied to predict various aviation-related parameters, such as delays, taxi times, aircraft turnaround times, and other operational performance indicators.

Current research shows an increasingly pronounced application of machine learning methods in the analysis and prediction of airport performance. Authors in [2] applied machine learning regression models for the prediction of airport delays, highlighting their potential application to support the decision-making process. In papers [3], [4], authors focused on aircraft flight time on the maneuvering surface, where the considered models were based on the theory of waiting, as well as the importance of individual input variables.

Other research extends the application of machine learning to operational decision support. Article [5] analysed the operational benefits of an arrival management system using KPI indicators and radar trajectories, while authors in [6] applied machine learning methods to predict flight delays. Study [7] additionally investigated the predictive classification of airport performance and the influence of weather conditions on airport operations. Authors in [8] applied machine learning techniques to predict reactionary delays and other key performance indicators in the British railway network, demonstrating the potential of data-driven approaches for performance prediction in transportation systems. Most previous studies have focused on predicting specific indicators, including delays, taxi times, and flight operation distributions, using both regression algorithms [11 – 17] and classification algorithms [9], [10], [18 – 20] (Table 1). It was found that there have been no previous applications of machine learning algorithms for predicting airport airside key performance indicators (KPIs) using data collected by EUROCONTROL. Most existing studies that employ machine learning techniques are based on supervised

classification algorithms, where delay is typically treated as a binary variable, i.e., whether a delay will occur or not [9], [10], and regression methods, on the other hand, have been used to estimate continuous values, such as the numerical magnitude of turnaround time, taxi-out or taxi–in time [11], [12].

Although recent years have witnessed significant advances in models and tools for airport performance analysis, a large proportion of the existing research focuses only on fragmented parts of the airport system. These studies generally consider a limited number of input attributes and performance indicators, which restricts their ability to provide a comprehensive assessment of airport operational performance. Most existing studies focus on individual airports, while only a limited number of studies consider large datasets covering multiple airports. Table 1 presents an overview of studies that have applied machine learning methods for air traffic analysis, including the primary input data, their sources, the machine learning techniques employed, and the evaluation tests used to assess the accuracy of the algorithms. The findings indicate that classification and regression models represent the most frequently applied machine learning approaches, while ensemble-based methods, particularly Random Forest and boosting algorithms, generally demonstrate superior predictive performance compared with other techniques.

The mentioned studies are mainly focused on individual flights, specific operational processes, or individual factors affecting performance. Less attention was paid to the simultaneous modelling of a larger number of airside KPI indicators at the aggregated monthly level of the airport, including infrastructural, traffic, and operational characteristics. This represents space for further research in terms of more comprehensive use of machine learning for performance analysis. It can be concluded that previous studies have not explicitly proposed an approach that integrates machine learning algorithms with airport airside key performance indicators (KPIs) for the analysis and prediction of airside performance. Therefore, this research provides a basis for further investigation of machine learning approaches in airport airside integral performance analysis. The model contributes to analysis, strategic planning, and benchmarking processes, enabling managers to develop scenarios and analyse the interdependencies between airport airside characteristics and airside performance through KPI behaviour. Furthermore, to provide an initial assessment of generalisation, the model was tested on an independent dataset of 10 airports that are not included in the training data.

3. Methodology and proposed model

The proposed framework (Figure 1) combines statistical methods, data analysis techniques, and supervised machine learning algorithms to identify patterns and predict airport airside key performance indicators (KPIs) based on historical infrastructure and operational data. The framework enables the analysis of statistical relationships between airport characteristics and performance indicators by generating numerical predictions from selected input variables describing airport airside characteristics.

The analysis begins with the selection of the target KPI to be predicted. The framework then accesses a historical database containing airport infrastructure and operational data, which is prepared for machine learning through data cleaning, preprocessing, and feature transformation procedures. Five supervised machine learning regression algorithms are trained and evaluated using standard performance metrics, and the best-performing model is selected for KPI prediction. Throughout the modelling process, expert oversight is applied to verify data quality, validate preprocessing procedures, and ensure the reliability and consistency of the analytical workflow. Simultaneously with the prediction process, an assessment of feature importance is performed to determine the contribution of each feature to the given prediction. The final outputs of the model are the predicted KPI values and a clear understanding of feature importance for that prediction.

The objective of the model is to predict airport airside key performance indicators (KPIs) from airport infrastructure and operational characteristics by learning statistical relationships from historical data. The KPIs are defined by the EUROCONTROL document Methodology for Assessing Airport Airside Performance (ATMAP framework) [21].

Table 1. Overview of studies using machine learning methods for air traffic analysis

Research area	Feature no.	Data Sources	Methods	Evaluation Method	Target Variable	Machine learning type
Taxi Performance Evaluation Times [13]	13	Airport data	Artificial Neural Network, Gradient Boosting, Decision Tree	R-squared, Mean Absolute Error, Mean Squared Error	Taxi time	Regression
Distribution Prediction of Strategic Flight Delays [14]	11	Airport data	ML regression models	Confidence level	Flight distributions in strategic scheduling	Regression
Taxi-Out Time Prediction Model [12]	42	Airport data	Neural network, reinforcement learning, multilayer perceptron (MLP)	Root mean Squared Error	Departure taxi time prediction	Regression
Predicting the turnaround time [11]	24	Skyguide, VEA	Process structure-aware prediction, neural networks	Accuracy score	Aircraft turnaround time	Regression
Taxi-out time prediction [15]	21	Civil Aviation Administration of China, air traffic control	ML regression models	R-squared	Departure taxi time	Regression
Probabilistic Flight Delay Predictions, [16]	20	Airport data	Mixture Density Networks and Random Forest	Continuous Ranked Probability Score	Probabilistic prediction of arrival and departure delays	Regression
Taxi-Out Time Prediction [17]	12	Aerobahn system	Supervised ML algorithms	RMSE, MAE	Taxi time prediction	Regression
Airline delay prediction [10]	42	Federal Aviation Administration (FAA)	ML algorithms (supervised and unsupervised)	Area Under the Receiver Operating Characteristic Curve (ROC)	Flight delayed or not	Classification
Predicting Ground Delay Program [9]	10	FAA	Logistic regression, Decision trees	ROC	Departure delay prediction	Classification
Prediction of airport delays in the US [18]	47	U.S. Bureau of Transportation Statistics (BTS)	Neural Network, Linear regression	Accuracy Stratified 5-fold CV	Airport delay status	Classification
Prediction of On-time Performance of Flights [19]	12	BTS	Random forest Gradient Boosting, AdaBoost Extra-Trees	Accuracy	Takeoff and landing delay prediction	Classification
Assessing strategic flight schedules and cancellation predictions [20]	**13**	**Airport data**	LightGBM, Random Forest	Accuracy	Strategic flight schedule	Classification

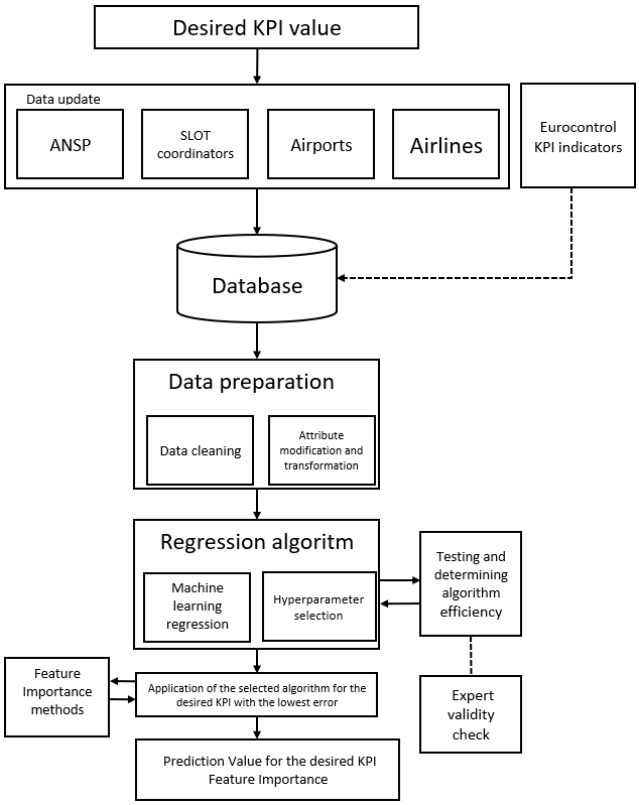


Figure 1. Description of the proposed conceptual prediction and analysis model.

The KPI data used in this study were obtained from the AIU (Aviation Intelligence Unit) Airport Performance Dashboard, which provides a consolidated monthly overview of the operational performance of European airports. It includes key indicators such as delays, flight punctuality, traffic throughput, operational efficiency, and other KPI metrics, enabling the monitoring and comparison of airport performance over time. All key performance indicators (KPIs) used in this study are continuous numerical variables. Therefore, no additional encoding was required for the target variables, as they were already represented in a numerical format suitable for machine learning regression models. The following Table 2 presents the target KPIs used in the model.

The study includes only those airports that provided complete reporting, as some airports report incomplete or irregular data. During the COVID-19 pandemic, there was a significant decline in air transport operations across European airports. The introduction of movement restrictions, border closures, and social distancing measures led to a reduced demand for air travel, which ultimately had a serious impact on the aviation industry, primarily through a decrease in the number of flight operations. For this reason, the dataset used for training in this study was taken from the years 2018 and 2019, as these periods were not affected by major disruptive events influencing air operations. Figure 2 shows the spatial distribution of the airports included in the study. The size of each airport icon is proportional to the total traffic at the airport during the observed time period. The map was created using the QGIS tool for the analysis and visualisation of geographic data.

Table 2. Target key performance indicators used in the model.

Abbrev.	KPI	Unit	Example Value
ATCPD	ATC_Pre departure delay	minutes	0.11
UTOT	Avg. Reference Taxi Out time	minutes	10.59
ATOT	Avg. additional Taxi Out time	minutes	1.28
RD	% of ATFM regulated departures	percentage	22.35
ATFMW	% of departures outside ATFM window	percentage	3.78
UTIT	Avg. Reference Taxi In time	minutes	4.42
ATIT	Avg. Additional Taxi In time	minutes	1.87
UAT	Total unimpeded reference Arrival Sequencing and Metering (ASMA) time	minutes	10.94
AAT	Total additional ASMA time	minutes	3.01

Independent variables in a machine learning model represent input data or features that are used to predict or classify the target variable, i.e., the key performance indicators (KPIs). These are the data that the algorithm uses to learn the relationships between inputs and outputs and to generalise to unseen data. Data were obtained from platforms provided by EUROCONTROL “Airport Corner” portal, AIP (Aeronautical Information Publication) publications, or directly through communication with airports.



Figure 2. Location of the airport data used within the EUROCONTROL area.

Some of the data on these platforms are not available in raw format, therefore, we obtained them directly in communication with EUROCONTROL and individual airports (Table 3). Since machine learning algorithms require numerical input data, all categorical variables were encoded before model development. For the preparation of categorical data for machine learning models, Label Encoding and One-Hot Encoding methods were applied. Label Encoding was used to transform categorical variables into numerical representations by assigning unique numerical labels to each category, while One-Hot Encoding was applied to variables without an inherent ordinal relationship between categories. The selection of the encoding method was based on the characteristics of the variables to ensure an appropriate data representation for the applied regression algorithms. The encoding process was applied to transform qualitative characteristics into numerical representations while preserving the meaning of the original variables. The categorical and binary variables were encoded to enable their use in the machine learning models. For example, the Coordination Level (CL) variable, which represents level of airport capacity coordination, indicating whether and to what extent airport capacity is subject to coordination procedures, was categorised using numerical labels (1, 2, and 3), where each value corresponds to the respective coordination level. The Runway Configuration Cross (RCC) variable was transformed into a binary feature, where a value of 1 indicates the presence of the analysed runway configuration, while a value of 0 represents its absence. Instrument Landing System categories are numeric counts and required no encoding.

4. Data structure

For the purposes of model training, 9 datasets were created (Figure 3), each corresponding to an individual performance indicator.

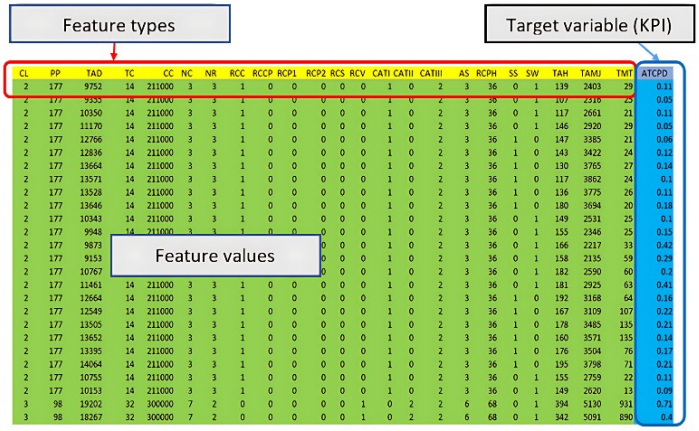


Figure 3. Example of one of the nine datasets used for model training.

Data from 76 airports were used to train the model. The selection criterion was the availability of complete monthly reports for the entire study period, therefore, only airports with complete monthly data were included in the analysis. There are nine datasets in total, corresponding to each of the KPIs examined. Each record consists of input data (features) - shown in the first 23 columns in Figure 3 and the final column representing the observed KPI. Each airport contributes 24 monthly records (12 months × 2 years), resulting in 76 × 24 = 1,824 records per KPI. The dataset contains no missing or duplicate values. The data used from the EUROCONTROL Airport Performance Dashboard represents monthly KPI values at the individual airport level. This means that a single value represents the aggregated result for a specific airport during a specific month, rather than data for an individual flight. For example, the ATC pre-departure delay value for Airport X in July 2019 represents the metric for all relevant operations at that airport during that month. Therefore, the unit of observation in the analysis is the airport-month.

Table 3. Input data (airport features) used in the model

Airport airside features	Definition	Type of variable	Value ex.
Coordination level (CL)	1 = non-coordinated, 2 = schedule-facilitated, 3 = fully coordinated	Categorical	2
Number of parking positions (PP)	No. of available aircraft parking stands	Continuous	177
Total arrivals and departures (TAD)	No. of total monthly flight movements (landings plus takeoffs)	Continuous	9752
Terminal Capacity (TC)	Maximum terminal passenger design throughput	Continuous	14
Global yearly current capacity (CC):	Declared total annual airport throughput capacity	Continuous	211000
Number of configurations (NC)	No. of possible operational runway configurations used at the airport	Continuous	3
Number of Runways (NR)	Total number of runways	Continuous	3
RC Cross (RCC)	Type of runway configuration	Binary	1
RC Cross Parallel (RCCP)		Binary	0
RC Parallel <2500(RCP1)		Binary	0
RC Parallel ≥2500 (RCP2)		Binary	0
RC Single (RCS)		Binary	0
RC V Formation (RCV)		Binary	0
Instrument Landing System (CATI)	Type of Instrument Landing System at the airport	Continuous	0
Instrument Landing System (CATII)		Continuous	2
Instrument Landing System (CATIII)		Continuous	0
Approach separation (AS)	Minimum longitudinal separation between aircraft on approach.	Continuous	3
Runway capacity per hour (RCPH)	Maximum number of aircraft movements that can be accommodated by the runway system per hour	Continuous	36
Season_Winter (SW)	Winter season is from November to March.	Binary	1
Season_Summer (SS)		Binary	0
TAH	Number of completed aircraft handling operations ((heavy (H), medium jet (MJ) and medium turboprop (MT))	Continuous	139
TAMJ		Continuous	2403
TMT		Continuous	29

5. Results

The following section presents the results obtained after applying machine learning algorithms to the given datasets. The analysis was conducted on 9 key performance indicators. Since this is a regression problem involving the prediction of continuous values, the following regression algorithms were used: Multiple Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB).

The model code was written in the Python programming language and implemented in the Google Colaboratory platform. The dataset was split into a training set and a test set using an 80:20 ratio. To avoid data leakage, the test set is kept separate from the training set and was not used during model training. It is used to evaluate the model's predictive performance on unseen observations. A fixed random seed of 42 was used to ensure the reproducibility of results. There were no missing values in the dataset. Consequently, no

imputation procedure was applied. Subsequently, 5-fold cross-validation was performed on the training set, incorporating random data shuffling and a fixed random seed of 42. Model performance during cross-validation was evaluated using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). Hyperparameters were predefined, where the RF model was configured with 500 trees (n_estimators = 500). GB model utilized 500 estimators, a learning rate of 0.05, and a maximum tree depth of 3. XGB model was configured with 500 estimators, a learning rate of 0.05, a maximum tree depth of 3, and an objective function based on squared error. LR and SVR utilized the default model parameters available in the scikit-learn library.

Feature standardization was performed within the pipeline for the LR and SVR models, whereas scaling was not applied to the decision tree-based models (RF, GB, and XGB). Scaling parameters were estimated separately within the training set of each validation subset to prevent data leakage. Following cross-validation, each model was retrained using the full training set and evaluated once on a previously set-aside independent test set. Final performance on the test set was assessed using R^2, MAE, and RMSE metrics. For each KPI, the best-performing algorithm was selected according to mean cross-validation R^2 on the training set. The test set was used only once, for the final evaluation of the selected model. RMSE (Root Mean Squared Error) represents the square root of MSE and expresses the error in the same units as the target variable, which makes the results easier to interpret. Lower RMSE values indicate better model performance.

The coefficient of determination (R^2) measures the proportion of the variance in the dependent variable that is accounted for by the model relative to the variance of the observed values. Values closer to 1 generally indicate better predictive performance. R^2 can also take negative values when the model performs worse than a baseline model that predicts the mean of the observed target values. RMSE test represents the root mean squared error calculated on the independent test set and provides a measure of prediction error in the same units as the target variable, with lower values indicating better predictive performance. Mean Absolute Error (MAE) measures the average absolute difference between observed and predicted values. It is expressed in the same units as the target variable and is less sensitive to large errors than RMSE. In this study, MAE was included primarily because of its straightforward interpretation as the average magnitude of prediction error, facilitating the practical interpretation and comparison of model performance. By combining these metrics, we observe the performance of the model, where models with lower RMSE and MAE values and higher R^2 values are considered more successful in predicting target values.

Figure 4 presents a comparison of the actual test set values and the predicted values of the key indicator Average Additional Taxi-Out Time (ATOT). Five machine learning algorithms were tested: Multiple Linear Regression, Random Forest, Support Vector Regression, Gradient Boosting, and Extreme Gradient Boosting. The actual values are marked in blue, while the predicted values are marked in red. Such graphs were obtained for all nine KPIs. However, due to page limitations in this article, only one graph for the ATOT indicator is presented, while the remaining values are shown numerically in Table 4. For the observed ATOT performance indicator, the algorithm that provided the best performance was Random Forest, with a CV R^2 of 0.87 ± 0.02 (test $R^2 = 0.91$, test MAE = 0.39)

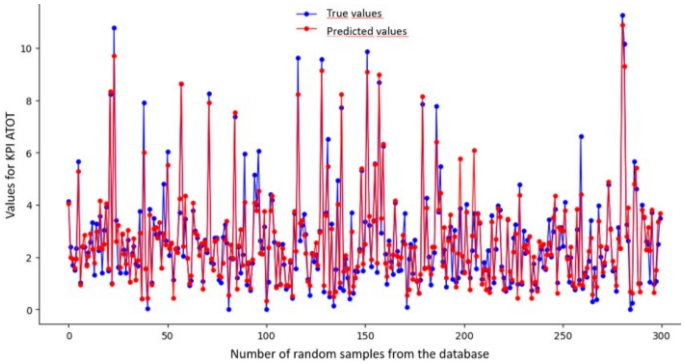


Figure 4. Prediction results using the Random Forest algorithm for the ATOT indicator.

The model's predictive capability was evaluated using both five-fold cross-validation and an independent test set, allowing its predictive accuracy and consistency to be assessed on both validation and unseen data. Table 4 contains results of best-performing machine learning models for the prediction of each 9 KPI indicators. The data is first divided into 80% for training and 20% for independent testing, while 5-fold cross-validation (CV) is performed on the training set to assess the stability and generalisation ability of the model.

Table 4. Cross-validation and test-set performance of the best-performing model for each KPI.

KPI	Best Model	CV R^2 (Mean ± SD)	CV R^2 95% CI	CV MAE (Mean ± SD)	CV MAE 95% CI	CV RMSE (Mean ± SD)	CV RMSE 95% CI	R^2 Test	MAE Test	RMSE Test
UTOT	RF	0.95 ± 0.02	[0.93, 0.97]	0.25 ± 0.02	[0.23, 0.27]	0.38 ± 0.04	[0.33, 0.43]	0.96	0.23	0.35
ATOT	RF	0.87 ± 0.02	[0.85, 0.90]	0.42 ± 0.04	[0.37, 0.47]	0.64 ± 0.07	[0.56, 0.73]	0.91	0.39	0.58
ATIT	RF	0.78 ± 0.04	[0.73, 0.82]	0.18 ± 0.02	[0.16, 0.20]	0.27 ± 0.03	[0.24, 0.31]	0.79	0.19	0.28
UTIT	RF	0.95 ± 0.00	[0.94, 0.95]	0.19 ± 0.01	[0.18, 0.20]	0.28 ± 0.01	[0.26, 0.29]	0.95	0.19	0.28
ATCPD	XGB	0.82 ± 0.03	[0.79, 0.85]	0.16 ± 0.01	[0.15, 0.17]	0.28 ± 0.03	[0.23, 0.32]	0.84	0.15	0.24
ATFMW	RF	0.66 ± 0.10	[0.53, 0.79]	1.36 ± 0.03	[1.32, 1.40]	2.24 ± 0.43	[1.70, 2.78]	0.70	1.39	2.25
AAT	RF	0.60 ± 0.09	[0.51, 0.69]	0.45 ± 0.06	[0.37, 0.53]	0.95 ± 0.26	[0.63, 1.27]	0.69	0.39	0.74
UAT	GB	0.90 ± 0.09	[0.77, 0.99]	0.23 ± 0.03	[0.19, 0.27]	0.48 ± 0.23	[0.19, 0.77]	0.97	0.17	0.24
RD	RF	0.76 ± 0.02	[0.74, 0.79]	5.15 ± 0.17	[4.94, 5.36]	6.62 ± 0.37	[6.16, 7.09]	0.81	4.75	6.14

The 95% confidence intervals (CIs) for the cross-validation performance metrics were calculated using the Student's t-distribution based on the results obtained from the five cross-validation folds. Confidence intervals were computed as the sample mean ± the critical t-value multiplied by the standard error ($SD/\sqrt{n}$), where SD is the sample standard deviation, n = 5 is the fold numbers, and the critical t-value was 2.776 for four degrees of freedom at a 95% confidence level. Accordingly, the intervals were calculated as mean ± 1.241 × SD. The reported confidence intervals were checked for consistency with the corresponding mean and standard deviation values. The performance of Linear Regression, Random Forest, SVR, Gradient Boosting, and XGBoost models is evaluated using R^2, MAE and RMSE indicators, with mean standard deviation, and 95% confidence interval (CI) calculated for cross-validation results. After that, each model is trained on the complete training set and evaluated on an independent test set, thus obtaining a separate assessment of the model's predictive performance. We observe that the best performance was provided mostly by ensemble algorithms, Random Forest, Gradient boosting and Extreme Gradient Boosting algorithms.

5.1. Out-of-group and temporal generalisation validation

To prevent airport-level information leakage, the predictive model was evaluated using an independent sample of 10 airports that did not appear in the training or test datasets (Table 5). Furthermore, the validation was performed using data from July 2024, representing a year that was not included in the datasets used for model development and testing.

Due to the pronounced seasonality of European airport operations, July was selected as the validation month. July 2024 was selected because European air traffic had substantially recovered from the COVID-19 disruption, providing a more representative operational period for out-of-sample validation. The upper part of the table presents the actual KPI values for July 2024, while the lower part presents the prediction results obtained using the proposed model. The obtained results are comparable to those reported in Table 4, with R^2 values between 0.86 and 0.97. This validation procedure provides independence in both the spatial and temporal dimensions.

The selected airports were not represented in the datasets used for model development, while the year 2024 was also not included in the training data. Therefore, the independent validation provides a more rigorous assessment of the model's ability to generalise to both previously unseen airports and a previously unseen time period, rather than only predicting new monthly observations from airports and years already represented in the dataset.

Table 5. Out-of-group and temporal validation results.

	Actual KPI values for July 2024								
Airport	ATCPD	UTOT	ATOT	RD	ATFMW	UTIT	ATIT	AAT	UAT
Vienna (LOWW)	0.84	6.72	3.27	60.35	1.96	4.79	1.27	3.4	11.2
Amsterdam (EHAM)	2.09	10.01	2.69	35.03	2.46	6.19	1.95	2.51	12.03
Paris (LFPG)	0.66	12.17	3.68	35.73	4.47	7.23	2.16	2.06	13.45
Frankfurt (EDDF)	1.01	10.86	3.15	44.16	5.92	7.23	2.26	2.57	12.96
Madrid (LEMD)	0.79	12.45	4.26	32.82	4.16	6.56	1.87	3.01	12.13
Rome (LIRF)	0.43	7.81	1.71	19.56	3.81	4.47	1.37	2.54	10.45
London (EGLL)	0.87	13.71	10.1	32.08	2.63	5.39	3.77	5.82	14.15
Munich (EDDM)	0.61	11.29	4.78	36.27	7.26	5.06	2.74	3.31	12.06
Malaga (LEMG)	0.71	9.04	2.94	41.49	6.32	3.26	1.64	3.09	11.6
Milan (LIMC)	0.73	10.91	4.89	45.11	3.2	4.68	2.33	4.53	10.21
	Predicted KPI values for July 2024								
Vienna (LOWW)	0.68	6.82	2.88	63.22	1.67	4.7	1.2	3	11.4
Amsterdam (EHAM)	2	10.09	2.98	31.12	2.01	6.34	2.13	2.74	12.26
Paris (LFPG)	0.7	12.03	3.19	40.67	5.23	7.1	2.36	1.78	13.42
Frankfurt (EDDF)	0.92	11.76	3.34	46.22	6.21	7.5	1.96	2.22	12.91
Madrid (LEMD)	0.89	12.33	3.87	27.55	4.67	6.23	1.67	3.45	12.2
Rome (LIRF)	0.47	7.64	1.94	17.44	4.33	4.45	1.2	2.76	10.43
London (EGLL)	0.75	13.83	9.81	30.12	2.44	5.55	3.57	5.76	14.25
Munich (EDDM)	0.65	10.2	4.68	34.28	6.56	4.99	2.55	3.1	11.97
Malaga (LEMG)	0.82	9.15	2.55	45.66	5.66	3.36	1.87	3.2	11.8
Milan (LIMC)	0.6	10.81	4.5	42.47	4	5.2	2.56	4.2	11.53
MAE	0.09	0.29	0.32	3.19	0.52	0.18	0.20	0.26	0.23
Test R^2	0.95	0.95	0.97	0.89	0.86	0.96	0.92	0.91	0.87

Model evaluation was conducted by randomly splitting the data into training and test sets using an 80:20 ratio. Subsequently, five-fold cross-validation was performed at the level of individual airport–month observations. Consequently, observations pertaining to the same airport could appear in both the training and test sets, although representing data from different months. This validation strategy assesses the model's ability to predict previously unseen observations for airports that may have already been represented in the training dataset. However, it does not allow for a comprehensive assessment of the model's ability to generalize results to entirely new airports.

To address these limitations, an additional validation was performed using airports that were not included in either the training or test sets (Table 5). This airport-level validation provides a more rigorous assessment of the model's generalisation capability when applied to airports that were not represented in the original training dataset.

Thus, the two validation approaches serve distinct purposes. The initial validation at the level of individual observations assesses predictive performance on previously unseen monthly airport observations, while the additional airport-level validation evaluates the model's applicability to airports not included in the training process. It should be noted that random splitting of observations may lead to overly optimistic estimates of model performance regarding generalisation to new airports.

5.2. Feature importance analysis

Examining feature importance in machine learning models is the process of identifying which input variables contribute most to the model’s ability to predict the target variable. Feature importance analysis is important for understanding the statistical relationships and dependencies between airport characteristics and key

performance indicators. It provides insights into which factors are most associated with the observed performance patterns in historical data. This method is tested on the KPI Average Additional Taxi-Out Time (ATOT).

Two feature-importance methods were applied. First, impurity-based feature importance (Figure 5) is extracted from the trained Random Forest model. This method quantifies the relative contribution of each predictor to the reduction in squared-error impurity across the decision trees. The first diagram (Figure 5) shows how individual input variables contributed to reducing prediction error during the construction of decision trees in the Random Forest model. A longer bar indicates greater estimated importance of a variable during model training. A shorter bar indicates lower estimated importance. The importance values are normalized so that their sum equals 1.

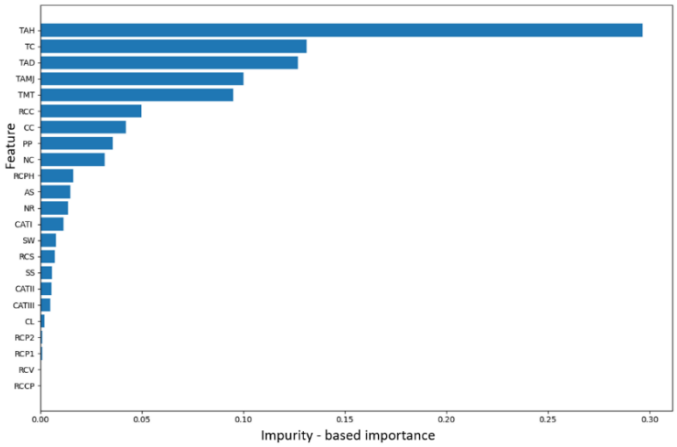


Figure 5. Random forest impurity-based feature importance.

Figure 6 shows permutation feature importance is calculated using the held-out test dataset, with 30 repetitions per predictor and (R^2) as the evaluation metric. This method measures the decrease in predictive performance when the values of an individual predictor are randomly shuffled. The mean decrease and standard deviation across repetitions are reported, providing an indication of the predictor's predictive relevance and the variability of the estimate.

Diagram in Figure 6 shows how much the model's predictive performance decreases when the values of an individual input variable are randomly shuffled in the test dataset. Unlike the first graph, this method is based on the test dataset, which was not used to train the model. Therefore, it provides additional insight into the importance of individual variables for predicting outcomes on unseen data. The horizontal bars show the mean permutation importance values, while the black lines represent the standard deviation across 30 permutations.

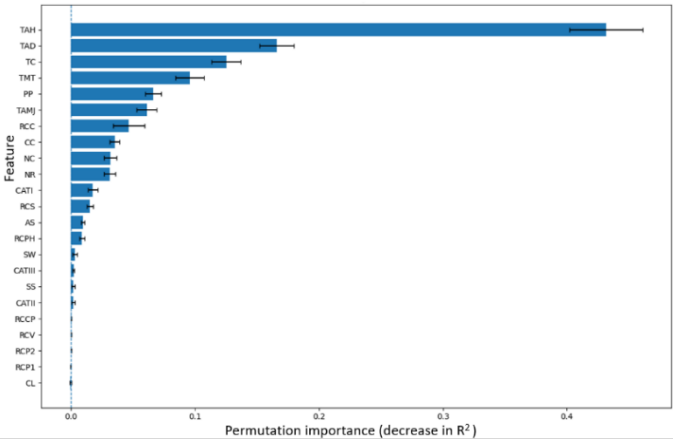


Figure 6. Random Forest permutation feature importance on the held-out test set.

Table 6 presents the results regarding the importance of prediction attribute values for both methods, along with the standard deviation. The achieved model performance is $R^2 = 0.91$. According to the impurity-based method, the variables with the highest importance are TAH (0.29), TC (0.13), TAD (0.12), and TAMJ (0.10), whereas according to the permutation importance method, the four attributes with the greatest impact on the prediction are TAH (0.43), TAD (0.16), TC (0.12), and TMT (0.09). The calculated standard deviation for the Permutation importance method ranges from 0.01 to 0.03. The difference in variable importance rankings depends on the measurement method used. Impurity-based importance is normalised (values sum to 1), while permutation importance is the mean decrease in test-set R^2. The two scales are different and only the rankings and not the absolute values should be compared. The results reflect the predictive importance of the variables rather than evidence of cause-and-effect relationships.

Table 6. Feature importance of the four most relevant predictors for ATOT prediction.

Feature	Impurity-Based Importance	Feature	Permutation Mean	Permutation SD
TAH	0.29	TAH	0.43	0.03
TC	0.13	TAD	0.16	0.01
TAD	0.12	TC	0.12	0.01
TAMJ	0.10	TMT	0.09	0.01

6. Discussion

The study demonstrated that machine-learning algorithms can be used to analyse complex systems such as the air transport system. The model, based on the assessment of statistical prediction accuracy, showed different predictive performance for the different KPIs examined in the study. It is important to emphasise that the analysis used aggregated monthly airport-level performance data rather than individual flight observations. Therefore, the developed model should be interpreted as a model of statistical relationships at the airport-month level.

Considering the regression algorithms tested in this study, the best performance was generally achieved by nonlinear ensemble methods. This finding reflects the complexity of the airport environment, where a large number of interdependent and potentially nonlinear relationships exist between infrastructure characteristics and operational demand. The results also showed that different algorithms can produce different levels of predictive performance for the same KPI. The performance of the individual algorithms is presented in Table 4, where five-fold cross-validation was performed. The ensemble algorithms achieved the best results for the majority of the examined KPIs, indicating their ability to capture complex statistical relationships within the available data.

However, cross-validation based on observations from the same overall dataset cannot completely address the possibility of airport-specific patterns. For this reason, the developed models were additionally evaluated using an independent sample of 10 airports that were not included in the training dataset. As presented in Table 5, the prediction results obtained for these airports remained within the ranges of variability and error represented by the standard deviations and error values reported in Table 4. The independent validation was performed using data from 2024, a year that was not included in the datasets used for model development. This provides a more demanding assessment of whether the statistical relationships learned from historical observations can be applied to previously unseen airports and to a substantially later time period. The results provide additional evidence that the predictive performance of the models was not substantially degraded when applied to the independent validation sample.

After applying the machine-learning algorithms, an impurity-based feature-importance analysis was performed to identify the predictive relevance of the input variables for each KPI (only ATOT was shown in the paper). The results indicate the relative importance of individual attributes for predicting the selected KPI within the fitted model. Lower importance values for certain runway configurations do not necessarily indicate that these configurations have limited operational significance. Rather, they indicate that, within the dataset used in this study, these variables provided less additional predictive information compared with other characteristics.

Therefore, the feature-importance analysis should be interpreted as an indicator of the predictive relevance of individual attributes for a given KPI, rather than as evidence of a causal relationship. The identified importance reflects how useful each variable was to the model in predicting the observed KPI values and should not be interpreted as demonstrating that a change in a particular airport characteristic directly causes a change in the corresponding KPI.

Model performance uncertainty was quantified using 5-fold cross-validation on the training set, reporting the mean, standard deviation (SD), and 95% Confidence Intervals (CI) for each metric. For most KPIs, the ensemble models demonstrated narrow confidence intervals and low standard deviations across folds (Table 4). For example, Average Additional Taxi-Out Time (ATOT) achieved a cross-validation MAE of 0.42 ± 0.04 minutes (95% CI: 0.37, 0.47). The narrow confidence intervals and low standard deviations across folds indicate stable performance across folds for most KPIs, although variability is larger for ATFMW, AAT and UAT.

The study also has several limitations. The KPI predictions are based on aggregated monthly values, therefore, the developed approach is primarily applicable at a strategic rather than tactical level. Furthermore, the analysis was based on data reported through EUROCONTROL sources, meaning that not all potentially relevant airport airside KPIs and attributes could be included. Furthermore, the variables considered over the two-year period did not change significantly,

particularly infrastructure-related variables such as the number of parking positions and the number of runways. Therefore, future research could extend the analysis to a longer period covering years in which significant infrastructure changes occurred.

In particular, the dataset does not fully capture factors such as weather conditions, air traffic control workload, detailed taxiway topology, runway occupancy characteristics, and temporary operational restrictions. The absence of these variables may limit the ability of the models to represent some short-term variations in airport airside performance.

Despite these limitations, the developed model provides a basis for further data-driven analysis of airport airside performance. Its main contribution is not the identification of causal effects, but the systematic examination of statistical relationships between airport infrastructure and operational characteristics and multiple continuous airside KPIs. The combination of five-fold cross-validation, independent airport-level validation, and feature-importance analysis strengthens the assessment of the predictive results and provides a foundation for further research into data-driven approaches for analysing airport airside performance.

7. Conclusion

The proposed framework is designed to support strategic airport performance monitoring, benchmarking, and scenario analysis related to capacity and infrastructure planning. In addition to predictive performance assessment, variable importance analysis is used to examine the statistical contribution of various airport characteristics to variations in the observed KPIs. The robustness and generalisability of the proposed framework were further tested through independent spatial and temporal validation, enabling an assessment of its applicability across different airports and time periods.

The model analyses and predicts key airport airside performance indicators and assesses the algorithms' ability to capture complex statistical relationships. Key operational and traffic-related airport variables were used as input data. The analysis utilized data from the EUROCONTROL database, covering 76 airports, with training and testing performed on data from 2018 and 2019. The algorithms demonstrated varying performance across different KPIs. Ensemble methods - specifically Random Forest, Gradient Boosting and Extreme Gradient Boosting yielded the best results, effectively capturing non-linear relationships between attributes and key performance indicators. A feature importance analysis was also conducted to identify the attributes contributing most significantly to the prediction of chosen KPIs (ATOT as an example). Following the training and testing phases, an additional analysis was performed on data from 10 airports not included in the training or test set, using data from 2024. This independent evaluation further provided supporting evidence of generalisation, although limited by the small sample. Static infrastructure features uniquely identify each airport across its 24-month timeline, tree-based models may partially memorize airport-specific patterns rather than learning generalizable relationships. Although the independent validation on 10 unseen airports from 2024 partially mitigates this issue, future work should apply airport-grouped cross-validation (e.g., GroupKFold) for the primary evaluation to ensure stricter spatial generalizability.

The findings indicate that machine learning can be used as an analytical tool for identifying patterns in airport airside performance and for supporting the interpretation of relationships among operational and traffic-related variables. Such models may provide additional quantitative information for airport performance monitoring and decision-support applications, but their outputs should complement rather than replace established operational knowledge and performance-assessment methods.

For future research, the proposed model can be extended by using a database with a larger number of attributes and key performance indicators. In this study, historical records covering a period of two years were considered. It is recommended to include data spanning multiple years, particularly those in which major infrastructural changes occurred (such as the construction of a new runway, implementation of a higher-category ILS, etc.), in order to improve the model’s ability to learn from the data. It is also recommended to develop a real-time database in which indicators would be monitored in real time. The input data would be dynamic in nature, such as current declared airport capacity, runway capacity per hour, number of air traffic controllers, number of

passengers, etc. The proposed model was used to analyse the airport airside, therefore, the next step should focus on integrating other components of the air transport system, such as the airport landside and the air traffic network, i.e., the airspace.

Furthermore, future research could include a comparative analysis between the European and the United States air transport systems by incorporating data obtained from the Federal Aviation Administration (FAA). Such a comparison would enable the evaluation of differences in airport infrastructure, operational procedures, air traffic management practices, and key performance indicators across the two aviation systems. The inclusion of FAA datasets would also increase the diversity of the training data and improve the model's robustness, generalisability, and applicability in different operational environments.

Competing Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the conduct, analysis, authorship, or publication of this work.

Data Availability Statement

The supplementary materials, research datasets, and analysis code used in this study are available upon reasonable request, subject to any applicable data-sharing restrictions. The complete datasets and analysis code may be obtained by contacting the corresponding author via email at edvin.simic@fsk.unsa.ba.

Funding Statement

This research was supported by the Federal Ministry of Education and Science - Federation of Bosnia and Herzegovina through the Support for Scientific Advancement During Doctoral and Postdoctoral Studies program.

References

[1] EUROCONTROL Performance Review Commission, "Performance Review Report 2025," EUROCONTROL, Belgium, Mar. 2026.

[2] E. Šimić and M. Begović, "Airport delay prediction using machine learning regression models as a tool for decision making process," i*n Proc. 45th Int. Conv. Information, Communication and Eectronic Technology (MIPRO)*, Opatija, Croatia, pp. 841–846, 2022, doi:10.23919/MIPRO55190.2022.9803809

[3] H. Idris, J.-P. Clarke, R. Bhuva, and L. Kang, "Queuing Model for Taxi-Out Time Estimation," *Air Traffic Control Q.*, vol. 10, no. 1, pp. 1–22, 2002, doi: 10.2514/atcq.10.1.1.

[4] X. Wang, A. E. I. Brownlee, J. R. Woodward, M. Weiszer, M. Mahfouf, and J. Chen, "Aircraft taxi time prediction: Feature importance and their implications," *Transp. Res. Part C Emerg. Technol.*, vol. 124, p. 102892, Mar. 2021, doi: 10.1016/j.trc.2020.102892.

[5] S. Liu, J. Zhang, Z. Peng, H. Guo, and A. Pi, "Identifying operational benefits of the arrival management system – a kpi-based experimental method by evaluating radar trajectories," *Promet - Traffic - Traffico*, vol. 33, no. 5, pp. 633–645, 2021, doi: 10.7307/PTT.V33I5.3786.

[6] M. Lu, P. Wei, M. He, and Y. Teng, "Flight Delay Prediction Using Gradient Boosting Machine Learning Classifiers," *J. Quantum Comput.*, vol. 3, no. 1, pp. 1–12, 2021, doi: 10.32604/jqc.2021.016315.

[7] M. Schultz, S. Reitmann, and S. Alam, "Predictive classification and understanding of weather impact on airport performance through machine learning," *Transp. Res. Part C Emerg. Technol.*, vol. 131, p. 103119, Oct. 2021, doi: 10.1016/j.trc.2021.103119.

[8] P. Taleongpong, S. Hu, Z. Jiang, C. Wu, S. Popo-Ola, and K. Han, "Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network," *J. Intell. Transp. Syst.*, vol. 26, no. 3, pp. 311–329, May 2022, doi: 10.1080/15472450.2020.1858822.

[9] A. Mukherjee, S. Grabbe, and B. Sridhar, "Predicting ground delay program at an airport based on meteorological conditions," *AIAA Aviat. 2014 -14th AIAA Aviat. Technol. Integr. Oper. Conf.*, no. June, pp. 1–18, 2014, doi: 10.2514/6.2014-2713.

[10] A. Dand, "Airline delay prediction using machine learning algorithms," Ph.D. dissertation, Wichita State Univ., Wichita, KS, USA, 2020.

[11] O. F. J. V. Hassel, "Predicting the Turnaround Time of an Aircraft: A Process Structure Aware Approach," Master's thesis, Eindhoven Univ. of Technology, Eindhoven, The Netherlands, 2019.

[12] F. Herrema, R. Curran, H. Visser, D. Huet, and R. Lacote, "Taxi-out time prediction model at Charles de Gaulle airport," *J. Aerosp. Inf. Syst.*, vol. 15, no. 3, pp. 120–130, 2018, doi: 10.2514/1.I010502.

[13] E. Çömez and O. İnan, "Performance evaluation of machine learning algorithms in estimating taxi times at

Istanbul Airport," *Intell. Methods Eng. Sci.*, vol. 3, no. 3, pp. 82–90, Sep. 2024, doi: 10.58190/imiens.2024.102.

[14] Z. Wang, C. Liao, X. Hang, L. Li, D. Delahaye, and M. Hansen, "Distribution Prediction of Strategic Flight Delays via Machine Learning Methods," *Sustain. Switz.*, vol. 14, no. 22, pp. 1–14, 2022, doi: 10.3390/su142215180.

[15] J. Yin, Y. Hu, Y. Ma, Y. Xu, K. Han, and D. Chen, "Machine learning techniques for taxi-out time prediction with a macroscopic network topology," *in Proc. 37th AIAA/IEEE Digital Avionics Systems Conf. (DASC)*, 2018, pp. 1–8, doi: 10.1109/DASC.2018.8569664.

[16] M. Zoutendijk and M. Mitici, "Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem," *Aerospace*, vol. 8, no. 6, 2021, doi: 10.3390/aerospace8060152.

[17] H. Lee, W. Malik, and Y. C. Jung, "Taxi-out time prediction for departures at charlotte airport using machine learning techniques," *16th AIAA Aviat. Technol. Integr. Oper. Conf.*, no. June, pp. 1–11, 2016, doi: 10.2514/6.2016-3910.

[18] K. R. Chandramouleeswaran, D. Krzemien, K. Burns, and H. T. Tran, "Machine learning prediction of airport delays in the US air transportation network," *2018 Aviat. Technol. Integr. Oper. Conf.*, pp. 1–10, 2018, doi: 10.2514/6.2018-3672.

[19] B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan, and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," *in Proc. 36th AIAA/IEEE Digital Avionics Systems Conf. (DASC)*, 2017, doi: 10.1109/DASC.2017.8102138.

[20] M. Lambelho, M. Mitici, S. Pickup, and A. Marsden, "Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions," *J. Air Transp. Manag.*, vol. 82, p. 101737, Jan. 2020, doi: 10.1016/j.jairtraman.2019.101737.

[21] EUROCONTROL Performance Review Commission, "ATM Airport Performance (ATMAP) Framework: Measuring Airport Airside and Nearby Airspace Performance," EUROCONTROL, 2009.