Journal of Air Transport Management 122 (2025) 102684

Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman

Exploring prediction accuracy for optimal taxi times in airport operations using various machine learning models

Simon Okwir [a,*], Kaveh Amouzgar [a,b], Amos HC. Ng [a,b]

[a] Division of Industrial Engineering and Management, Uppsala University, 75310 Uppsala, Sweden
[b] Division of Intelligent Production Systems, School of Engineering Science, University of Skövde, 54128 Skövde, Sweden

ARTICLE INFO

Keywords:
Prediction accuracy
Turnaround operations
Collaborative decision making
Airport operations
Machine learning

ABSTRACT

Understanding delay conditions and making accurate predictions are essential for optimizing turnaround and taxi times, which in turn reduces fuel consumption and lowers CO_2 emissions in airport operations. However, while existing research has explored the impact of various prediction models on airport operations, it often overlooks the performance of Collaborative Decision Making (CDM) variables when discussing delay conditions. The implementation of CDM at major European airports has led to a milestone-based approach within airport operations, particularly in the turnaround operations, segmenting these operations with unique features. The purpose of this paper is to systematically investigate the efficacy of various machine learning techniques, such as linear regression, regression trees, random forests, elastic nets, and multi-layer perceptrons (MLP), in accurately predicting delay categories within the CDM framework. For this purpose, we analyzed CDM operational data from Madrid Airport, with at least 166,185 flight observations. Our findings illustrate a training methodology on how different models vary in prediction accuracy when applied to CDM operational data. We applied the SHAP (SHapley Additive exPlanations) method for feature importance analysis of all our independent variables to interpret the output of our machine learning models. Our results indicate that linear regression and elastic nets are the most effective machine learning models for achieving high prediction accuracy within the CDM framework. To test their robustness, we extended the analysis with predictions for better schedule times for taxi times on arrival and depature for selected runways using a different dataset. Our results contribute by showcasing a training methodology, highlighting how elastic net model as the best-performing model can be adopted for turnaround operations. In conclusion, we discuss the implications of our results for runway demand policies and use of airport resources such as gate & runaway allocation.

1. Introduction

In the context of current global challenges, including the aftermath of the COVID-19 pandemic and escalating geopolitical tensions, recent scholarly research has highlighted a notable resurgence and sustained expansion in the civil aviation sector (Lemetti et al., 2023; Michelmann et al., 2023). As a result, optimal airport capacity utilization continues to be a significant challenge, particularly within turnaround operations at major airports (Thiagarajan et al., 2017; Yu et al., 2019b). Moreover, the propagation of delays within the turnaround operations adversely impacts on-time performance of flights (Lin, 2023). In this paper, we propose to leverage machine learning models to improve prediction of delays in turnaround operations, including activities such as ground handling, fueling, cleaning, and the allocation of runways and gates. These operations are currently managed through Collaborative Decision Making (CDM) (Zuniga and Boosten, 2020).

CDM is an optimization standard that has been successfully implemented at major airports in Europe. Despite of its greater benefits, CDM redefines the way how prediction accuracy is determined. Through a milestone approach, CDM brings upon new performance indicators as shown in Table 2. However, there is not much evidence how different machine learning models may be leveraged within the CDM framework for prediction accuracy. The purpose of this paper is to systematically investigate the efficacy of diverse machine learning techniques, such as linear regression, regression trees, random forests, elastic net, and multi-layer perceptrons (MLP), in accurately predicting delay categories within the CDM framework. Our work focuses on evaluating the precision and reliability of these models, with an aim to enhance the

* Corresponding author.
E-mail address: simon.okwir@angstrom.uu.se (S. Okwir).

https://doi.org/10.1016/j.jairtraman.2024.102684
Received 7 April 2022; Received in revised form 18 July 2024; Accepted 19 September 2024
Available online 23 October 2024
0969-6997/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

predictive capabilities and operational efficiency in the CDM process.

For instance, the overall Target Off Block Time (TOBT) is key indicator within the CDM framework that is crucial for on time-performance in the turnaround (Alonso Tabares et al., 2021; Eurocontrol, 2006; Okwir et al., 2017; Zuniga and Boosten, 2020a). TOBT is when an aircraft operator or ground handler estimates that an aircraft will be ready, all doors closed, boarding bridge removed, push back vehicle available, and ready to start up upon reception of clearance from the tower. Within the CDM context, TOBT is a measure that is collaboratively achieved by airport actors who share timely information to increase situational awareness. But due to a poor prediction accuracy of these new KPIs such at TOBT, adversely impacts the overall demand and supply of runways throughput, making the management of airside operations in the turnaround constrained.

In this study, we leverage machine learning to improve the error reductions in the airport operations management, an aspect that has received less attention with CDM airports. While CDM milestones offer an optimization process with various KPIs, there remains limited understanding of the extent to which available machine learning models offer optimal accuracy within the new CDM working framework. Furthermore, considering the interdependency network of turnaround operations, our inquiry focuses on how collaborative measures, involving shared roles among airport turnaround actors, contribute to segmenting the milestones provided by the CDM framework.

The significance of prediction accuracy in turnaround operations involves multiple airport actors, such as airline operators, ground handling companies, airport operators, and Air Navigation Service Providers (Eurocontrol, 2006). These actors operate with interdependent processes and shared resources, causing delays in their individual operations that ultimately affect the entire turnaround process. However, the limited comprehension of prediction accuracy with new CDM features may restrict the ability of airport actors to reduce turnaround times. Additionally, optimizing collaborative activities for pre-planning necessitates accurate predictions. We observe that there are three major changes within the turnaround operations due to the current CDM framework.

First, CDM brings significant transformations that it replaces take-off slot allocation and offers a new operational methodology in the turnaround. CDM has facilitated the shift from a “first come, first served” approach to a “best planned, best served” strategy. As an implication, this transition presents a challenge to airport actors to find new ways to predict their performances and network performance that is shared in real time. It is within this realm that the potential for well-managed, accurate predictions can be realized, particularly through the examination and application of machine learning models.

Secondly, CDM milestone approach requires various levels of collaboration among airport stakeholders. This collaboration is enhanced by the distribution of information, ensuring that all actors can share decision-making responsibilities, particularly when allocating scarce resources. High-level decisions require precise predictions. For example, in CDM operations, the accuracy of the Target Off Block Time (TOBT) heavily relies on collaborative processes and is used for milestone monitoring, which is crucial for on-time performance. These collaborations often a prone to delays and hence the need for prediction accuracy.

Thirdly, is the impact of turnaround delays to the EU air space. Due to the single European sky, the EU air space becomes a dependency network. Since CDM operations now aim to connect airports through the Flight Update Message (FUM) about the progress of each flight. Accurate predictions are essential not only for turnaround operations but also for network operations. This leads to improve en-route sector planning, resulting in slot compliance and the reduction of wasted slots at other CDM airports.

Our paper offers three contributions. Firstly, departing from prior studies such as those by Herrema et al. (2019a) and Diana (2018) on features such as taxi time prediction and demand features, our work introduces new features within the CDM milestone approach. We evaluate the effectiveness of various machine learning techniques, such as linear regression, regression trees, random forests, elastic nets, and multi-layer perceptrons (MLP), specifically in predicting delay categories with CDM features. Notably, our findings reveal that among these models, Linear Regression and Elastic Net exhibit the highest accuracy in predicting delays using airport data within the CDM context.

Secondly, our research not only introduces a method to improve prediction accuracy using the CDM milestone approach but also identifies critical features that are prone to increase delays. These features, including ASRT-TSAT, ASAT-ASRT, and ASAT-TSAT, consistently demonstrate a significant impact on the variability of delay categories prone to delays. This emphasizes the importance of optimizing collaborative activities, particularly during turnaround, by focusing on these specific indicators in the pre-planning phase.

Our key contribution lies in applying a tested machine learning model out of the many to a new dataset, aimed at minimizing taxi times for arrivals and takeoffs. Our findings demonstrate enhanced accuracy in scheduled times and reduced variability. We illustrate our analysis by employing a separate data set. We show that robust performance evaluation can be acchived within airport operations, leading to more effective planning, delay reductions and less CO2 emissions in the overall airport operations management.

In overall terms, our results contribute to the understanding of delays with CDM features and operations. This understanding is to also leveraged by not only airport actors on how to improve prediction accuracy but vast air space users to optimize collaborative activities in the pre-planning phase at various airports.

The paper is structured as follows: In the next section, we describe further the concept of CDM as conducted at CDM airports and discuss various prediction methods mentioned in the literature. In the third section, we present our method section and an extended analyis of our best performing model to a new data set. Finally, we conclude with a discussion of our findings and emphasize the importance of error reductions in prediction accuracy in understanding the current capacity constraints.

1.1. *Collaborative Decision Making at airports and its application*

According to the proposed European Air Traffic Management (ATM) master plan, four key areas are being implemented, as identified by Zuniga and Boosten (2020a). These areas include optimized ATM Network Services, Advanced Air Traffic Services, High Performing Airport Operations, and Enabling Aviation Infrastructure. Airport Collaborative Decision Making (A-CDM) is a seamless process through which the future ATM system plans to integrate airports into a common network. Once A-CDM is effectively implemented in European airports, the anticipated benefits will be realized at both the airport and network levels. At the airport level, benefits include enhanced runway throughput, efficient resource use, and reduced fuel consumption. These improvements will subsequently impact efficiency, environment, safety, and capacity at the network level, with European skies being the primary target area for 2025 (Eurocontrol, 2006).

However, despite its increasing adoption at several European airports, little has been established regarding post-implementation performance measurement of A-CDM. To maintain high performance and demonstrate success with A-CDM, further studies are required. There is a scarcity of studies demonstrating network performance through real airport operations data. The CDM implementation manual explains how CDM performance can be extended by presenting performance drivers, indicators, and measurements to be assessed, which are considered a conceptual model. This model deals with generic CDM objectives for the network and airport partner objectives. Other studies suggest that performance measurement can be carried out by developing specific KPIs that reflect the success of the system within a dependent network (Lohman et al., 2004).

For airports, these objectives should measure the effects of Airport CDM in two ways: 1) Performance can be compared against the same indicators utilized before implementation, and 2) Performance can be measured across the entire European airport network, necessitating standardized methods at both local and European regional levels. Despite the extended nature of this performance measurement, CDM airports have gradually moved away from this milestone approach, adopting new local measures (Okwir et al., 2017). The differentiation suggested by context-specific operations, such as weather conditions at various airports, underscores the need for accuracy in prediction. However, despite the extensive implementation of CDM, there is no consensus on a set of KPIs to measure CDM benefits to date. Therefore, there is a need to identify updated frameworks for prediction to determine if a common set of indicators can be derived from real empirical evidence. This may be challenging, as each airport has unique features and limitations, making the generalization of a taxonomy with a similar set of KPIs impractical.

Nevertheless, within this study's context, we explore the potential of leveraging machine learning models to improve prediction accuracy. In the next section, we discuss the use of big data and machine learning models within airport operations.

1.2. Big data and machine learning in airport operations

Faced with the challenge of limited capacity within existing airport infrastructure and the mounting pressure to meet the ever-growing demand for flight transportation, while ensuring passenger satisfaction and safety, there is an urgent need to develop more intelligent and automated civilian airport operations management systems (Wang et al., 2021a). The abundance of real-time data concerning flight information, performance, airport and air traffic conditions, weather, passenger data, and feedback, available from diverse sources, including social media platforms, is fueling rapid developments in Big Data and data analytics applications (Chung et al., 2020). The primary advantage of Smart Airports lies in leveraging big data analysis to enhance operational efficiency—a concept aiming to achieve "man-machine integration" by redefining airport operations and services based on IoT, mobile networks, and Big Data analytics. This paradigm shift enables remote control and monitoring of airport operations, unlike traditional airports (AlMashari et al., 2018; Rajapaksha and Jayasuriya, 2020). Noteworthy applications of Smart Airport initiatives are evident in Changi Airport in Singapore (Lee and Miller, 2019), Incheon Airport in Korea (Rajapaksha and Jayasuriya, 2020), and Heathrow Airport in London, UK (Guo et al., 2020). Through the integration of sensing and network infrastructure, data management systems, data analytics, as well as Artificial Intelligence (AI) and Machine Learning (ML) capabilities, these airports have significantly enhanced their operational efficiency levels.

Advances in data-analytic technologies, when applied to Big Data (i.e., large volumes of data from a broad variety of data sources collected in real-time), have given companies and organizations access to a large set of ML algorithms to extract useful knowledge from these unprecedented amounts and varieties of data (Wang et al., 2021b). ML is a sub-field of computer science, very often linked to the field of Artificial Intelligence (AI), which provides machines with the ability to learn without explicit programming, which is evolved from pattern recognition and computational learning theory (Nibareke and Laassiri, 2020). Dependent on the amount and type of supervision they get during the learning process, ML algorithms can be classified into four major categories: supervised, unsupervised, semi-supervised and Reinforcement learning. Dependent on the type of supervised training, it also decides the application of an ML-based system (Géron, 2019). Typical applications of supervised ML algorithms are classification and predicting a target numeric value. In the former, the training data fed into the algorithm are labelled as different classes of solutions (e.g., desired or undesired) so that the trained algorithm is able to classify new, "unseen" data into different classes. For the latter, given a set of features called predictors, or independent variables, the algorithm can predict the value of a dependent variable. Unlike the case of supervised learning, training data for unsupervised learning algorithms is unlabeled. Mixing labelled data with a lot of unlabeled data, using a combination of supervised and unsupervised algorithms, give rise to semi-supervised learning. Greasley (2019) further classifies data analytics using machine learning algorithms into the following three categories of analytics: descriptive, predictive and prescriptive.

Among these three categories, predictive analytics is by far the most popular in the state of the art related to airport operations management. In terms of the target of predicting, taxi-out time and flight delay draw the greatest interest in predictive analytic studies for improving airport operations. The concept of predicting taxi-out time using queuing models or simulation (Pujet et al., 2000) is not new, already started in early 2000, but most recent publications in the literature are clearly dominated by ML-based approaches, due to the above-mentioned massive of collecting real-time data and Big Data analytics trend. Pujet (2000) define taxi-out time as the sum of travel time, from gate to entry into the departure queue, and the waiting time for take-off on the runway. They are the observed times from pushback to take-off and provide an indicator of how congested an airport is if they exceed nominal values, i.e., taxi-out delays, as the difference between actual and unimpeded taxi-out times (Diana, 2018). According to (Diana, 2018), accurately predicting taxi-out times and delays are crucial for airport operations and planning because it enables to identify of the choke points between gate-out and take-off in order to recommend and implement infrastructure improvements for optimal surface movement flows. Similarly, flight delay is another critical performance indicator of airport operations widely recognized by the commercial aviation industry. It is one of the primary measures of efficiency in the system and is a major cost to traffic managers, airlines, passengers and even the environment because of extra fuel burns and re-scheduling (Esmaeilzadeh and Mokhtarimousavi, 2020).

As said, ML-based predictive analytic techniques have become extremely popular in recent years when compared to classic models because they, in general, can lead to higher prediction accuracy based on the massive amount of collected data. Table 1 below summarizes the recent ML-based studies in predicting taxi-out time and flight delay, followed by more detailed reviews.

1.3. Taxi-out time prediction

(Diana, 2018) compared the predictive accuracy of ten selected supervised ML models, including regressions (Ordinary Least Squared or OLS and regularized/penalized ones), SVM and ensemble-based algorithms such as Random Forest, Bagging, Adaptive Boosting, Extra Trees and Gradient Boosting. These methods are tested on actual hourly data from the Seattle/Tacoma International Airport (SEA) with five independent operational variables: departure demand, departure throughputs, percentage of total airport capacity utilized, approach conditions and runway configurations with an aim to explain variations in taxi-out times. The models utilized in the study were designed to support only short to medium-term predictions. The study started with a hypothesis that ensemble learning through the combination of several learners would improve the prediction accuracy of taxi-out times when compared with a single learner in the OLS, penalized and SVR. The results turn out suggested that the OLS and Ridge models performed better in terms of lower root-mean-squared error (RMSE) when there were fewer instrument approach conditions and more predictable use of runway configurations in the dataset. On the other hand, the ensemble-based algorithms, especially Gradient Boosting, appeared to provide the best balance of bias as there was less difference between the training and testing RMSE values, even though the OLS and Ridge models came as the second-best in the cross-validation results with another dataset. Among the examined algorithms, SVR has been found to be the worst in terms of both RMSE and bias/variance balance. His

Table 1
Summary the recent ML-based studies in predicting taxi-out time and flight delay.

Authors and Years	ML Techniques	Airports	Features and Major Outcomes
Diana (2018)	Supervised ML models: OLS regularized/ penalized regressions, SVM and ensemble-based algorithms like RF, Bagging, Adaptive Boosting, Extra Trees and Gradient Boosting (GB).	SEA	Short to medium-term predictions of taxi-out times and variations with five independent operational variables: departure demand, departure throughputs, percentage of total airport capacity utilized, approach conditions and runway configurations.
Herrema et al. (2018)	Comparing ANN, RT, RL and multilayer perceptron (MLP).	CDG	Taxi time prediction and feature importance identification: 10 out of the 42 features, e.g., unimpeded taxi time, congestion level and number of departures in the last 20 min were found as the most influential features.
Herrema et al. (2019a)	GBRT	VIE	Predicting the procedural and non-procedural runway exit; GBRT performs best when a maximum tree depth of 11 and minimum leaf size of 135 are adopted.
Yu et al. (2019a)	DBN-SVN, comparing with k-NN, SVR and LR.	PEK	Novel factors, such as air route situation and crowdedness degree of airports, are of high importance for flight delay prediction accuracy.
Nibareke and Laassiri (2020)	Comparing LR, Naïve Bayes and DT.	9 Airports, including ORD, ATL, SFO, EWR, DEN, IAH, LGA, MIA & BOS	DT is the best, Naïve Bayesian is the worst for flight delay prediction.
Esmaeilzadeh and Mokhtarimousavi (2020)	SVM	Three major New York airports: EWR, JFK & LGA	Features such as pushback delay, taxi-out delay, ground delay program, and demand-capacity imbalance are significantly associated with flight departure delay.
Wang et al. (2021a)	MLP, Linear Regression, Polynomial Regression (PR), GBRT and RF	CDG; MAN; HKG	Among the 33 features, aircraft departure/arrival, sum of taxi distance and number of other aircraft on the way to runway when current aircraft pushes back are the most influential features identified; RF has been found to outperform for taxi time prediction.

Table 2
Description of Independent variables as new features within the CDM framework.

Actual -Estimated Times	Description of KPIs according to CDM framework
AOBT-SOBT	Actual off Block time - Scheduled off block time
AOBT-TOBT	Actual off Block time - Target off block time
AXOT-EXOT	Actual taxi out time - Estimated taxi out time
ASAT-TSAT	Actual start-up approval time - Target start-up approval time
ASAT-ASRT	Actual start-up approval time - Actual start-up request time
TSAT-TOBT	Target start-up approval time- Target off block time
AOBT-ASAT	Actual off Block time- Actual start-up approval time
TOBT-SOBT	Target off block time - Scheduled off block time
ASRT-TSAT	Actual start-up request time- Target start-up approval time
AXIT- EXIT	Actual taxi- in time - Actual taxi- out time
ASRT-TOBT	Actual start-up request time - Target off block time

results also indicate, in the case of SEA, that algorithms fit data differently when comparing two samples. When a model is too complex for a given training dataset, it overfits the training data. Consequently, it cannot generalize well to new, unseen data.

(Herrema et al., 2018) compared ANN, Regression Tree (RT), RL and multilayer perceptron (MLP) methods for the taxi time prediction at Charles de Gaulle Airport (CDG), in which RT turned out to be the most efficient method. In a later study (Herrema et al., 2019), they assessed the Gradient Boosting Regression Tree (GBRT) method on its predictive performance and computational time for predicting the procedural and non-procedural runway exit using data consisting of 54,679 arrival flights at Vienna airport (VIE). The GBRT method they utilized has obtained an accuracy of 79%. They also observed that the GB technique performs best when the maximum tree depth and minimum leaf size are 11 and 135, respectively. Apart from being transparent and interpretable due to the RT, they also concluded that the developed GB method is fast enough to provide prediction updates when new data is extracted from the airport in real-time.

In another more recent study in predicting taxi-out times by Wang et al. (2021a), five models, including MLP, Linear Regression, Polynomial Regression (PR), GBRT and RF, on real data with up to 33 features from three international airports in Europe and Asia: Manchester Airport (MAN) in the UK, Zurich Airport (ZRH) in Switzerland and the Hong Kong International Airport (HKG). Their extensive experiments (totally 10210, 11248 and 33060 track records from the three airports, respectively, after removing any taxi routes longer than 45 min as outliners) demonstrate that the Random Forest model outperforms other prediction models on the datasets from the three airports. In their conclusions, they point out that the range of ML approaches is constantly growing so it would be interesting to evaluate newer regression models, specifically Lasso Regression and Elastic Net Regression. This is one of the research tasks that the current study aims at taking on.

1.4. Flight delay prediction

A practical prediction model that aims to address the airport surface congestion problems by accurately estimating flight delays in real-time is described in (Yu et al., 2019b). With the ultimate goal to enable

airports to collectively alleviate flight delay propagation within their network through collaborative efforts (e.g., synchronizing their delay predictions at regular intervals), such a prediction model is based on a multi-factor (i.e., macro and micro-level) approach using a novel DBN method to mine the inner and deep patterns of flight delays. Their system also embeds SVR into the top layer of the DBN to perform a supervised fine-tuning, as an exemplar use case of a semi-supervised approach using operational data obtained from the Peking International Airport (PEK). Such a DBN-SVR model is also compared its performance based on accuracy, robustness and parameter tuning when compared to three other benchmark methods, namely k-NN, SVR and LR. Forecasting results of the proposed DBN–SVR model indicate 99.3% of the predicted values are within 25 min deviation from the observed value, implying an acceptable prediction performance and satisfying the key requirement of PEK, 98%. The benchmarking results also demonstrated that the proposed approach outperforms the other three methods and provides a high-performance prediction based on large-volume and high-dimensional data. (Nibareke and Laassiri, 2020) compare three ML models, LR, Naïve Bayes and DT, in their predictive accuracy in two distinctive use cases: diabetes and flight delays. In terms of their calculated score and RMSE, DT is the best while Naïve Bayesian gave the worst for both areas.

1.5. Feature importance identification and CDM variables

In addition to forecasting expected taxi-out times and flight delays, many of these studies aim to utilize ML approaches to explore deeper into characterizing delay factors. Their focus is on examining the impacts of various explanatory variables to aid decision-making regarding infrastructural and operational enhancements (Esmaeilzadeh and Mokhtarimousavi, 2020). Generally, linear and penalized models, such as Linear and Ridge regressions, offer clearer insights into the individual impacts of explanatory variables on taxi-out time variations compared to non-parametric ensemble-based learning models like Random Forests and SVM, which might appear opaque (Diana, 2018). In a study by Herrema et al. (2018), out of 42 features, the top 10—such as unimpeded taxi time, congestion levels, and the number of departures in the last 20 min—were identified as the most influential in the feature selection process. Wang et al. (2021a) stress the importance of identifying feature importance to better comprehend which features and to what extent they affect taxi-out times. Their study demonstrates that prediction models for taxi-out times, solely incorporating aircraft departure/arrival, distance, existing departure flights, recent aircraft speeds, and specific features for target airports, yield high accuracy.Wang et al. (2021a) also developed a feature importance identification procedure. Their analysis highlights that aircraft departure/arrival, total taxi distance, and the number of other aircraft on the way to the runway during the current aircraft's pushback are the most crucial features across three airports. These are followed by turning angle totals, straight taxi distances exceeding 500 m, and departure-related speed features. In the flight delay prediction tool developed by Yu et al. (2019a), novel, influential factors, such as air route situation and crowdedness degree of airports, based on the number of flights and passenger flows, were introduced and examined in the previously mentioned multi-factor approach. Their findings underscore the significance of these new factors in accurately predicting flight delays. Despite the introduction of new standards at major airports, CDM continues to enhance efficiency and cooperation among stakeholders in the turnaround process. Consequently, new features have emerged, as shown in Table 2, altering the dynamics of turnaround operations and refining accuracy (Corrigan et al., 2015; Okwir and Correas, 2014; Zuniga and Boosten, 2020a).

In contrast to previous research, feature importance identification within CDM integrates various variables across the system, encompassing all operational aspects of involved actors. CDM features have thus set a new benchmark for optimizing airport operations and resources, benefiting the performance of all stakeholders. Fig. 1 illustrates a schematic representation depicting actors with contextual, contractual, and work-related interdependencies. Turnaround services operate within a network of actors with diverse interests and mutual dependencies, resulting in complex and embedded collaborations.

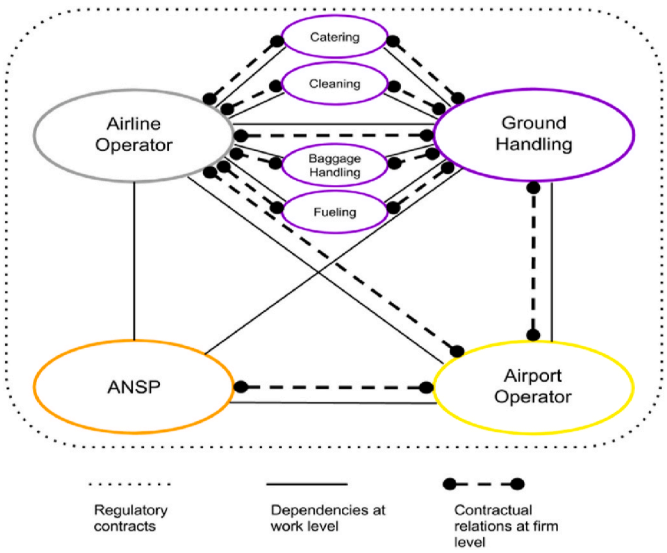

Fig. 1. Airport turnaround activities that are monitored by CDM milestone approach.[11].

2. Method and results

We collected turnaround operations data from Madrid-Barajas Airport after one year it implemented CDM milestone approach. This data was obtained through the airport's information-sharing system, chosen for its intricate nature and the availability of real-time operational data. Our primary source of information was the expertise of 11 CDM actors from the CDM office. The focus of our data is specifically on turnaround movements.

To establish the average number of flights for each month, we opted for a representative single day—specifically, the last day of the month. Our analysis of independent variables centered on the features or attributes outlined in Table 2. The ultimate objective of our study was to predict the "Target off-block time - Scheduled off-block time" for any given flight.

We train different machine learning algorithms to the data set to obtain the most accurate and representative predictive model of CDM data. The total number of observations in the data set approximately 166185 flights, which is enough to present a methodology for predicting off-Block time based on CDM data. All the attributes represent a delay caused by different actors in terms of minutes, consequently, they are on the same scale and there is no need to perform feature scaling. Since most machine learning algorithms cannot work with missing features, the next step is to clean the data set by removing the instances with missing attributes.

Table 2 shows an overview of the data after cleaning, which is reduced to 162955 instances. The last column is the target (dependent) variable and the other six columns show represent the independent variables or features. A histogram plot, in Fig. 2 is another quick way of depicting an overview of each attribute and the target variables, in which the vertical axis shows the number of instances that have a value range given on the horizontal axis.

In the next step the dependent variable (off-Block delay) is categorized into three groups presented in Table 3. According to Eurocontrol, the difference between Scheduled Times (ST) and the Actual Times (AT) in minutes for any activity is considered as a delay. For example, for a flight with a scheduled off-block time set to 8:30, if the actual off-block

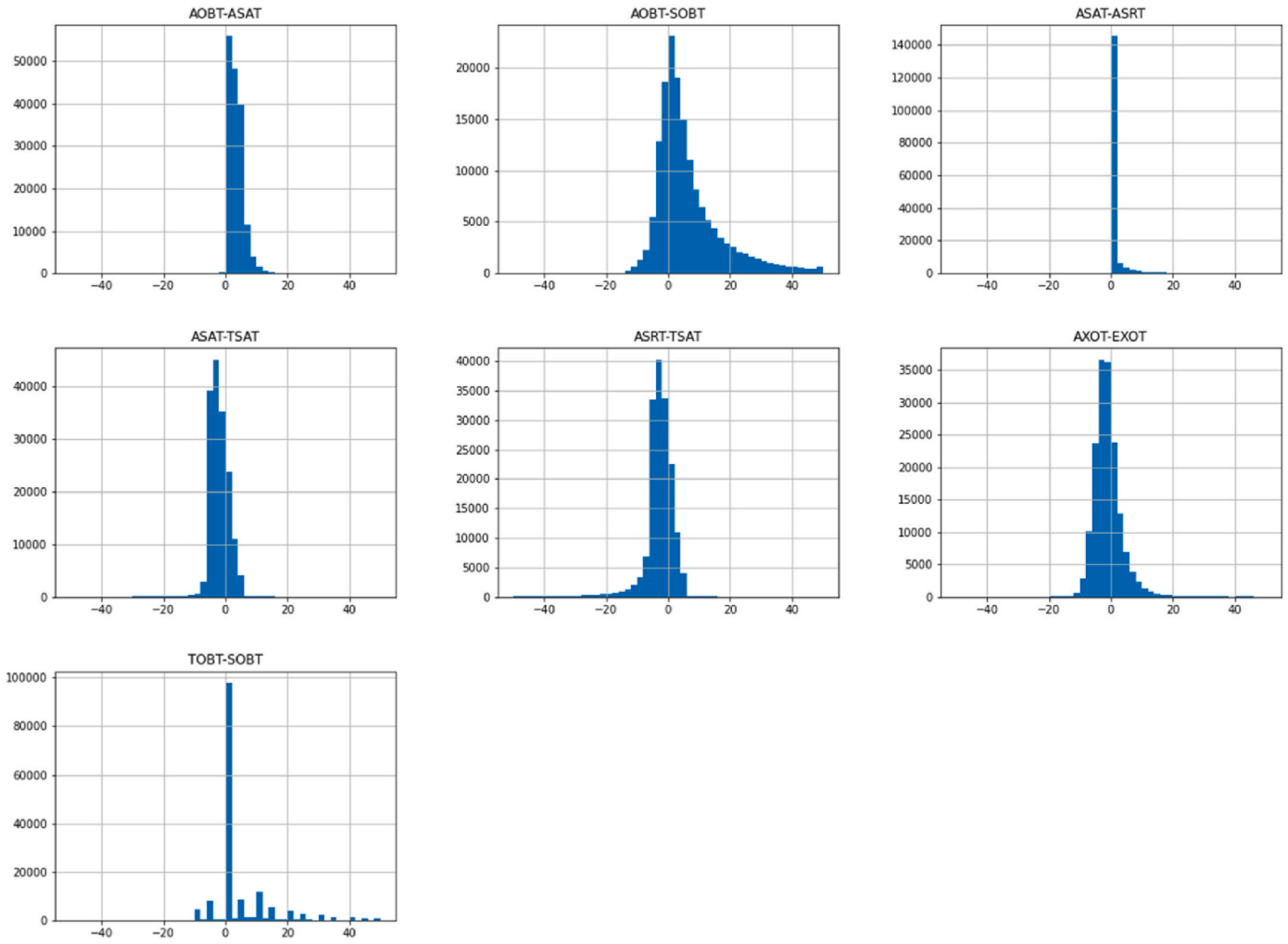

Fig. 2. A histogram of the data set.

Table 3
Summary of the numerical attributes of the data set.

	AOBT-ASAT	AOBT-SOBT	ASAT-ASRT	ASAT-TSAT	ASRT-TSAT	AXOT-EXOT	TOBT-SOBT
count	162955	162955	162955	162955	162955	162955	162955
mean	2.87	9.20	0.89	−2.53	−3.43	−1.43	7.47
std	3.43	22.62	3.93	16.54	17.09	9.60	39
min	−274	−109	−27	−1440	−1442	−68	−10
max	164	790	153	4302	4302	1440	2890

Table 4
Delay categories for the dependent variable.

DEFINITION	CATEGORY IN MINUTES
OVER TIME (OT)	delays < -0.5
ON TIME (T)	$-0.5<$ delays <0.5
EARLIER THAN TIME (ET)	delays <0.5

occurs at 8:35 the delay is categorized as OT (delay = −5 min), and if it occurred at 8:25 the delay is in the ET category (delay = 5 min).

In this study, we use five different machine learning methods to predict the delay categories. The linear method, Regression Tree, Random Forest, Elastic Net and Multi-layer Perceptron (MLP) implemented in the popular Scikit-learn library in Python are used in this study. The hyperparameter of different methods, shown in Table 5, are tuned in two levels. First, a random search with a 3-fold cross-validation evaluation study on a wide range of values for each

Table 5
Hyperparameters for different methods.

Machine learning methods	Hyperparamters
Linear	No parameters
Regression Tree	max_depth; min_samples_leaf; max_features; min_weight_fraction_leaf; max_leaf_nodes
Random Forest	max_depth; min_samples_leaf; n_estimators; min_samples_split
Elastic Net	Alpha; L1_ratio;
Multi-layer Perceptron	Alpha; solver; hidden_layer_sizes; learning_rate

[1] In this figure, ANSP stands for Air Navigation Service Providers.

hyperparameter is performed to narrow the search space. After that, a more focused and thorough grid search with a 3-fold cross-validation method is applied to search and evaluate every combination of hyperparameters values.

Before moving forward to training the data set to different machine learning algorithms, a procedure for comparing the performance of different methods needs to be defined. For this purpose, the data is usually divided into three different sets: training, validation, and test sets. As the names imply, the training set is used to train the models, while the validation set is generally used for the hyperparameter setting of the best models. In the end, the models with the validated hyperparameters are tested against the test to evaluate how they will perform and generalize to new cases. The *generalization error* or *out-of-sample error* is the error rate on new instances, the value shows how well the model predicts the instances it has never seen before. A low training error with a high generalization error indicates that the models in overfitting the training data. However, because the size of the available data set is low and to avoid wasting too much training data on the validation set, the cross-validation method (*k-fold* validation) is used. In cross-validation, the training set is split into *k-subset* or folds, and each model is trained with a different combination of *k-1* folds and validated against the remaining single set. When the hyperparameters are fine-tuned and selected, the models are trained for the final time on the full training set before the generalization error is measured on the test set.

Since the data set is not large enough to include a sufficient number of observations from each category (OT, T, and ET) on both training and test sets, using random sampling methods in creating the two sets might introduce the *significant sampling bias*. For this reason, a stratified sampling method is used, where the data set is divided into homogenous subsets called "strata", in this case, three strata corresponding to the three delay categories, and based on the proportion of the categories in the data set the correct number of instances are sampled from each stratum. This procedure will create a training and a test set that are representative of the original data set in terms of delay categories.

Table 3 shows the percentage of each delay category in the original data, training, and test set using the stratified sampling method. The training and test sets are created with the same percentage of delay categories as in the original data set.

The size of the test set is typically chosen to 20% of the dataset (Table 6).

2.1. Performance metrics

For measuring the performance of different machine learning models, the typical metric used in the field is the root mean square error (RMSE), which is the standard deviation of errors in the prediction of the system. However, in some contexts, the mean absolute error (MAE) might be preferred. The RMSE and MAE are calculated by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (y_i - \widehat{y_i})^2}{m}}$$

$$MSE = \frac{\sum_{i=1}^{m} |y_i - \widehat{y_i}|}{m}$$

Where m is the number of instances in the data set that the metric is measured on and $\widehat{y_i}$ and y_i are the target variable in the data set and the predicted target variable of the i_{th} instance of data.

For calculating the cross-validation error, the Scikit-Learn library in Python uses a utility function denoted by Score, where the Score for each of the k-folds and the mean and the standard deviation of all folds are measured. Table 7 and Fig. 2 summarize the RMSE scores from the 10-fold cross-validation for all methods. Based on the Score results we can observe that the linear, random forest and elastic net method yield a lower error, followed by MLP with a slightly higher error and regression tree with the least accurate. However, the standard deviation of Scores, representing the spread of the average Scores across the 10 folds, shows an opposite ranking of different methods with the regression tree having the lowest deviation in prediction error and the linear, random forest and elastic net showing the higher spread in errors.

However, analyzing the regression errors might not be a suitable metric for evaluating the performance of different regression methods when used in a problem with a classification nature. The main aim of this comparison is to identify the best machine learning method for predicting the categorical delay, instead of the delay values in minutes.

Therefore, the predicted dependent variable can be converted to one of the categories based on the predicted value. Consequently, the accuracy of the prediction models can also be evaluated by converting the predicted dependent variable to the corresponding category (OT, T, ET) and employing classification performance metrics.

2.2. Accuracy

The accuracy of classification predictions can be obtained by calculating the ratio of correct predictions. However, this measure is not preferred when some classes occur much more frequently than others, or the dataset is skewed. The ratio of correct predictions in percentage for all ML methods is shown in Fig. 3. The regression tree predicted the correct categorical delay in 81% of the test set, followed by random forest with 62%, MLP with 52%, and linear and elastic net with 53%. Now by comparing the accuracy in terms of the percentage of correct categorical prediction (Fig. 4) with the Score error metric (Table 7), one can observe the difference in the performance of different methods. However, since most of the data is in the "T" delay class (Table 4), the accuracy metric is not a good representative of the performance in classification problems.

2.3. Confusion matrix

The performance of a classifier model is evaluated more effectively by using the confusion matrix. The confusion matrix calculates how many times a class or category is confused with another category. For example, how many times a "T" class delay is confused with an "ET" category in the prediction. The confusion matrix is a binary performance metric, therefore for each category, a confusion matrix is calculated, by using the multi-label confusion matrix function in the Scikit-Learn library.

An illustration of confusion matrix for a two-category classification problem (positive and negative) is shown in Fig. 5.

Table 6
Percentage of each delay category in different data sets by using the stratified sampling method.

	OT	T	ET
Original data set	8.7	59.63	31.67
Training set	8.7	59.47	31.83
Test set	8.77	59.6	31.63

Table 7
Summary of the RMSE scores from 10-folds Cross-validation.

RMSE score	Linear	Regression Tree	Random Forrest	Elastic Net	MLP
Mean	32.12	40.07	32.19	32.12	33.38
Std. dev.	10.54	8.75	10.53	10.53	9.01

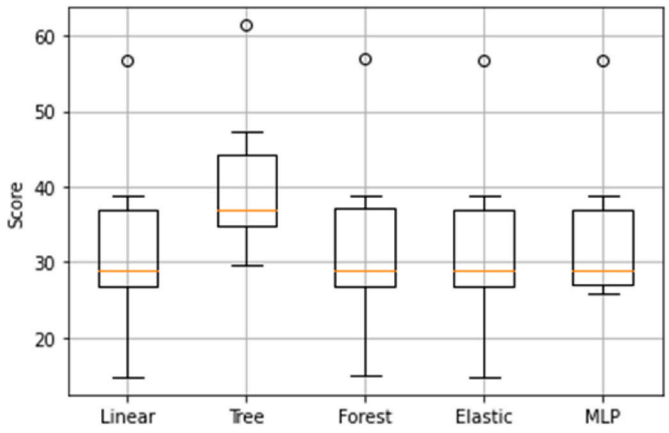

Fig. 3. RMSE score from 10-fold cross-validation.

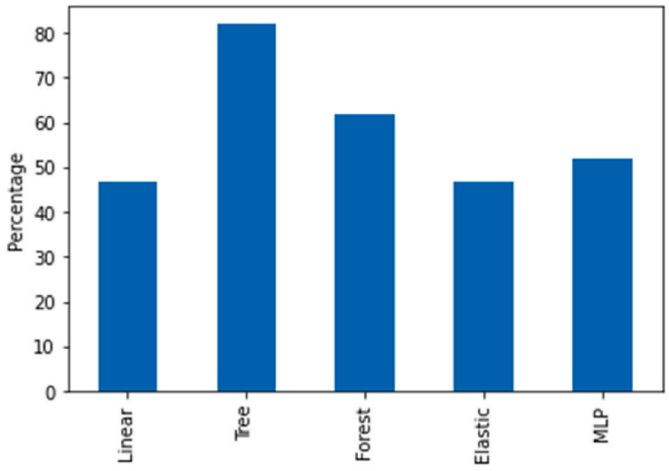

Fig. 4. Percentage of correct categorical prediction on the test set.

		Actual	
		Positive	***Negative***
Predicted	***Positive***	**True Positive (TF)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

Fig. 5. Representation of confusion matrix for a two a category (positive, negative) classification.

2.4. Precision and recall

More concise measures for the performance of classifiers are *precision* (also called positive predicted value) and *recall* (called sensitivity or true positive rate) calculated by:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

where TP, FP and FN are the number of true positives, false positives and false negatives, correspondingly.

2.5. F-score

Precision and recall form a trade-off in which higher-level information from decision-makers is required to interpret and select a proper model. Therefore, the two aforementioned metrics are combined into a single metric with a parameter to set the trade-off threshold, namely F_β score, which can be interpreted as a weighted harmonic mean of the precision and recall. The F_β score is calculated by

$$F_\beta = (1+\beta)^2 \frac{precision.recall}{(\beta^2.precision) + recall}$$

in which β is the factor that weights recall more than precision. For giving equal importance to precision and recall β is set to one ($\beta = 1$) and F_1 score is calculated by

$$F_1 = 2\frac{precision.recall}{precision + recall}$$

Table 8 shows the precision, recall and F1 score of the test set for all ML methods. A decision-maker can now select the best performing method based on their requirements by studying Table 8. If predicting the on-time (T) flights precisely is the most important criteria for a decision-maker, the regression tree with 79% precision is the best method. On the other hand, if predicting the lowest proportion of OT or AT flights falsely as on time is the focus, the linear, elastic net and MLP have the high recall percentage of 95%. In general, if both precision and recall of on-time flight predictions are desirable, then by looking at the F_1 score in the table, one can conclude that the regression tree is the most suitable method. A similar type of analysis can be performed for other delay categories based on the data presented in Table 8.

2.6. Feature importance analysis of independent variables with SHAP method

We utilized the SHapley Additive exPlanations (SHAP) method to ascertain the feature importance of each independent variable in our machine learning models. This method, is based on the concept of Shapley values from cooperative game theory. The method facilitates the interpretation of model outputs. SHAP values elucidate the contribution of each feature to a specific prediction, thereby enhancing the interpretability of complex models (Ekanayake et al., 2022).

Fig. 6 illustrates the dispersion of SHAP values across various features for five distinct machine learning models. The x-axis in each plot denotes the SHAP values, reflecting the influence of individual features on the model's predictions. The y-axis lists the features used in our model, comprising eight independent variables, with each feature corresponding to a row of dots in the plots. A comparative analysis of features across these plots reveals that, with the exception of the Multilayer Perceptron (MLP), the ASRT-TSAT feature is uniformly acknowledged as having a significant impact on variability with delay by all models. Additionally, an evaluation of consistency across all models indicates that all features exhibit uniformly high SHAP values, with magnitudes ranging from −15 to 25 for the MLP model, suggesting their potential as strong predictors. This consistency also implies a stable relationship of these features with the dependent variable. Notably, certain features such as ASRT-TSAT, ASAT-ASRT, and ASAT-TSAT maintain identical positions across multiple model plots, while other features such as AOBT-SOBT, AOBT-ASAT, and AXOT-EXOT show a similar, thinly distributed positioning across various models, including Linear, Tree, Forest, and Elastic models. This consistency signifies a robust predictive power across these models. The MLP model, however, exhibits a unique range of SHAP values across features, indicating more intricate interactions and non-linear relationships within the model.

Additionally, from the analysis of the five models, we have determined that Linear Regression, Random Forest, and Elastic Net models exhibit enhanced proficiency in predicting delay categories within the

Table 8
Precision, recall and F-score on the test set.

	Precision			Recall			F_1 score		
	OT	T	ET	OT	T	ET	OT	T	ET
Linear	0.931	0.13	0.98	0.26	0.95	0.52	0.40	0.22	0.68
Regression Tree	0.81	0.79	0.87	0.74	0.90	0.73	0.77	0.84	0.80
Random Forrest	0.777	0.40	0.99	0.90	0.94	0.47	0.83	0.56	0.63
Elastic Net	0.93	0.13	0.98	0.26	0.95	0.51	0.40	0.22	0.68
MLP	0.85	0.22	0.99	0.68	0.95	0.42	0.75	0.36	0.59

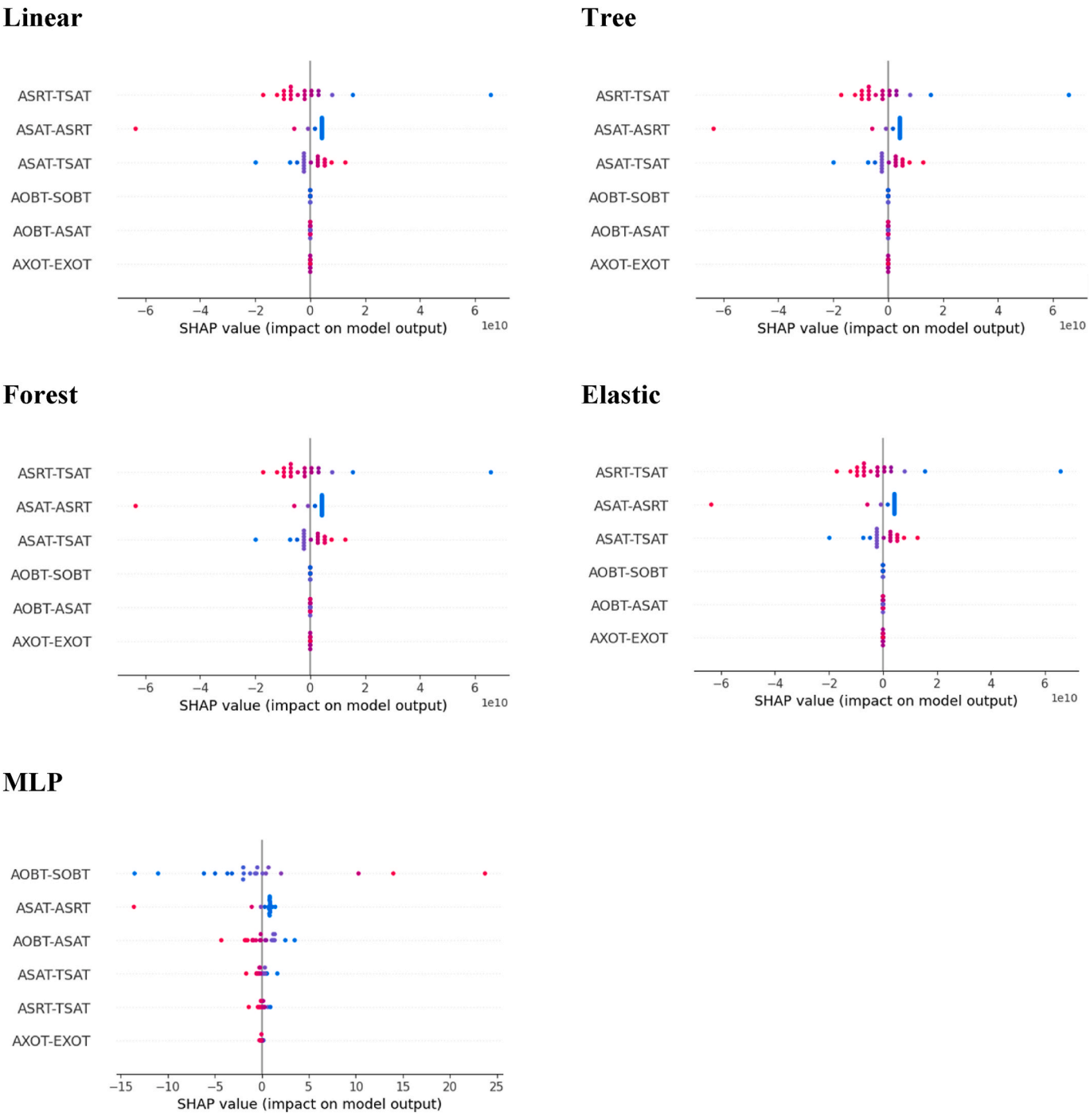

Fig. 6. SHAP plots of the five models showing feature value importance in all our independent variables.

CDM framework. This finding is evidently seen by the RMSE scores presented in Table 7, where these models demonstrate lower scores, indicating better performance. However, despite its low RMSE score, the Random Forest model has been excluded from further consideration. Although it shows promising predictive accuracy, the Random Forest model requires significantly higher computational time, which could necessitate elevated energy consumption. This trade-off between accuracy and computational efficiency is a crucial factor in our decision to exclude it from our preferred set of models. As such, after a thoroughly evaluation of our findings, one can conclude that Linear Regression and Elastic Net are the most preferred machine learning models for achieving high prediction accuracy with airport data within the CDM framework.

Building on prior research, like Yu et al. (2019a) who introduced influential factors assessing air route conditions and airport congestion levels based on flight and passenger data, and Wang et al. (2021a) who underscored the significance of understanding feature importance for effective decision-making in airport enhancements and operational efficiency (Esmaeilzadeh and Mokhtarimousavi, 2020), our study takes a different path. We concentrate on pinpointing important features within the CDM framework, integrating diverse variables across turnaround operations. Our research highlights specific features within turnarounds that display substantial variability, offering insights into managing them for continuous improvement and better predictive accuracy. Table 9 delineates these selected CDM performance features, providing actionable information for pre-planning.

2.7. Elastic model: Feature importance analysis

After evaluating the performance of various predictive models utilizing CDM data, we identified the elastic model as the superior predictor (refer to Figs. 4 and 6). To further assess its robustness, we extended our analysis by applying the elastic model to predict taxi times using an alternative dataset. This involved conducting supplementary analyses and examining the impact of employing the elastic machine learning model on taxi time prediction. For this extended analysis, we compiled taxi times involving four runways for arrivals and 8 runways for departures. The eight runways in second dataset are described as follows: 18L, 18R, 32L, 32R, and 14L for arrivals; and 14R, 36L, and 36R for departures. (R stands for right lanes and L stands for left lanes). This dataset is publicly available in the attached appendix. The new dataset comprised actual taxi-in times (AXIT) and actual taxi-out times (AXOT), along with their scheduled (expected) times, enabling the calculation of prediction errors. The data was randomly selected from flights, including their corresponding runway allocations, resulting in a dataset of over 15,000 observations.

In exploring the associations within our data, we applied the optimal machine learning model, identified as the elastic model, for training purposes. Utilizing a training dataset consisting of 4290 flights with taxi time information, we assessed the robustness and generalization performance of our model by incorporating three independent variables: TSAT, ASRT, and AXOT. Additionally, we introduced a star variable, following Okwir et al. (2017), to indicate delay categories, as presented in Table 4, which represents the optimal condition with no delays.

To minimize error in our new training set, we employed a function that generated random numbers with a mean close to zero minutes, thereby creating a dataset for training. Cross-validation was implemented by partitioning the dataset into multiple folds. The elastic model, identified as the best-performing, was then used to predict times with a different test set comprising 2964 instances. In Fig. 7a, we present data for 8 different runways and their associated taxi times during both arrival and takeoff. A higher standard deviation compared to the mean indicates increased variability in taxi times. However, in Fig. 7b, after applying the elastic prediction, the standard deviation relative to the mean is considerably reduced. This highlights the effectiveness of the elastic model in minimizing variability in taxi times.

These findings suggest that at CDM airports, the elastic model can be leveraged to predict scheduled times in different parts of the turnaround process, thereby improving delay conditions.

Furthermore, for the training process, the dataset was partitioned, allocating 20% to the test set and the remaining 80% to the training set. This partitioning procedure was iteratively performed eight times, resulting in the creation of eight distinct training subsamples. Each of these subsamples then underwent a training regimen, iterating over the training procedure 50 times. The learning curve, depicted in Fig. 8, provides a visual representation of the model's performance throughout this iterative training process. It validates the findings presented in Fig. 7b, offering insights into how the model evolves and refines its predictive capabilities over multiple training iterations.

Table 9
Selected CDM performance features for prediction accuracy [1,2].

Performance driver	Performance indicator	Feature analysis
Optimize the turn-round time predictability	Turn-round compliance	Measure ATTT vs. (SIBT-SOBT) Compare MTTT to ATTT
Improve the ARR predictability	EIBT predictability Estimated Landing Time predictability	Measure EIBT Measure Estimated Landing Time
Improve the departure predictability	TOBT accuracy & predictability	Compare TOBT to ARDT
Improve the departure predictability	TSAT accuracy & predictability	Compare TSAT to AOBT
Reduce average delay of departure flights	Average delay of departure flights	Compare AOBT to SOBT
Reduce noise on ground and emission reduction	Group engine time	EXIT vs. AXIT EXOT vs. AXOT

3. Discussion: implications for practice and future work

We started this paper with the purpose of exploring how five different machine learning methods—the linear method, Regression Tree, Random Forest, Elastic Net, and Multi-layer Perceptron (MLP)—predict delay categories within the CDM framework. We argued that, although previous research has demonstrated the impact of various models on prediction in different airport segments (Lian et al., 2018; Wang et al., 2021a; Yu et al., 2019b), many studies still lack the application of machine learning models, particularly those incorporating CDM performance variables when describing delay conditions. Therefore, we expanded upon this work to advance current knowledge regarding prediction accuracy and how different machine learning methods enhance prediction accuracy within the CDM turnaround operations.

Furthermore, while other studies focus on predictions influenced by external factors like weather (Liu et al., 2019) or runway prediction (Herrema et al., 2019a), our study specifically focuses on prediction within the CDM milestone approach, which segments the turnaround process using standard KPIs. Our results as shown in Fig. 7A and B, indicate that the Elastic Net model is the most effective machine learning model for achieving high prediction accuracy within the CDM framework. Below, we present implications for practice.

Our results contribute to the understanding of the current usage of the CDM milestone approach for airport managers. Following the CDM implementation manual, we selected important features, as shown in Table 9, and below we explain how enhanced prediction accuracy for turnaround operations impacts delay conditions for taxi times.

Firstly, within the CDM framework, TOBT/TSAT is a measure aiding the integration of on-time performance into other airports, although the accuracy of TOBT/TSAT remains an unresolved topic. Our results demonstrate that utilizing machine learning leads to greater accuracy, enabling the prediction of network imbalances and more rational preplanning for all actors, both at the local airport and destination

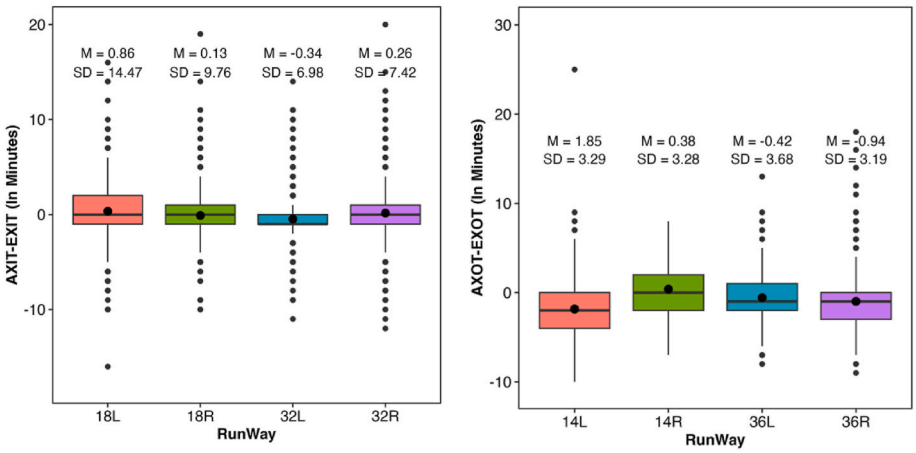

Fig. 7a. The figure illustrates the variability in taxi times upon arrival and departure for selected runways, showing deviations from the scheduled predicted times to actual taxi times.

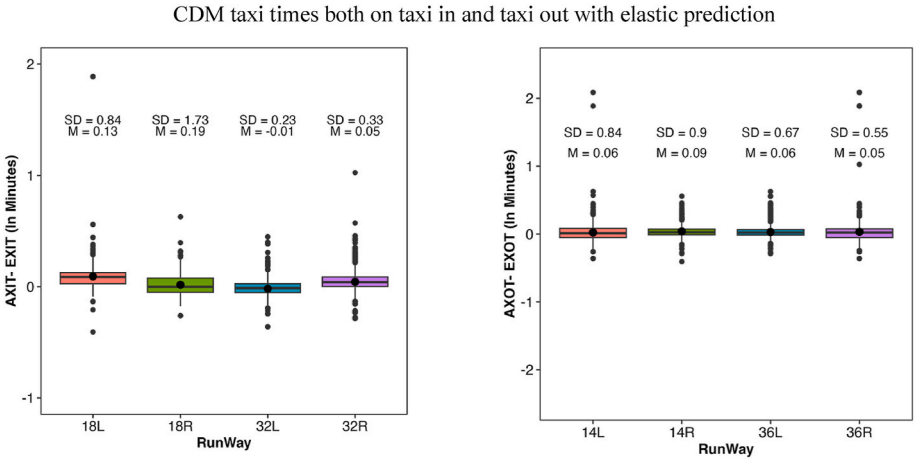

Fig. 7b. The figure demonstrates a decrease in error, leading to a reduction in delay conditions, achieved through the implementation of the elastic prediction model.

airports. This results in a more rational Dynamic Demand & Capacity Balancing (dDCB) (Schefers et al., 2020). Consequently, accurate prediction in each turnaround segment enables airport actors to estimate performance between actual operational capacity and declared capacity within their operations. Apart from TOBT prediction accuracy, activities collaboratively generated by several actors can also be evaluated by adopting the best-performing model. Our results contribute to automation structures in ground handling operations, as proposed by Alonso Alonso Tabares et al., 2021a.

Secondly, to the best of our knowledge, this is the first paper using turnaround data from a CDM airpot to apply machine learning for prediction. Our results contribute by showcasing a training methodology using the elastic model as the best-performing model. In the training phase, we partitioned the dataset, allocating 20% to the test set and reserving the remaining 80% for the training set. Our contribution lies in the nature of our training process, where we not only test different models but also extend this analysis to offer valuable insights into the dynamics of overall error reduction and performance improvement achieved through our chosen methodology. With error reduction, there is less waiting time, lower CO2 emissions, and reduced fuel costs.

Thirdly, the primary strategic objective of every airport is to enhance operational efficiency. When evaluating success following the implementation of airport CDM, accuracy plays a pivotal role in distinguishing the performance of various airport stakeholders and the overall system. Previous research has proposed two approaches to boost collaborative performance at CDM airports. The first involves reviewing agreed-upon objectives with all CDM partners, defining a set of Key Performance Indicators (KPIs), or collectively developing new ones. The second approach is to establish pre-CDM operations as a basis for comparison, creating a performance baseline, and fostering a no-blame culture that includes reporting and feedback mechanisms for all stakeholders. To our knowledge, shared objectives for establishing performance indicators have been categorized into four main areas applicable to all CDM airport partners: Efficiency, Environment, Capacity, and Safety. These measures aim to eliminate bias and variance associated with machine learning while linking each airport's indicators to a common objective (Diana, 2018c). However, these areas lacked a basis for continuous improvement. Our results, driven by machine learning predictive accuracy, demonstrate that airport CDM operations can function as a learning system. Machine learning enables continuous

[2] **Notes**TOBT- Target Off-Block Time (The time that an Aircraft Operator or Ground Handler estimates that an aircraft will be ready, all doors closed, boarding bridge removed, push back vehicle available and ready to start up/push back immediately upon reception of clearance from the TWR)ATTT- Actual Turn-round Time which is expressed as AOBT – AIBTARDT- Actual Ready Time for movement - When the aircraft is ready for start up/push back or taxi immediately after clearance delivery, meeting the requirements set by the TOBT definitionELDT- Estimated Landing Time (The estimated time that an aircraft will touch-down on the runway. (Equivalent to ATC ETA –Estimated Time of Arrival = landing)MTTT- Minimum Turn-round Time (The minimum turn-round time agreed with an AO/GH for a specified flight or aircraft type).

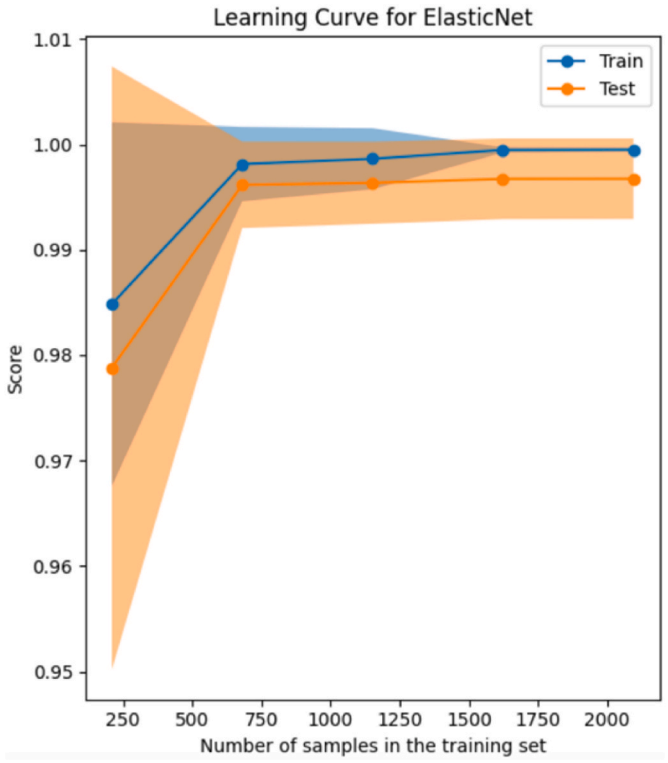

Fig. 8. Training learning curve.

monitoring and facilitates adaptations based on high-accuracy data in real time, influencing stakeholder behavior. Consequently, continuous improvement programs can be seamlessly implemented to allow the flow of operations in the turnaround. We therefore propose by consulting the CDM implementation manual the most critical drivers (refer to Table 9) that enhance accuracy when utilizing machine learning methods.

4. Future work

Our findings demonstrate how different machine learning techniques—such as linear regression, regression trees, random forests, elastic nets, and multi-layer perceptrons (MLP)—predict delay categories within the CDM framework. Our study highlights several crucial areas for future research. First, upcoming work could compare non-CDM airports with our prediction model. Using similar models to better understand additional features, such as TOBT efficiency at CDM airport is crucial. Secondly, we have illustrated how machine learning can be applied to CDM operational data by categorizing the dependent variable (off-block delay) into three groups, as shown in Table 3. Optimizing this approach could test how effectively airport capacity can be fully utilized with high accuracy and efficiency during turnaround operations. Lastly, future research could explore enhancing these models to account for adverse weather conditions, such as snow delay, bird strike delays, and other non performance related features, in order to maintain prediction accuracy. In Table 9, we present key peformance indicators that are commonly used during delay reduction performance evalations according to the CDM implementation manual.

CRediT authorship contribution statement

Simon Okwir: Writing – review & editing, Writing – original draft, Second data set anaysis, Project administration, Formal analysis, Data curation, Conceptualization. **Kaveh Amouzgar:** Validation, Software, Methodology. **Amos HC. Ng:** Writing – review & editing, Validation, Supervision, Conceptualization.

Data availability

This appendix provides the necessary materials to replicate the results presented in our paper. Three PDFs are attached: Stage 1_Analysis.pdf, Stage 2_Analysis.pdf, and Stage 3_Analysis.pdf.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jairtraman.2024.102684.

References

AlMashari, R., AlJurbua, G., AlHoshan, L., Al Saud, N.S., BinSaeed, O., Nasser, N., 2018. IoT-based Smart airport solution. 2018 International Conference on Smart Communications and Networking (SmartNets), pp. 1–6. https://doi.org/10.1109/SMARTNETS.2018.8707393.

Alonso Tabares, D., Mora-Camino, F., Drouin, A., 2021. A multi-time scale management structure for airport ground handling automation. J. Air Transport. Manag. 90, 101959. https://doi.org/10.1016/j.jairtraman.2020.101959.

Chung, S.-H., Ma, H.-L., Hansen, M., Choi, T.-M., 2020. Data science and analytics in aviation. Transport. Res. E Logist. Transport. Rev. 134, 101837. https://doi.org/10.1016/j.tre.2020.101837.

Corrigan, S., Mårtensson, L., Kay, A., Okwir, S., Ulfvengren, P., McDonald, N., 2015. Preparing for airport collaborative decision making (A-CDM) implementation: an evaluation and recommendations. Cognit. Technol. Work 17 (2). https://doi.org/10.1007/s10111-014-0295-x.

Diana, T., 2018. Can machines learn how to forecast taxi-out time? A comparison of predictive models applied to the case of Seattle/Tacoma International Airport. Transport. Res. E Logist. Transport. Rev. 119, 149–164. https://doi.org/10.1016/j.tre.2018.10.003.

Ekanayake, I.U., Meddage, D.P.P., Rathnayake, U., 2022. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Stud. Constr. Mater. 16 (April), e01059. https://doi.org/10.1016/j.cscm.2022.e01059.

Esmaeilzadeh, E., Mokhtarimousavi, S., 2020. Machine learning approach for flight departure delay prediction and analysis. Transport. Res. Rec.: J. Transport. Res. Board 2674 (8), 145–159. https://doi.org/10.1177/0361198120930014.

Eurocontrol, 2006. Airport CDM implementation. In: Eurocontrol. Brussels, Belgium.

Greasley, A., 2019. Simulating Business Processes for Descriptive, Predictive, and Prescriptive Analytics. Walter de Gruyter GmbH & Co KG.

Guo, X., Grushka-Cockayne, Y., De Reyck, B., 2020. London Heathrow airport uses real-time analytics for improving operations. INFORMS Journal on Applied Analytics 50 (5), 325–339. https://doi.org/10.1287/inte.2020.1044.

Herrema, F., Curran, R., Hartjes, S., Ellejmi, M., Bancroft, S., Schultz, M., 2019. A machine learning model to predict runway exit at Vienna airport. Transport. Res. E Logist. Transport. Rev. 131, 329–342. https://doi.org/10.1016/j.tre.2019.10.002.

Herrema, F., Curran, R., Visser, H., Huet, D., Lacote, R., 2018. Taxi-Out Time Prediction Model at Charles de Gaulle Airport. J. Aero. Inf. Syst. 15 (3), 120–130. https://doi.org/10.2514/1.I010502.

Lee, S., Miller, S.M., 2019. AI gets real at Singapore's Changi Airport (Part 1).

Lemetti, A., Hardell, H., Polishchuk, T., 2023. Arrival flight efficiency in pre- and post-Covid-19 pandemics. J. Air Transport. Manag. 107, 102327. https://doi.org/10.1016/j.jairtraman.2022.102327.

Lian, G., Zhang, Y., Desai, J., Xing, Z., Luo, X., 2018. Predicting taxi-out time at congested airports with optimization-based support vector regression methods. Math. Probl Eng. 2018, 1–11. https://doi.org/10.1155/2018/7509508.

Lin, P.-C., 2023. The propagation of European airports' on-time performance and on-time flights via air connectivity prior to the Covid-19 pandemic. J. Air Transport. Manag. 109, 102382. https://doi.org/10.1016/j.jairtraman.2023.102382.

Liu, Y., Liu, Y., Hansen, M., Pozdnukhov, A., Zhang, D., 2019. Using machine learning to analyze air traffic management actions: ground delay program case study. Transport. Res. E Logist. Transport. Rev. 131, 80–95. https://doi.org/10.1016/j.tre.2019.09.012. December 2018.

Lohman, C., Fortuin, L., Wouters, M., 2004. Designing a performance measurement system: a case study. European journal of operational research 156 (2), 267–286.

Michelmann, J., Schmalz, U., Becker, A., Stroh, F., Behnke, S., Hornung, M., 2023. Influence of COVID-19 on air travel - a scenario study toward future trusted aviation. J. Air Transport. Manag. 106, 102325. https://doi.org/10.1016/j.jairtraman.2022.102325.

Nibareke, T., Laassiri, J., 2020. Using Big Data-machine learning models for diabetes prediction and flight delays analytics. Journal of Big Data 7 (1). https://doi.org/10.1186/s40537-020-00355-0.

Okwir, S., Correas, A., 2014. Collaborative decision making (CDM) in airport surface: Europe vs USA implementations, challenges and best practices. In: Integrated Communications, Navigation and Surveillance Conference. https://doi.org/10.1109/ICNSurv.2014.6819989. ICNS.

Okwir, S., Ulfvengren, P., Angelis, J., Ruiz, F., Núñez Guerrero, Y.M., 2017. Managing turnaround performance through collaborative decision making. J. Air Transport. Manag. 58. https://doi.org/10.1016/j.jairtraman.2016.10.008.

Rajapaksha, A., Jayasuriya, Dr N., 2020. Smart airport: a review on future of the airport operation. Global J. Manag. Bus. 25–34. https://doi.org/10.34257/GJMBRAVOL20IS3PG25.

Schefers, N., Amaro Carmona, M.A., Ramos González, J.J., Saez Nieto, F., Folch, P., Munoz-Gamarra, J.L., 2020. STAM-based methodology to prevent concurrence events in a Multi-Airport System (MAS). Transport. Res. C Emerg. Technol. 110 (August 2019), 186–208. https://doi.org/10.1016/j.trc.2019.11.012.

Thiagarajan, B., Srinivasan, L., Sharma, A.V., Sreekanthan, D., Vijayaraghavan, V., 2017. A machine learning approach for prediction of on-time performance of flights. 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), pp. 1–6. https://doi.org/10.1109/DASC.2017.8102138.

Wang, X., Brownlee, A.E.I., Woodward, J.R., Weiszer, M., Mahfouf, M., Chen, J., 2021a. Aircraft taxi time prediction: feature importance and their implications. Transport. Res. C Emerg. Technol. 124, 102892. https://doi.org/10.1016/j.trc.2020.102892.

Wang, X., Brownlee, A.E.I., Woodward, J.R., Weiszer, M., Mahfouf, M., Chen, J., 2021b. Aircraft taxi time prediction: feature importance and their implications. Transport. Res. C Emerg. Technol. 124. https://doi.org/10.1016/j.trc.2020.102892.

Yu, B., Guo, Z., Asian, S., Wang, H., Chen, G., 2019a. Flight delay prediction for commercial air transport: a deep learning approach. Transport. Res. E Logist. Transport. Rev. 125, 203–221. https://doi.org/10.1016/j.tre.2019.03.013.

Yu, B., Guo, Z., Asian, S., Wang, H., Chen, G., 2019b. Flight delay prediction for commercial air transport: a deep learning approach. Transport. Res. E Logist. Transport. Rev. 125, 203–221. https://doi.org/10.1016/j.tre.2019.03.013. October 2018.

Zuniga, C., Boosten, G., 2020. A practical approach to monitor capacity under the CDM approach. Aerospace 7 (7). https://doi.org/10.3390/AEROSPACE7070101.

Simon Okwir is a researcher and lecturer at Uppsala University, Sweden. Simon holds a double Ph.D. in Industrial Management and Economics from KTH-Kungliga Tekniska Hogskolan in Sweden and UPM -Universidad Politécnica de Madrid in Spain. Research focuses on Airport operations & Technology in organizations. Simon has worked with Madrid airport and IBERIA airlines and published several works on the CDM implementation.

Kaveh Amouzgar is an Assistant Professor at the Division of Industrial Engineering and Management, Uppsala University, Sweden. He is also a part-time Assistant Professor at the Division of Intelligent Production Systems, University of Skövde, Sweden. His research interests include the application of machine learning, simulation-based multi-objective optimization and extended reality (XR) in different industrial fields.

AMOS HC NG is a Professor of Automation Engineering with the Division of Intelligent Production Systems, University of Skövde, Sweden. He is also a Visiting Professor at the Division of Industrial Engineering and Management, Uppsala University, Sweden, and the CEO of Evoma AB. His research interests include the applications of simulation, multi-objective optimization, and artificial intelligence in production/health-care/business systems. His email address is amos.ng@his.se.