

A Generalisable Machine Learning Framework for Taxi Time Prediction at A-CDM Airports

Angelika Swatowska, Lucie Gabagnou
EUROCONTROL Innovation Hub (EIH)
Brétigny-sur-Orge, France

Abstract—Accurate taxi time estimates are central to the Airport Collaborative Decision Making (A-CDM) framework, where the sharing of timely information underlies network efficiency and predictability. Many airports still rely on historical averages as operational baselines. However, these values fail to capture the variability inherent to daily operations. This paper presents a generalisable machine learning model based on Gradient Boosted Decision Trees (GBDT) for both taxi-out (AXOT) and taxi-in (AXIT) operations, trained on 4.1 million flights (2022–2024) across six European airports, of which five are fully implemented under A-CDM. The model draws on standard airport data sources, combining AODB operational airport records with METAR weather observations and aircraft characteristics from the Base of Aircraft Data (BADA). Evaluation against operational baselines (Standard Taxi Times) shows MAE reductions of 30–44% at regional airports and 1–6% at major airports, with typical mean absolute percentage errors of 17–20%. Feature importance analysis confirms the crucial role of airport infrastructure and operational factors, supporting the cross-applicability of this approach with minimal local calibration. The approach offers a practical path to improving the accuracy of shared operational information, strengthening both local decision support and network-wide flow management.

Keywords—Air Traffic Control; A-CDM; airport operations; network predictability; taxi time; AODB data; machine learning; gradient boosting

I. INTRODUCTION

As of 2025, 34 European airports are fully integrated into the Airport Collaborative Decision Making (A-CDM) framework, and together with Advanced ATC-Tower airports, they now account for nearly 50% of all departures in the EUROCONTROL Network Manager (NM) area [1], [2]. A-CDM is a structured process that tries to improve network predictability by sharing the key milestones between the stakeholders and the NM [3]. For departures, these milestones include the Target Off-Block Time (TOBT), Target Start-Up Approval Time (TSAT), and Target Take-Off Time (TTOT); for arrivals, they include the Estimated Landing Time (ELDT) and Actual Landing Time (ALDT) [4]. Accurate taxi time predictions are needed because departure and arrival taxi times directly determine when an aircraft will be ready for take-off or reach the parking stand. Taxi time estimation is thus crucial to local surface operations—in particular, sequencing, runway and stand allocation—and the wider ATM network, where it underpins demand–capacity balancing and flow management.

Despite its importance, many airports, including A-CDM airports, still rely on reference taxi times derived from historical averages [5]. While these nominal values provide a convenient baseline, they fail to capture day-to-day variability caused by factors such as runway configuration changes, adverse weather, de-icing needs, or surface congestion. As a result, inaccurate taxi time estimates propagate through milestone updates and Departure Planning Information (DPI) messages, reducing network-wide predictability and constraining ATM efficiency.

This paper addresses the challenge of enhancing taxi time prediction by presenting a generic machine learning model developed and evaluated across six European airports. Five of these—Brussels (EBBR), Prague (LKPR), Paris Orly (LFPO), Vienna (LOWW) and Zurich (LSZH)—are A-CDM airports, while Istanbul (LTFM) is currently undergoing A-CDM deployment¹. The key novelty of this work lies in its adaptability to different airport environments, enabled by reliance on standardised airport data flows such as the Airport Operational Database (AODB). Whereas many existing approaches are tailored to a single airport and require extensive calibration, the proposed model is designed to be adaptable with minimal configuration, making it well suited for large-scale adoption within the A-CDM community. By leveraging a shared data foundation, the model strengthens both local decision support and network-wide predictability.

The paper is organized as follows: Section II provides a literature review on taxi times predictions in existing work, section III introduces the methodology, including the dataset and model design. Section IV presents the results as well as a discussion regarding their relevance. Finally, section VI concludes with a discussion on the findings and directions for future work.

II. LITERATURE REVIEW

An aircraft's taxi time—whether taxi-out (AXOT, from stand to take-off) or taxi-in (AXIT, from runway touchdown to stand)—is affected by numerous operational conditions. Runway configuration, surface congestion, weather disruptions, and the need for remote de-icing are the most obvious

¹Istanbul (LTFM) airport is currently in an "ongoing" A-CDM deployment phase, including the implementation of DPI messages exchange with the NM and the application of A-CDM specifications. Source: <https://www.eurocontrol.int/concept/airport-collaborative-decision-making>



[3], [6], [7], since de-icing operations often require off-route taxiing and generate additional queues. Airline-specific taxiing policies can also have an impact [8]. Some operators aim to minimize taxiing duration through route optimization in coordination with ATC, while others prefer fuel-saving techniques such as single-engine taxiing, which may lengthen taxi time [9]–[12]. Aircraft characteristics matter as well: heavier types generally require more thrust when maneuvering, which can lead to longer times on turns or when accelerating after stops [13], [14]. Since such policies and performance effects are not explicitly available in the data, they are typically approximated through proxy features such as operator, aircraft type, or weight class [15], [16]. These features, while simple to derive from tabular AODB data, capture part of the variability otherwise hidden in averages.

Early work on taxi time estimation relied primarily on statistical values, such as historical averages or percentiles [16]. These nominal, or unimpeded, times established a baseline but ignored day-to-day operational variability. EUROCONTROL's Additional Taxi-Out Time KPI [17] formalized the distinction between unimpeded and variable components of taxi times resulting from congestion, providing a framework that has become standard practice. While useful for benchmarking, this method largely overlooks operational variability.

In response, researchers turned to simulation and analytical models. Microscopic simulators such as the Surface Operations Simulator and Scheduler (SOSS) were able to reproduce observed taxi times in great detail [18]. However, the computational and data demands made them impractical for operational use. Queueing-based models offered a lighter alternative [19], capturing runway congestion factors, yet required airport-specific calibration that restricted broader applicability.

Subsequent work emphasized the role of airport layout and route structure. Approaches based on regression [20], node–link surface models [15], or spatio-temporal graph representations of trajectories [21] demonstrated that incorporating structural airport surface characteristics improved predictions. These methods estimated particularly well unimpeded times, but most remained single-airport case studies and did not examine variability.

With the rise of data-driven approaches, machine learning methods showed evident advantages over static references. Models using macroscopic features such as stand–runway pairs and airport traffic loads [22] improved accuracy, while operational predictors for pushback and ramp times [23] proved that simple supervised models performed well in practice. These efforts shifted the focus to predictive decision-making, though generalisation remained difficult.

Most recently, research has shifted towards deep learning architectures. Wide-and-Deep networks combining categorical and continuous features [24], Graph Neural Networks applied to airport surface routing [25], and hybrid CNN–GRU models capturing spatio-temporal dependencies in flows [26] all showed improvements over classical methods. The use of spatial and temporal airport structure features improved accuracy, but limited interpretability and airport-specific tailoring hinder

use in operations.

Overall, the literature demonstrates a clear progression: from static references to simulation and analytical approaches, to graph-based and machine learning models, and finally to deep learning. However, most contributions remain focused on taxi-out prediction, with relatively little attention given to taxi-in times, and are constrained to single-airport deployments. At the same time, evolving A-CDM requirements—notably the 2025 Specification [4], which enforces stricter milestone accuracy—are creating strong operational demand for models that are both precise and portable. This is where the present work contributes: by leveraging standardised airport data flows through AODB and other sources consistently available across A-CDM airports, it introduces a generic model for both arrivals and departures, developed and validated across six airports, thereby bridging the methodological gap of cross-airport generalisation.

III. METHODOLOGY

This section details the methodology for developing and evaluating a generalisable machine learning framework for taxi time prediction. The approach first describes standardised data sources available across A-CDM airports and the construction of the supervised dataset, including preprocessing and feature selection procedures. The model architecture is then specified, followed by the validation protocol designed to assess performance under realistic deployment conditions.

A. Data Sources

The model leverages standardised data sources commonly available across A-CDM airports to capture the operational, meteorological, and infrastructure factors influencing taxi time. Flight arrival and departure data were obtained from EUROCONTROL's Performance Review Unit via the Airport Operator Data Flow (APDF) database [27]. Since 2020, APDF has been updated monthly and includes scheduled and actual times, runway and stand assignments, operator, aircraft type, de-icing status, and additional flight-level details. As APDF is fed directly by Airport Operational Database (AODB) systems, it mirrors precisely the data recorded at A-CDM airports.

Complementing the operational data, weather features were extracted from Meteorological Aerodrome Reports (METARs) supplied by the Secure Aviation Data Information System (SADIS). These weather reports contain real-time weather observations at the aerodrome level and include temperature, precipitation, and visibility information relevant to this study. Infrastructure constraints were captured through taxiway and runway maintenance work data sourced from the EUROCONTROL Airport Corner [28], with the exception of Paris Orly (LFPO) and Václav Havel Prague Airport (LKPR), for which data were provided directly by the airport operators. Finally, aircraft-specific technical specifications were extracted from EUROCONTROL's Base of Aircraft Data (BADA) [29], including aircraft length, maximum take-off weight, wing area and number of engines. Although comparable information is available from open-source data sources such as the FAA's



Aircraft Characteristics Database [30], BADA was used here for convenience and consistency. These data do not originate from airport systems; they are provided by external services but remain consistently available across A-CDM airports.

B. Dataset

The dataset comprises 4.1 million flights spanning the 2022–2024 period. Each observation corresponds to a single flight operation and includes a target variable—the actual taxi time—and the predictor features described below.

The target variable represents actual taxi times calculated using A-CDM standard definitions, derived from operational timestamps in the APDF database.² The calculations distinguish between departure and arrival operations as follows:

$$\text{AXOT} = \text{ATOT} - \text{AOBT} \quad (\text{Departures}) \quad (1)$$

$$\text{AXIT} = \text{AIBT} - \text{ALDT} \quad (\text{Arrivals}) \quad (2)$$

where:

- ATOT = Actual Take-Off Time
- AOBT = Actual Off-Block Time
- AIBT = Actual In-Block Time
- ALDT = Actual Landing Time

To ensure data reliability, for each airport–phase (AXIT/AXOT) pair, the upper 0.1% and lower 0.1% taxi time values were removed, with the thresholds validated by airport operations experts.³

Regarding features, context features (*airport*, *phase*) are incorporated by design to enable a single unified model across A-CDM airports, thereby simplifying deployment and maintenance requirements. The *phase* feature distinguishes between AXIT and AXOT, while the *airport* feature encodes site-specific operational characteristics. Additional features were selected using a combination of domain expertise and SHAP (SHapley Additive exPlanations) [31] importance rankings, ensuring a reduced feature set that mitigates overfitting and enhances generalisation. Complete feature specifications are detailed in Table I.

C. Model

This study addresses taxi time prediction as a supervised machine learning regression problem using Gradient Boosting Decision Trees (GBDT). GBDT represents an ensemble learning technique that sequentially fits weak learners to build a strong predictive model, with each learner focusing on reducing residual errors from previous iterations. This approach generally outperforms basic statistical models and artificial neural networks on tabular data [32], while supporting interpretability through SHAP analysis.

Let $x_i \in \mathbb{R}^d$ denote the feature vector and $y_i \in \mathbb{R}$ the taxi time (minutes). The predicted taxi time $\hat{y}_i$ is estimated as:

$$\hat{y}_i = f_0 + \sum_{m=1}^M \nu f_m(x_i),$$

²All temporal data were standardised to UTC to ensure consistency across airports operating in different time zones.

³Part of the study was conducted within the EATIN programme, enabling consultation on operational aspects.

where f_0 is an initial constant (e.g., the mean of y), f_m denotes the m -th regression tree, M is the number of boosting rounds, and $\nu \in (0, 1]$ is the learning rate.

Model parameters are estimated by minimizing the empirical loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N l(y_i, \hat{y}_i),$$

where $l(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$ represents the Mean Absolute Error (MAE) loss. This choice provides robustness to outliers compared to squared-error alternatives.

Among available implementations, CatBoost was selected for its native handling of categorical variables (e.g., airport, phase, airline) and competitive performance compared to XGBoost and LightGBM [33]. This study uses a single modeling approach to emphasize interpretability and domain-specific insights over extensive machine learning model comparisons.

D. Evaluation Setup

The evaluation follows a validation protocol with training conducted on 2022–2023 data and testing performed on 2024 data. This temporal split simulates real-world deployment, where models trained on historical data must predict future operations without access to future patterns during training. While temporal splits can introduce year-specific effects, by 2022 European aviation had largely normalized following acute COVID disruptions (2020–2021), with traffic recovering to 83% of pre-pandemic levels [34], and the two-year training window allows to capture diverse operational conditions. This approach provides a more conservative and operationally realistic evaluation than random train-test splits, as it tests whether the model generalizes to genuinely unseen future operational contexts.

During the training phase, model hyperparameters—including the number of trees, maximum depth, and number of leaves per tree—were optimized using stratified K-Fold Cross-Validation (CV) with five folds. The stratification ensures each fold maintains proportional representation of all airport-phase combinations (six airports $\times$ two phases = twelve strata), preventing scenarios where certain folds might disproportionately contain flights from specific airports or phases. Hyperparameters were selected to minimize the Mean Absolute Error (MAE) metric on validation folds:

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

where y_i represents the observed taxi time and $\hat{y}_i$ the model prediction. Following hyperparameter optimization, the final model was fitted on the complete 2022–2023 dataset using the selected parameters (described in Table II).

The testing phase evaluates model performance against a baseline representing current airport operational practices through Standard Taxi Times (STT). These STT values are fixed reference times that airports and airlines use for operational planning, typically computed as median or mean historical times for specific stand–runway pairs, sometimes stratified

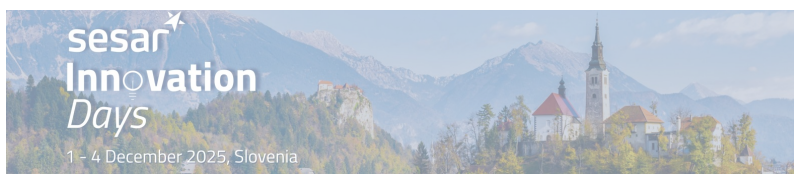


TABLE I. INPUT FEATURES USED IN THE TAXI TIME PREDICTION MODEL

Category	Feature	Description	Type	Unit	Source
Calendar	Hour	Hour of day at movement time	Categorical	–	APDF
	Month	Month of year at movement time	Categorical	–	APDF
	Weekday	Day of week (Mon–Sun)	Categorical	–	APDF
Aircraft	Wing Area	Wing planform area	Numerical	m ²	BADA
	Max Take-Off Weight	Maximum certified take-off weight	Numerical	kg	BADA
	Length	Fuselage length	Numerical	m	BADA
	Number of Engines	Installed engine count	Numerical	–	BADA
	Type	Aircraft ICAO designator	Categorical	–	APDF/BADA
Flight	Phase	Operational phase (Arrival or Departure)	Categorical	–	APDF
	Airline ICAO	Three-letter airline ICAO code	Categorical	–	APDF
	De-/Anti-Icing	De-/anti-icing status (None, Stand, Remote)	Categorical	–	APDF
Airport	Airport	Airport ICAO code	Categorical	–	APDF
	Runway	Assigned arrival/departure runway	Categorical	–	APDF
	Stand/Gate	Assigned stand or gate	Categorical	–	APDF
	Works	Indicator of scheduled airport works	Binary	–	Airport Corner
Meteorological	Cloud Ceiling	Height of lowest cloud layer	Numerical	ft	METAR
	Dew Point	Dew point temperature	Numerical	°C	METAR
	Temperature	Air temperature	Numerical	°C	METAR
	Visibility (CAVOK)	Indicator of clear visibility conditions	Categorical	–	METAR
	Visibility Distance	Reported horizontal visibility	Numerical	m	METAR
	Wind Direction	Wind direction (cardinal)	Categorical	–	METAR
	Wind Speed	Wind speed	Numerical	kt	METAR

TABLE II. FINAL CATBOOST HYPERPARAMETERS, OPTIMIZED USING THE OPTUNA HYPERPARAMETER OPTIMIZATION FRAMEWORK TO MINIMIZE VALIDATION MAE.

Parameter	Value	Description
Iterations	300	Number of boosting trees
Learning rate	0.10	Step size at each iteration
Depth	8	Maximum tree depth
L2 leaf reg.	6.6	L2 regularization strength
Random strength	6.0	Noise added for regularization
Loss function	MAE	Objective minimized

by aircraft type. For each test flight, the STT prediction is obtained by looking up the airport’s operational STT value corresponding to that flight’s stand-runway pair (and aircraft type where applicable).

Evaluation is conducted at two complementary levels: flight level and airport level, with both the proposed model and the STT baseline assessed using identical metrics to enable direct comparison. At the flight level, performance is assessed using two complementary metrics. Mean Absolute Error (MAE), defined previously as the model’s loss function and CV objective, quantifies prediction accuracy in absolute terms (minutes), providing directly interpretable operational insights into typical prediction deviations. Mean Absolute Percentage Error (MAPE) expresses errors as percentages of actual taxi times:

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

At the airport level, aggregated absolute error is computed as the sum of absolute prediction errors across all flights in

the 2024 test set:

$$\text{Aggregated Error} = \sum_{i=1}^{N_{2024}} |y_i - \hat{y}_i|,$$

where N_{2024} denotes flights per airport in 2024. This annual metric quantifies cumulative impact in total error minutes, with the model–baseline difference representing operational minutes saved while accounting for varying traffic volumes across airports.

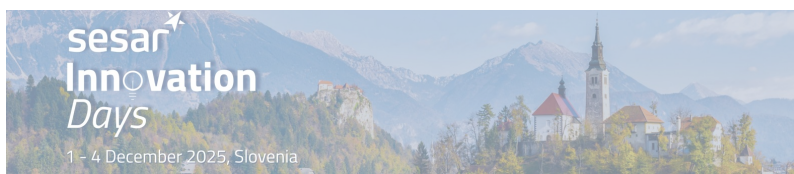
Finally, SHAP was employed both during feature selection and for post-hoc interpretation of the final model to quantify the contribution of each predictor and to reveal operational patterns driving predictions. The relative importance (share) of feature j is defined as:

$$\text{Importance Share}_j = 100 \times \frac{|\phi_j|}{\Phi},$$

where $\phi_j(x_i)$ is the SHAP value of feature j for prediction i , $|\phi_j| = \frac{1}{N} \sum_{i=1}^N |\phi_j(x_i)|$ is the mean absolute SHAP value of feature j , $\Phi = \sum_{k=1}^d |\phi_k|$ is the total explanation magnitude, N is the number of predictions, and d is the number of features.

IV. RESULTS

Following the evaluation protocol established in Section III-D, this section presents model performance across the six airports in the 2024 test set. This evaluation comprises three components: model benchmarking results, performance metrics assessing prediction accuracy at flight and airport levels, and interpretability analysis identifying key predictive factors.



A. Model Selection

To support the selection of CatBoost model, several models were benchmarked following the training and validation protocol outlined in Section III-D. Table III reports performance in terms of MAE, RMSE, and inference time (in seconds) for predictions of AXIT and AXOT taxi times combined. Linear models (Linear Regression and Ridge) achieved the lowest performance, with MAE values at 2.92 minutes. Gradient boosting methods demonstrated strongest performance for this task, with CatBoost obtaining the lowest MAE (2.73 minutes). Based on these results, CatBoost was selected as the final model, and all subsequent analyses are based on this implementation.⁴

TABLE III. MODEL BENCHMARK RESULTS. PERFORMANCE ON THE 2024 TEST SET.

Model	MAE	RMSE	Inference Time (s)
Linear Regression	2.92	4.21	2.84
Ridge	2.92	4.21	2.86
Random Forest	2.92	4.34	11.53
CatBoost	2.73	4.05	7.09

B. Performance Metrics

At the flight level, model accuracy was assessed using MAE and MAPE across the six European airports. The model consistently outperformed the baseline Standard Taxi Times, achieving MAE reductions ranging from 1% to 52% across airports and phases.

TABLE IV. AXIT (ARRIVALS): MODEL VS. BASELINE

Airport	MAE (Model)	MAE (Baseline)	MAPE (Model)
EBBR	1m 05s	1m 31s	18.4%
LFPO	1m 20s	1m 24s	19.7%
LKPR	1m 02s	1m 51s	20.5%
LOWW	0m 52s	1m 19s	16.8%
LSZH	0m 54s	1m 37s	16.5%
LTFM	2m 21s	2m 25s	17.4%

For arrivals (AXIT), reported in Table IV, MAE reductions ranged from 2% to 44%: Prague and Zurich (LKPR, LSZH) achieved the strongest improvements at 44%, followed by Vienna (LOWW) at 34% and Brussels (EBBR) at 29%. Paris-Orly and Istanbul (LFPO, LTFM) demonstrated more modest gains of 5% and 2%, respectively. Airports exhibiting lower absolute MAE values (LOWW, LSZH) also demonstrated the most favorable MAPE values of approximately 17%, indicating that predictions deviate by roughly one minute for every six minutes of actual taxi time. This pattern suggests improved effectiveness for shorter and more standardised taxi-in operations.

For departures (AXOT), reported in Table V, greater variability in performance was observed, with MAE reductions ranging from 1% to 52%. Vienna (LOWW) demonstrated the strongest improvement at 52%, followed by Zurich and

⁴RMSE values are reported for completeness but were not used as the primary model selection criterion.

TABLE V. AXOT (DEPARTURES): MODEL VS. BASELINE

Airport	MAE (Model)	MAE (Baseline)	MAPE (Model)
EBBR	1m 56s	2m 41s	17.2%
LFPO	2m 02s	2m 03s	19.5%
LKPR	1m 55s	2m 48s	17.8%
LOWW	1m 42s	3m 31s	19.0%
LSZH	2m 18s	3m 34s	19.2%
LTFM	3m 32s	3m 45s	20.1%

Prague (LSZH, LKPR) at 35% and 31%, respectively. Brussels (EBBR) achieved a 28% reduction, whereas Paris-Orly and Istanbul (LFPO, LTFM) showed minimal improvements of 1% and 6%. Model MAPE values cluster between 17% and 20%, indicating that predictions typically deviate by roughly one-fifth of actual taxi-out times.

Because MAPE is scale-independent, it enables direct comparisons across phases. Across the six airports, mean MAPE is 18.2% for arrivals and 18.8% for departures, indicating comparable relative accuracy. By contrast, MAE exhibits wider dispersion: arrivals range from 0m 52s (LOWW) to 2m 21s (LTFM), whereas departures range from 1m 42s (LOWW) to 3m 32s (LTFM). The higher and more variable departure MAE reflects added operational complexity—runway sequencing, ATC coordination, slot management, and de-/anti-icing—while the near-equivalent MAPE across phases shows that errors scale with phase-specific durations rather than indicating a loss of relative accuracy in departures.

Error distributions (defined as $Actual - Predicted$, in minutes) were examined at the airport-phase level to assess the predictive behaviour of the model for taxi-in (AXIT) and taxi-out (AXOT) operations. Figures 1–4 present the error distributions for Vienna Airport (LOWW) and Istanbul Airport (LTFM), representing respectively one of the lowest-error and highest-error cases in the dataset. These two airports were selected to illustrate the range of model performance observed across all airports.

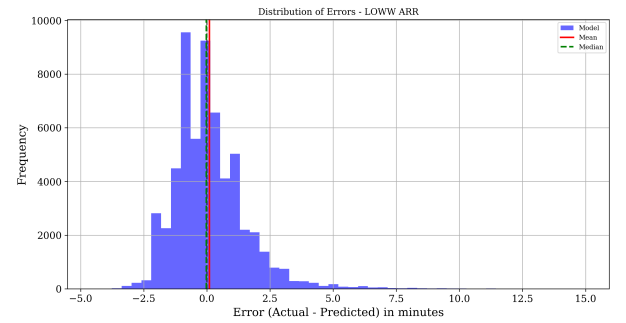


Figure 1. Taxi-in (AXIT) error distribution (LOWW airport)

Across both airports, the error distributions are centred near zero, with mean and median closely aligned, indicating an absence of systematic over- or under-prediction. Vienna LOWW displays a narrower distribution—particularly for arrivals—with most errors falling within ± 3 minutes, whereas Istanbul LTFM shows a longer right tail in both arrivals (AXIT) and departures (AXOT), reflecting occasional under-

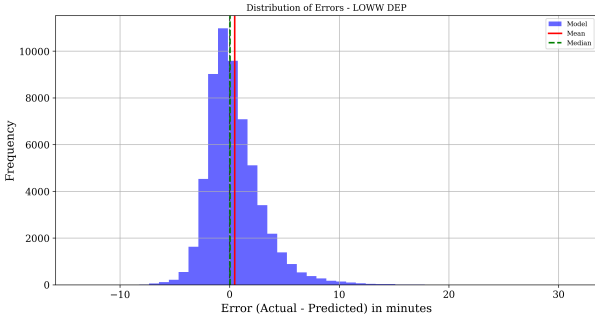


Figure 2. Taxi-out (AXOT) error distribution (LOWW airport)

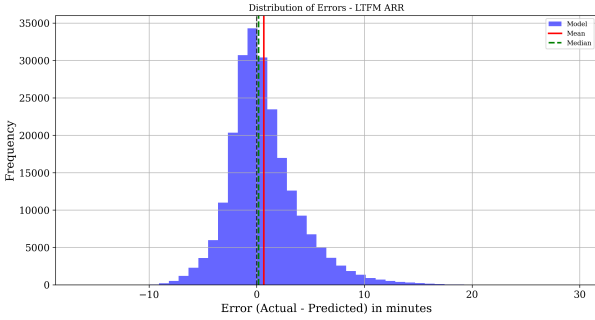


Figure 3. Taxi-in (AXIT) error distribution (LTFM airport)

predictions during periods of extended taxi times typical of large hub operations. Departures at both airports show higher variability than arrivals, consistent with the additional operational uncertainty arising from pushback procedures, queuing, and runway sequencing.

Overall, the distributions indicate stable predictive behaviour, with errors concentrated around zero and only limited occurrence of extreme deviations.

Beyond flight-level metrics, airport-level aggregated absolute error provides a cumulative perspective on operational impact across facilities of different scales. Figure 5 reports totals in thousands of minutes (k), with the gap between baseline (orange) and model (blue) bars representing operational time saved. Istanbul LTFM, as the major hub with the highest traffic volume, exhibits the largest aggregated error at approximately 780k minutes for the model versus 810k minutes for the baseline—a savings of 30k minutes corresponding to a 4% relative improvement that nonetheless represents substantial operational value given the airport’s scale and coordination complexity. Regional airports demonstrate markedly stronger performance: Vienna LOWW and Prague LKPR each record approximately 80k minutes of model error compared to 130k minutes for the baseline, achieving savings of 50k minutes and 38% improvement, while Zurich LSZH achieves 100k versus 155k minutes with savings of 55k minutes and a 35% reduction. These facilities benefit from simpler taxiway networks and more predictable operational patterns. Brussels EBBR and Paris-Orly LFPO occupy an intermediate position with aggregated errors of 160k and 200k minutes, achieving savings of 65k and 11k minutes with improvements of 29%

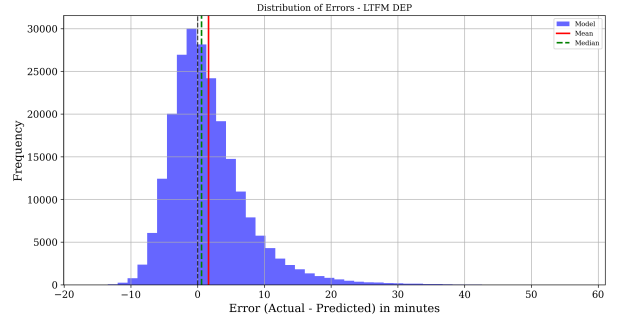


Figure 4. Taxi-out (AXOT) error distribution (LTFM airport)

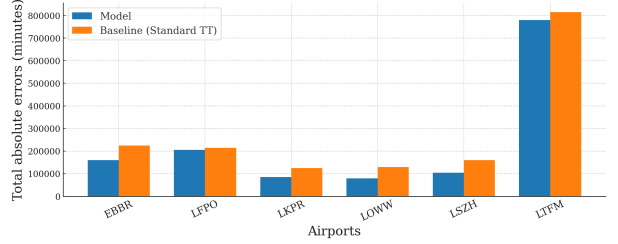


Figure 5. Aggregated absolute errors across airports showing total error minutes for combined arrival and departure operations

and 5%, respectively.

C. Model Interpretation

TABLE VI. GLOBAL SHAP IMPORTANCE SHARES BY FEATURE FOR DEPARTURES (AXOT) AND ARRIVALS (AXIT)

Feature	Share (AXOT / AXIT) %
Phase	29.07 / 23.39%
Runway	19.12 / 31.60%
Stand	13.50 / 16.30%
Airport	9.32 / 8.31%
De-/Anti-Icing	7.37 / 4.94%
Airline	5.44 / 4.70%
Aircraft Max Takeoff Weight	3.00 / 2.03%
Hour of Day	2.90 / 1.97%
Aircraft Type	1.79 / 0.98%
Temperature Dew Point	1.55 / 0.99%
Scheduled Works	1.24 / 0.86%
Aircraft Length	1.12 / 0.67%
Wing Area	1.00 / 0.83%
Visibility	0.93 / 0.72%
Temperature	0.76 / 0.52%

SHAP analysis, in Table VI, reveals distinct predictive hierarchies between departure and arrival operations, with facility and operations-related variables dominating both phases. For departures (AXOT), the *phase* indicator leads at 29%, followed by *runway assignment* (19%) and *stand location* (14%), which together account for 62% of total importance. Aggregated by category, facility and operations variables (phase, runway, stand, airport, de-/anti-icing, airline, and temporal indicators) comprise 87%, compared with 7% for aircraft characteristics and 3% for meteorology. For arrivals (AXIT), *runway* (32%), *phase* (23%), and *stand* (16%) total 71%; by category, facility and operations reach 91%, with aircraft at 5% and meteorology

at 2%. The 5% de-/anti-icing share on arrivals (all coded as “None”) arises from correlated encodings and interactions rather than causality. Notably, the *airport* identifier contributes approximately 9%; site-specific patterns are captured primarily via runway and stand assignments rather than the airport code itself. Finally, the temporal features—*hour of day*, *weekday*, and *month*—act as proxies for surface-demand regimes (e.g., peak and off-peak waves, seasonal traffic), thereby explaining part of the observed congestion signal without explicit surface-occupancy inputs.

V. DISCUSSION

Accurate taxi time prediction remains challenging in A-CDM operations when reliance on historical references fails to reflect variability from runway configuration, weather, de-/anti-icing, and surface demand. Errors propagate through milestone updates and DPI messages, degrading network efficiency. This study demonstrates that standardised A-CDM data flows can support accurate cross-airport prediction without site-specific calibration.

The framework relies on tabular inputs commonly available at A-CDM airports: operational data from Airport Operational Databases (accessed via EUROCONTROL’s APDF), METAR observations, and centrally maintained specifications (BADA for aircraft characteristics, Airport Corner for planned works). Evaluated on 2024 operations across six European airports, the single unified model achieves substantial improvements over Standard Taxi Times, with performance varying by airport operational scale. Regional platforms demonstrate the strongest relative gains, while major hubs show more modest improvements that nonetheless translate to meaningful operational benefits given higher traffic volumes.

Model interpretability reveals why cross-airport generalisation succeeds. Infrastructure assignment variables—runway, phase, and stand—dominate the predictive signal, collectively accounting for approximately two-thirds of total model importance. Airport-specific characteristics, airline operational patterns, and temporal factors contribute secondary effects, while aircraft physical properties and meteorological conditions play minimal roles. Because predictive power derives primarily from standardised assignments inherent to A-CDM data pipelines, a single model generalises effectively across diverse airport environments and operational phases. This contrasts fundamentally with Standard Taxi Times, which rely on static stand–runway averages and cannot capture day-of-operation variability in weather, de-icing requirements, or surface demand.

VI. CONCLUSION

The macroscopic, portable design of the model in this study entails limitations that warrant future attention. The model does not incorporate real-time surface conditions such as queue lengths, which may limit its accuracy at congested airports like Istanbul. Integrating surveillance-derived congestion indicators or developing strong operational proxies represents a natural extension. The single-algorithm approach favors

interpretability and deployment simplicity but would benefit from systematic benchmarking against alternative learners under the same evaluation protocol. Finally, generalisability is demonstrated across six European airports. Extending validation to airports with divergent procedures or outside the A-CDM framework would establish broader external validity and clarify the minimal feature set required for effective cross-airport prediction. Addressing these directions would strengthen the foundation for local decision support and network-wide predictability, furthering core objectives of collaborative air traffic management.

REFERENCES

- [1] EUROCONTROL, *Airport Collaborative Decision Making (A-CDM) Concept Page*. Available: <https://www.eurocontrol.int/concept/airport-collaborative-decision-making>, Last accessed September 2025.
- [2] EUROCONTROL, *Annual Network Operations Report 2024 – Annex II: Airports*, May 2025. Available: <https://www.eurocontrol.int/publication/network-operations-report-2024>.
- [3] EUROCONTROL, *Airport CDM Implementation Manual*, Version 5, 2017.
- [4] EUROCONTROL, *EUROCONTROL Specification for Airport Collaborative Decision Making (A-CDM)*, Edition 1.0, January 2025. Available: <https://www.eurocontrol.int/publication/eurocontrol-specification-airport-collaborative-decision-making>.
- [5] EUROCONTROL, *ATFCM Users Manual*, Edition 27.0, June 2024. Available: <https://www.eurocontrol.int/sites/default/files/2024-05/eurocontrol-atfcm-users-manual-maint-2-wef-20240618.pdf>.
- [6] I. Simaiakis and H. Balakrishnan, “Impact of congestion on taxi times, fuel burn, and emissions,” *Transportation Research Record*, vol. 2300, no. 1, pp. 36–44, 2013.
- [7] J. Park, H. Kim, and J. Lee, “Influential factors to aircraft taxi time in airport,” *Journal of Air Transport Management*, vol. 107, 102362, 2023.
- [8] M. E. J. Stettler, S. Eastham, and S. R. H. Barrett, “The impact of single engine taxiing on aircraft fuel consumption and pollutant emissions,” *The Aeronautical Journal*, vol. 122, no. 1255, pp. 1547–1566, 2018.
- [9] S. Neto, A. Silva, and G. Gonçalves, “Optimization of aircraft taxiing strategies to reduce the environmental footprint,” *Sustainability*, vol. 14, no. 15, p. 9692, 2022.
- [10] EUROCONTROL, *Sustainable Taxi Operations Concept of Operations*, 2024.
- [11] Air France, *Air France Halves CO2 Emissions on Two of its Flights with a Series of Actions to Limit its Environmental Footprint*, May 2022. Available: <https://corporate.airfrance.com/en/news/air-france-halves-co2-emissions-two-of-its-flights-series-actions-limit-its-environmental>.
- [12] NASA, *Taxi-Out Time Prediction for Departures at Charlotte Douglas International Airport*, NASA Report, 2017.
- [13] J. Wang, Y. Xu, and J. Yin, “Aircraft taxi time prediction: Feature importance and their impacts,” *Journal of Air Transport Studies*, vol. 12, no. 2, pp. 56–74, 2021.
- [14] Y. Yin et al., “Prediction and analysis of airport surface taxi time: Classification, features, and methodology,” *Applied Sciences*, vol. 14, no. 3, p. 1306, 2024.
- [15] M. Jeong, H. Lee, and I. Hwang, “Unimpeded taxi-time prediction based on the node–link model,” *Journal of Aerospace Information Systems*, vol. 17, no. 8, pp. 411–424, 2020.
- [16] Federal Aviation Administration, *U.S./Europe Comparison of ATM-Related Operational Performance 2010*, FAA Report, 2012.
- [17] L. Capelleras, *Additional Taxi-Out Time Performance Indicator Documentation*, EUROCONTROL Performance Review Unit, 2015.
- [18] R. Windhorst, A. Odoni, and T. Reynolds, “Validation of simulations of airport surface traffic with the Surface Operations Simulator and Scheduler,” in *13th AIAA Aviation Technology, Integration, and Operations Conference*, 2013.
- [19] S. Badrinath, H. Li, and H. Balakrishnan, “Integrated surface–airspace model of airport departures,” *Journal of Guidance, Control, and Dynamics*, vol. 42, no. 5, pp. 1049–1063, 2019.
- [20] Y. Zhang and Q. Wang, “Methods for determining unimpeded aircraft taxiing time and evaluating airport taxiing performance,” *Chinese Journal of Aeronautics*, vol. 30, no. 2, pp. 523–537, 2017.



- [21] H. Tran, D. T. Pham, M. Schultz, and S. Alam, "Taxi-speed prediction by spatio-temporal graph-based trajectory representation," in *International Conference on Research in Air Transportation (ICRAT)*, 2020.
- [22] J. Yin, Y. Ma, Y. Xu, and J. Sun, "Machine learning techniques for taxi-out time prediction with a macroscopic network topology," in *37th Digital Avionics Systems Conference (DASC)*, pp. 1–8, IEEE, 2018.
- [23] H. Lee, W. Malik, and Y. C. Jung, "Prediction of pushback times and ramp taxi times for departures at Charlotte Airport," in *AIAA Aviation Forum*, 2019.
- [24] X. Du, Y. Chen, and X. Li, "Using deep learning method to predict taxi time of aircraft: A case of Hong Kong Airport," *Transportation Research Part C: Emerging Technologies*, vol. 114, pp. 233–251, 2020.
- [25] K. Lim, A. Jardines, and R. J. Hansman, "Variable taxi-out time prediction using graph neural networks," in *SESAR Innovation Days*, 2021.
- [26] J. Yuan, Y. Wang, and X. Chen, "Predicting additional taxi-out time using CNN–GRU at Shanghai Pudong Airport," *Transportation Research Part C: Emerging Technologies*, vol. 158, p. 104432, 2024.
- [27] EUROCONTROL, *EUROCONTROL Specification for Operational ANS Performance Monitoring*, 2020. Available: <https://www.eurocontrol.int/publication/eurocontrol-specification-operational-ans-performance-monitoring>.
- [28] EUROCONTROL, *Airport Corner (Public)*. Available: https://ext.eurocontrol.int/airport_corner_public.
- [29] EUROCONTROL, *BADA: An Advanced Aircraft Performance Model for Present and Future ATM Systems*, Technical Report, 2009. Available: <https://www.eurocontrol.int/publication/bada-advanced-aircraft-performance-model-present-and-future-atm-systems>.
- [30] Federal Aviation Administration (FAA), *Aircraft Characteristics Database*, Online Resource, 2025. Available: https://www.faa.gov/airports/engineering/aircraft_char_database.
- [31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [32] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.
- [33] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, pp. 6639–6649, 2018.
- [34] EUROCONTROL, *Performance 2022 – Outlook 2023*, EUROCONTROL Analysis Paper, 2023. Available: <https://www.eurocontrol.int/publication/eurocontrol-analysis-paper-2022-year-european-aviation-bounced-back>.
- [35] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: An empirical study on performance estimation methods," *Machine Learning*, vol. 109, no. 11, pp. 1997–2028, 2020.