Vol. 11, No. 2, Oktober 2026, Page. 644-656
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Integrating Operational, Traffic Density, and Meteorological Features for Aircraft Ground Departure Duration Prediction

Devi V. Waas[1], Emanuella M. C. Wattimena*[2], Jefri E. T. Radjabaycolle[3], Shinta R. Mangaluk[4]

[1,2,3,4]Program Studi Ilmu Komputer, Universitas Pattimura, Indonesia
Email: [1]devi.waas@lecturer.unpatti.ac.id, [2*]emanuella.wattimena@lecturer.unpatti.ac.id,
[3]jefri.radjabaycolle@lecturer.unpatti.ac.id, shintamangaluk123@gmail.com

Abstract

Accurate prediction of aircraft departure duration is important for improving the efficiency of airport surface operations. This study develops a *machine learning* model to predict Ground Departure Duration (GDD) at a regional airport by integrating operational, temporal, schedule-adherence, short-term traffic-density, and meteorological features. The dataset consists of 10,084 observations with 31 predictor features. Traffic conditions are represented using multi-window features over 15- and 30-minute intervals, scheduled departure demand, and the number of active aircraft on the ground. Three algorithms were evaluated: Random Forest Regressor, Extra Trees Regressor, and CatBoost Regressor. The data were chronologically divided into 80% training data and 20% testing data, with *five-fold time-series cross-validation* used for model optimization. Performance was evaluated using RMSE, MAE, and R^2. Random Forest achieved the lowest RMSE of 5.1295 minutes and the highest R^2 of 0.3230, while CatBoost obtained the lowest MAE of 2.9685 minutes. The models were able to capture general GDD patterns under normal operating conditions but still tended to underestimate extreme durations. *Permutation Feature Importance* analysis showed that RIU_DEP, aircraft type, airline, and pushback time were among the most important features. These findings demonstrate the potential of *ensemble machine learning* for GDD prediction at regional airports, although further improvement is needed to better predict extreme operating conditions.

Keywords: CatBoost, extra trees, ground departure duration, machine learning, random forest, regional airport, traffic density

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

The growth in air transport activity demands increasingly efficient, timely, and predictable airport operations. One critical aspect of airport operations is aircraft movement during the departure phase, from the pushback process through to takeoff. This phase is dynamic, as it is influenced by traffic conditions, flight characteristics, operational schedules, and meteorological conditions. Uncertainties in aircraft ground movements can affect the use of airport facilities, the management of departure flows, and overall operational efficiency [1]. Furthermore, inefficient ground operations can increase wait times, fuel consumption, and aircraft emissions while on the ground [2].

This issue is not limited to high-traffic airports but is also relevant to regional airports, which have their own operational patterns and traffic characteristics. Changes in the number of flights at a given time can lead to varying runway surface conditions throughout the day. Uncertainty regarding departure times from parking positions can also alter the sequence of flights and the number of aircraft operating simultaneously [3]. Therefore, the ability to estimate the duration of the departure process more accurately is crucial for understanding operational conditions and managing aircraft movements at regional airports.

The duration of the departure process is influenced by various interrelated factors. Previous research has shown that taxi time is related to flight characteristics, aircraft type, operating time, traffic conditions, and meteorological factors [4]. Traffic conditions are a key factor because an increase in the number of arriving and departing flights can increase interactions between aircraft and put pressure on ground facilities [1], [4]. Meteorological conditions are also related to aircraft movement. Wind speed and direction, visibility, and precipitation can affect operating conditions and aircraft movement times

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

on the ground [1], [5]. This indicates that predicting departure duration requires a comprehensive consideration of various types of information.

In addition to traffic and meteorological conditions, the relationship between the schedule and actual operations also provides important information about the state of the aviation system. Differences between the planned departure time and the actual time may indicate disruptions or changes in operating conditions. Uncertainty regarding off-block time is known to affect the effectiveness of ground operations control [3]. Thus, information regarding deviations from the schedule can be used in conjunction with traffic conditions to describe the situation an aircraft faces when beginning the departure process.

Advances in the availability of flight operations and meteorological data have opened opportunities for the application of machine learning to study complex and nonlinear relationships between variables. Approaches that combine temporal information and actual operational factors have demonstrated strong performance in predicting taxi time [6]. CNN-GRU models have also been used to leverage temporal patterns and departure flow information [7]. Other studies have shown that predictive performance is influenced by the choice of algorithm as well as the quality of the features used [8]. More recent approaches even utilize spatio-temporal traffic information to improve the model's ability to capture changes in operational conditions on the airport tarmac [9]. These findings indicate that the quality of the representation of operational conditions plays a crucial role in addition to the selection of the predictive algorithm.

This study utilized three ensemble tree-based algorithms. Random Forest is capable of handling nonlinear relationships in tabular data, has good generalization ability, and allows for feature importance analysis [10]. Extra Trees Regressor employs a higher degree of randomization in tree construction and can be used to capture complex patterns in regression problems [11]. Meanwhile, CatBoost uses a gradient boosting approach with an ordered boosting mechanism designed to improve the learning process and reduce certain forms of bias in the model [12], [13]. These three algorithms are relevant for comparison because flight operation data consist of various numerical and categorical features that may have nonlinear relationships.

Although various approaches have been developed for aircraft taxi-time prediction using aircraft movements, real-time information, departure flow, gate-level time-series, deep-learning models, and spatiotemporal feature [4], [6]-[9], the representation of short-term traffic conditions remains limited. Airport surface activity can change rapidly, and a single traffic-density measure may not adequately capture variations across different time horizons, particularly at regional airports where traffic intensity can fluctuate substantially. Moreover, previous studies have not explicitly integrated multi-window short-term traffic density, scheduled departure demand, active ground aircraft, schedule adherence, and meteorological conditions for GDD prediction in a regional-airport context. This study addresses this gap by integrating these factors using 15- and 30-minute traffic windows.

Based on these gaps, This study integrates operational, temporal, schedule-adherence, short-term traffic-density, ground-activity, and meteorological features to predict Ground Departure Duration (GDD) at a regional airport. Its novelty lies in using multi-window short-term traffic density to capture changing traffic conditions across different intervals. The study compares Random Forest, Extra Trees, and CatBoost models and identifies the most influential predictors. The proposed approach is expected to provide both accurate GDD predictions and interpretable insights into factors affecting aircraft departure duration.

2. MATERIALS AND METHODS

2.1. Research Design and Data Source

This study employs a predictive quantitative approach based on machine learning to predict Ground Departure Duration (GDD) at a Pattimura International Airport, Ambon, Indonesia. The research stages include data collection and integration, preprocessing, target variable formulation, feature engineering, chronological data splitting, model development and optimization, performance evaluation, and model interpretation. The overall research workflow is shown in Figure 1.

Vol. 11, No. 2, Oktober 2026, Page. 644-656
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

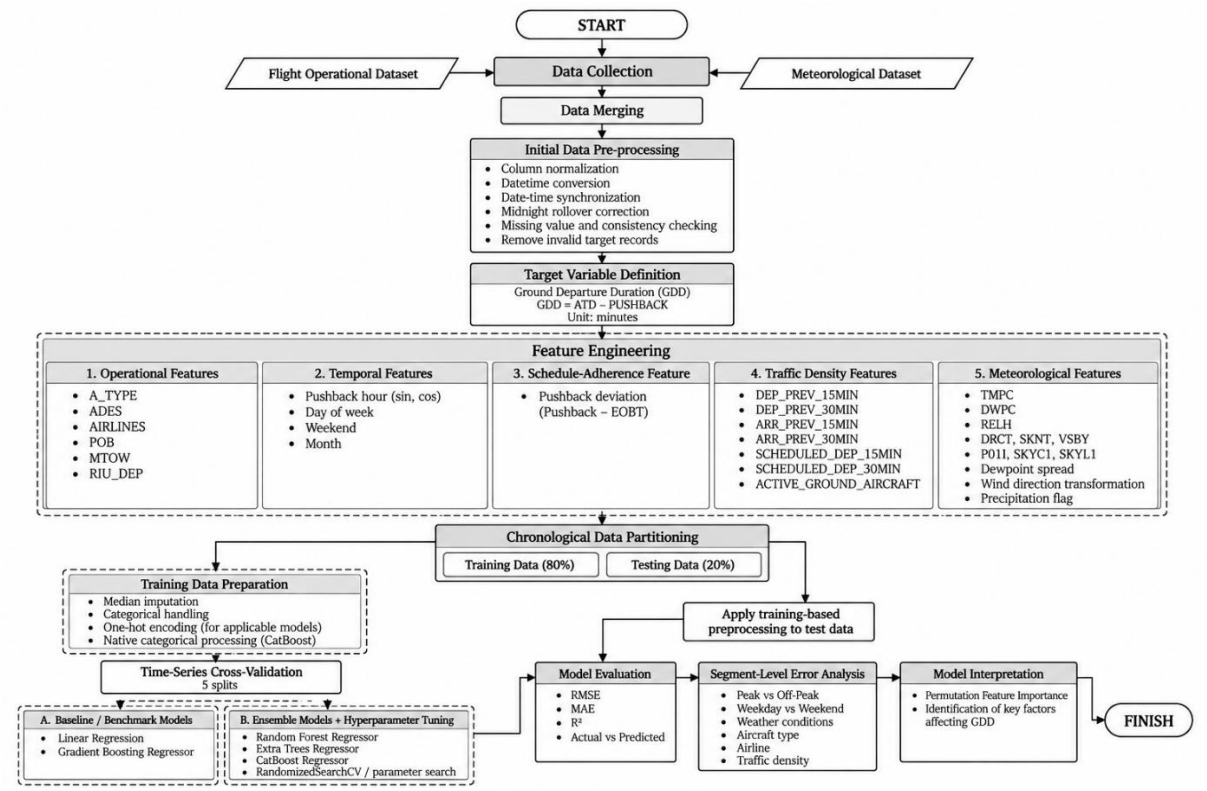


Figure 1. Research workflow

The research data consist of two main categories: flight operational data and meteorological data. The flight operational data were obtained from official airport operational records provided for research purposes under restricted-access and confidentiality provisions. To comply with the data-sharing agreement, the identity of the operational data provider is not disclosed in this study. The dataset contains information on flight schedules and operational characteristics, including Estimated Off-Block Time (EOBT), pushback time, Actual Time of Departure (ATD), Actual Time of Arrival (ATA), aircraft type (A_TYPE), flight destination (ADES), airline (AIRLINES), Passengers on Board (POB), Maximum Take-Off Weight (MTOW), and Runway in Use for Departure (RIU_DEP).

Meteorological data were obtained through the ASOS-AWOS-METAR Data Download service managed by the Iowa Environmental Mesonet (IEM) at Iowa State University. This platform provides an archive of METAR observations and has been used as a source of aviation meteorological data in previous studies [14]. Operational and meteorological data were then integrated into an analysis dataset based on time-synchronized information. The features used in this study are shown in Table 1.

Table 1. Summary of Predictor, Reference, and Target Variables and Their Temporal Availability

Variable group	Variables	Definition and temporal availability
Operational	A_TYPE, ADES, AIRLINES, POB, MTOW, RIU_DEP	A_TYPE represents aircraft type; ADES, the destination aerodrome; AIRLINES, the aircraft operator; POB, persons on board; MTOW, maximum take-off weight; and RIU_DEP, the runway-in-use configuration for departure. These variables are available before or at the reference pushback time.
Temporal	PUSHBACK_HOUR_SIN, PUSHBACK_HOUR_COS, DAY_OF_WEEK, WEEKEND, MONTH	Derived from the pushback timestamp to represent daily and calendar patterns; available at the prediction reference time.

Variable group	Variables	Definition and temporal availability
Schedule adherence	PUSHBACK_DEVIATION_MIN	Difference between actual pushback and EOBT. EOBT is known before pushback, while actual pushback defines the prediction reference time.
Traffic density	DEP_PREV_15MIN, DEP_PREV_30MIN, ARR_PREV_15MIN, ARR_PREV_30MIN, SCHEDULED_DEP_15MIN, SCHEDULED_DEP_30MIN, ACTIVE_GROUND_AIRCRAFT	Represents recent departures/arrivals, scheduled departure demand, and active aircraft on the ground. Historical departure/arrival counts use only events observed before the reference pushback; scheduled demand is derived from EOBT.
Meteorological	TMPC, DWPC, DEWPOINT_SPREAD, RELH, SKNT, WIND_DIR_SIN, WIND_DIR_COS, VSBY, P01I, PRECIPITATION_FLAG, SKYC1, SKYL1	METAR-based atmospheric conditions and derived weather features matched to the operational reference time.
Reference variables	PUSHBACK, EOBT	PUSHBACK is the actual start of pushback and defines the prediction reference time; EOBT is the scheduled off-block time known beforehand.
Outcome / target	ATD, GDD	ATD is the actual time at which the aircraft takes off from the runway. GDD is the target variable, calculated as GDD = ATD - PUSHBACK (minutes). The target flight's ATD is not used as a predictor.

2.2. Data Pre-processing

The preprocessing stage is performed to ensure the consistency of the data structure and quality before the data is used in modeling. Column names are first normalized, and then date and time information is converted to the datetime format. Date and time synchronization is performed because some time variables in the original data are available only in hours and minutes.

Observations lacking pushback or ATD information were excluded because both variables are required for target generation. Missing values in the predictor features were handled during the modeling phase using information derived solely from the training data. This approach is necessary to reduce the risk of bias and prevent the use of information- from the test data in the training process [15].

The target variable of this study is Ground Departure Duration (GDD), which is the time interval between pushback and ATD. GDD is calculated in minutes as

$$GDD_i = (ATD_i - PUSHBACK_i)\ minutes \quad (1)$$

where GDD_i is the departure process duration for flight i. After the time correction process is performed, observations with negative GDD values are removed from the dataset.

2.3. Feature Engineering

The *feature engineering* stage was performed to create features that comprehensively represent flight operations and improve the model's ability to predict Ground Departure Duration (GDD). The features were grouped into five categories: operational, temporal, schedule adherence, traffic density, and meteorological features. Operational features describe flight characteristics and departure conditions, including aircraft type, destination, airline, parking position, passenger count, Maximum Take-Off Weight (MTOW), and flight status. These variables capture operational differences that may influence GDD.

Temporal features were derived from pushback time. Because time is cyclical, pushback hour was transformed using sine and cosine functions to preserve the proximity between times near the daily cycle boundary, such as 11:00 p.m. and 12:00 a.m., as follows:

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

$$PB_{sin} = sin\left(2\pi\frac{h}{24}\right) \quad (2)$$

$$PB_{cos} = cos\left(2\pi\frac{h}{24}\right) \quad (3)$$

where h represents the pushback time in hours. This transformation is used to preserve the periodic nature of time and maintain the proximity between values at the boundaries of the daily cycle [16]. In addition to the cyclic representation of pushback times, other temporal features include DAY_OF_WEEK, WEEKEND, and MONTH. The DAY_OF_WEEK variable indicates the day of the week on which a flight occurs; WEEKEND is used as an indicator to distinguish between weekdays and weekends; and MONTH is used to represent potential variations in operational patterns from month to month.

Furthermore, the degree of alignment between the flight schedule and actual operational conditions is represented by the pushback deviation feature. This feature is calculated as the difference between the actual pushback time and the Estimated Off-Block Time (EOBT), which is defined as:

$$PD_i = PUSHBACK_i - EOBT_i \quad (4)$$

where PD_i is the pushback deviation value for flight ii, expressed in minutes. Positive values indicate pushback after EOBT, while negative values indicate earlier pushback. This feature represents the deviation between scheduled and actual operations.

Surface traffic conditions are represented using short-term traffic-density features, including departures and arrivals within 15- and 30-minute windows, scheduled departures, and active aircraft on the ground. This multi-window approach captures traffic variations across different time intervals before pushback.

To evaluate the contribution of the window-based traffic-density features, an ablation analysis was performed using four configurations: no temporal-window features, 15-minute features only, 30-minute features only, and the combined 15- and 30-minute features used in the full model. ACTIVE_GROUND_AIRCRAFT was retained in all configurations because it represents the instantaneous ground-traffic state rather than a temporal-window feature. All configurations were evaluated using the same chronological time-series cross-validation procedure and model settings to ensure a consistent comparison.

Meanwhile, meteorological features represent atmospheric conditions during operations and include temperature, dew point, relative humidity, wind speed and direction, visibility, precipitation, and cloud conditions. Derived variables are also generated to improve atmospheric representation. The dew point spread is calculated as the difference between air temperature and dew point:

$$DPS = TMPC - DWPC \quad (5)$$

where TMPC represents the air temperature and DWPC represents the dew point temperature.

Wind direction is also cyclic in nature because the values 0° and 360° correspond to the same direction. Therefore, wind direction is transformed using the sine and cosine functions:

$$WD_{sin} = sin\left(2\pi\frac{DRCT}{360}\right) \quad (6)$$

$$PB_{cos} = cos\left(2\pi\frac{DRCT}{360}\right) \quad (7)$$

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

where DRCT represents wind direction in degrees. In addition, the variable PRECIPITATION_FLAG was created as a binary indicator to show the presence or absence of precipitation at the time of observation.

Through this feature engineering process, the raw data is transformed into a feature representation better suited for modeling needs. The combination of operational, temporal, schedule adherence, traffic density, and meteorological features is expected to provide a more comprehensive picture of the conditions affecting variations in Ground Departure Duration.

2.4. Multi-Window Traffic Density

Short-term traffic conditions are represented using traffic-density features calculated over 15- and 30-minute temporal windows. This representation is intended to capture recent variations in departure and arrival activity at different short-term temporal scales around the reference pushback time.

In addition to actual activity, scheduled departure demand is calculated based on EOBT. The number of flights with an EOBT within $a \pm 15 - minute$ and $\pm 30 - minute$ window relative to the reference flight's pushback time is represented as SCHEDULED_DEP_15MIN and SCHEDULED_DEP_30MIN.

To assess whether the temporal traffic-density windows provide additional predictive value, an ablation analysis was conducted using four feature configurations: no window-based traffic features, 15-minute features only, 30-minute features only, and the combined 15- and 30-minute features used in the full model. ACTIVE_GROUND_AIRCRAFT was retained in all configurations because it represents the instantaneous ground-traffic state rather than a temporal-window feature. All configurations were evaluated under the same chronological time-series cross-validation procedure and model settings to ensure a consistent comparison.

2.5. Dataset Splitting and Preparation

The dataset has a temporal structure, so the data is not split randomly. All observations are sorted by pushback timestamp. The first 80% of the data is used as training data, while the last 20% is used as test data.

The chronological hold-out strategy was adopted to preserve the temporal ordering of the observations so that the models were trained using data from earlier periods and evaluated using observations occurring afterward. This approach is more appropriate for temporal data than random splitting because it reduces the likelihood of future information being indirectly introduced into the training process [17], [18]. Preventing data leakage is important to ensure that the reported model performance reflects realistic predictive capability rather than access to information unavailable at the intended prediction time [19].

Within the training set, TimeSeriesSplit with five splits was applied while maintaining chronological order. In each iteration, the validation fold always followed the corresponding training fold temporally. To further prevent information leakage, all data-dependent preprocessing operations were fitted independently within each training fold. For Random Forest and Extra Trees, missing numerical values were imputed using the median calculated only from the corresponding training fold. Missing categorical values were replaced with "Unknown", and one-hot encoding was fitted exclusively on the training fold before the resulting transformation was applied to the corresponding validation fold.

For CatBoost, numerical missing-value imputation was likewise performed using statistics derived only from the corresponding training fold. Categorical variables were retained in their original categorical form and were processed directly by the CatBoost algorithm [12], [13]. Thus, no information from the validation folds was used during preprocessing or model fitting. After model selection, the final models were retrained using the complete 80% training set and subsequently evaluated on the independent 20% chronological test set.

2.6. Model Development and Optimization

Three regression algorithms were used and compared: the Random Forest Regressor, the Extra Trees Regressor, and the CatBoost Regressor. Random Forest and Extra Trees are tree-based ensemble learning approaches, while CatBoost is a gradient boosting approach. These three algorithms were selected because they are capable of handling non-linear relationships and complex interactions in tabular data [10], [11], [12], [13].

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Hyperparameter optimization was performed using only the training data. Proper hyperparameter tuning is necessary to achieve a balance between the model's ability to learn from the training data and its ability to generalize to new data.

Random Forest and Extra Trees were optimized using RandomizedSearchCV with 20 parameter combinations and five time-series splits. The parameter space used is shown in Table 2. The best configuration is selected based on the lowest validation MAE.

Table 2. Hyperparameter Search Space

Hyperparameter	Search Space
n_estimators	200, 300, 500, 700
max_depth	None, 8, 12, 16, 20
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4
max_features	sqrt, log2, 0.5, 0.8

The final Random Forest and Extra Trees configurations used 500 estimators, a maximum tree depth of 12, a minimum of five samples for node splitting, a minimum of four samples per leaf, and a maximum-feature fraction of 0.5.

For CatBoost, combinations of tree depth, learning rate, and L2 leaf regularization were evaluated using the same five-fold time-series validation framework. The final CatBoost model used 700 boosting iterations, a tree depth of 6, a learning rate of 0.03, an L2 leaf-regularization coefficient of 5, and MAE as the optimization loss.

All experiments used a fixed random seed of 42 to ensure reproducibility. The models were implemented in Python using scikit-learn and CatBoost; the software versions used in the final experiments were recorded to support reproducibility.

To provide a meaningful performance reference, two benchmark models were included. Linear Regression was used as a conventional linear baseline, while Gradient Boosting Regressor served as a standard nonlinear boosting benchmark. All benchmark and ensemble models were evaluated using the same chronological training and testing partitions and the same RMSE, MAE, and R^2 metrics to ensure a fair comparison.

2.7. Model Performance Evaluation

The performance of the final model was evaluated using 20% of the test data that was not used in the training or optimization process. Three metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination R^2. Lower MAE and RMSE values indicate smaller prediction errors. MAE represents the mean absolute error, while RMSE imposes a greater penalty on prediction errors with high values [20]. Meanwhile, a higher R^2 value indicates the model's better ability to explain the variation in GDD.

2.8. Model Interpretation

Model interpretation was performed using Permutation Feature Importance (PFI). This method measures changes in model performance after the value of a feature is randomized, while other features are held constant. The greater the decline in performance after a feature is randomized, the higher that feature's predictive contribution to the model [21].

PFI was performed with 10 iterations and used MAE as the basis for evaluating performance changes. Subsequently, the 15 features with the highest permutation importance values were used to identify the operational, temporal, schedule compliance, traffic condition, and meteorological factors that contributed most significantly to GDD predictions.

To ensure consistency between evaluation and interpretation, feature importance analysis should be applied to the model with the best performance on the test data. With this approach, the identified factors truly originate from the model that provides the best predictive capability.

3. RESULTS AND DISCUSSION

After completing all stages of preprocessing, target variable creation, and feature engineering, the dataset used in the modeling consisted of 10,084 observations with 31 predictor features. The dataset was then divided chronologically into 8,067 observations (approximately 80%) for the training set and 2,017 observations (approximately 20%) for the test set. The training data covers the period from November 30, 2024, at 9:58 PM to October 22, 2025, at 10:59 PM, while the test data covers the period from October 22, 2025, at 11:42 PM to December 31, 2025, at 10:43 PM. This separation preserves the temporal order so that the model is tested using observations that occurred after the training period.

The evaluation results for the three models are shown in Table 3. Random Forest achieved an RMSE of 5.1295 minutes, an MAE of 3.1152 minutes, and an R^2 of 0.3230. Extra Trees produced an RMSE of 5.2775 minutes, an MAE of 3.1506 minutes, and an R^2 of 0.2834. Meanwhile, CatBoost produced an RMSE of 5.4629 minutes, an MAE of 2.9685 minutes, and an R^2 of 0.2322.

Table 3. Performance of the Ground Departure Duration Prediction Model

Model	RMSE	MAE	R^2
CatBoost	5.4629	2.9685	0.2322
Random Forest	5.1295	3.1152	0.3230
Extra Trees	5.2775	3.1506	0.2834
Linear Regression	5.5772	3.3234	0.1997
Gradient Boosting	5.2570	3.2098	0.2890

The benchmark models provide a clearer reference for evaluating the ensemble methods. Linear Regression achieved an RMSE of 5.5772 minutes, MAE of 3.3234 minutes, and R^2 of 0.1997, while Gradient Boosting improved the performance to an RMSE of 5.2570 minutes, MAE of 3.2098 minutes, and R^2 of 0.2890. Among all evaluated models, Random Forest achieved the lowest RMSE of 5.1295 minutes and the highest R^2 of 0.3230, whereas CatBoost obtained the lowest MAE of 2.9685 minutes. These results show that the proposed ensemble models provide competitive performance compared with both linear and nonlinear benchmark approaches.

To further evaluate the contribution of the temporal traffic-density features, an ablation analysis was conducted using four configurations: no window-based traffic features, 15-minute features only, 30-minute features only, and the combined 15- and 30-minute features used in the full model. All non-window predictors were kept unchanged, and ACTIVE_GROUND_AIRCRAFT was retained in all configurations because it represents the instantaneous ground-traffic state rather than a temporal-window feature. Each configuration was evaluated using the same five-fold time-series cross-validation procedure and final model settings to ensure a consistent comparison. The results are summarized in Table 4.

Table 4. Ablation analysis of temporal traffic-density window configurations

Model	Configuration	MAE	RMSE	R^2
CatBoost	No window	2.756	4.742	0.245
	15 min	2.766	4.731	0.249
	30 min	2.772	4.733	0.248
	15+30 min	2.770	4.737	0.247
Extra Trees	No window	2.897	4.620	0.284
	15 min	2.891	4.618	0.284
	30 min	2.894	4.614	0.285
	15+30 min	2.891	4.614	0.285
Random Forest	No window	2.949	4.601	0.287
	15 min	2.941	4.599	0.288
	30 min	2.948	4.601	0.287

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

The Random Forest prediction pattern compared to the actual values is shown in Figure 2. In general, the prediction line is able to follow the main pattern of the actual data for most observations. However, some actual GDD values experience a very sharp increase up to about 60–80 minutes, while the Random Forest predictions remain within a lower range. This indicates that the model is able to predict normal operating patterns quite well but still struggles when faced with extreme GDD events.

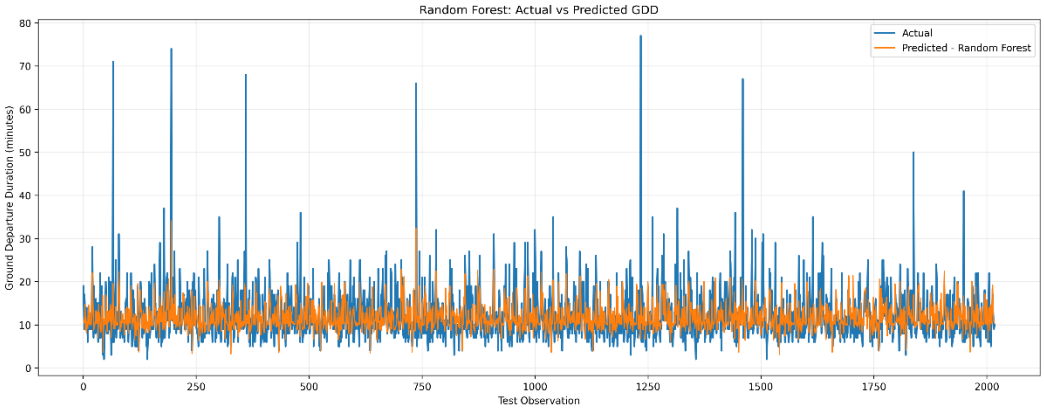


Figure 2. Comparison of actual GDD and Random Forest predictions

A similar pattern is observed in the Extra Trees model, as shown in Figure 3. The Extra Trees predictions follow the main variations within the typical GDD range but are not fully able to capture the spikes in the actual values. These characteristics correspond to an RMSE of 5.2775 minutes and an R^2 of 0.2834. Compared to the Random Forest model, Extra Trees demonstrates a slightly lower ability to explain the variation in the target variable.

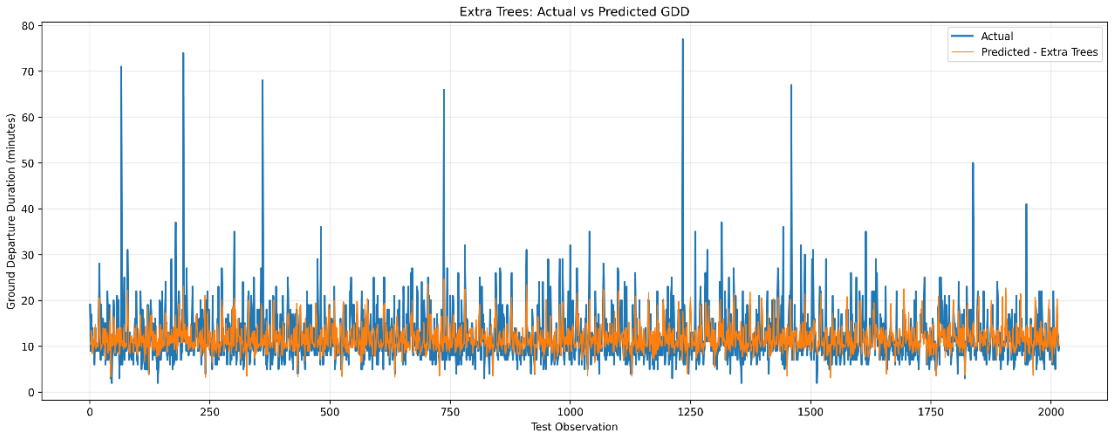


Figure 3. Comparison of actual GDD and Extra Trees predictions

The CatBoost results in Figure 4 show a prediction pattern that is relatively concentrated within the typical range of GDD values. This aligns with CatBoost's lowest MAE value. CatBoost's predictions tend to be close to most actual values under normal operating conditions, but it is less capable of tracking GDD spikes in some extreme observations. As a result, although its mean absolute error is the lowest, CatBoost's RMSE is the highest and its R^2 value is the lowest among the three models.

Vol. 11, No. 2, Oktober 2026, Page. 644-656
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

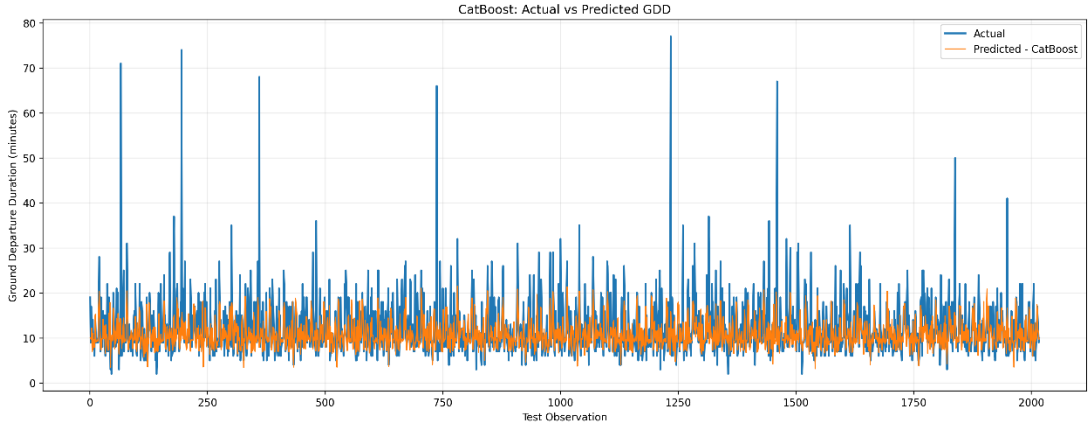


Figure 4. Comparison of actual GDD and CatBoost predictions

The three graphs in Figures 2–4 show the same trend: the model performs better at predicting GDD within the range of frequently occurring values than at predicting extreme values. For some observations with actual GDD values exceeding 50 minutes, the predicted results tend to be significantly lower. This pattern indicates a tendency toward underestimation in events with long departure durations. One possible cause is that there are fewer extreme observations compared to observations within the normal GDD range, so the patterns learned by the model better represent typical operational conditions.

To quantify the uncertainty in the performance differences among the ensemble models, paired bootstrap analysis was applied to the independent chronological test-set predictions. Random Forest achieved a lower RMSE than Extra Trees ($\Delta RMSE = -0.146, 95\%$ confidence interval [CI]: −0.288 to −0.023) and CatBoost ($\Delta RMSE = -0.330, 95\%\ CI: -0.591\ to\ -0.125$). Extra Trees also achieved a lower RMSE than CatBoost ($\Delta RMSE = -0.184, 95\%\ CI: -0.323\ to\ -0.066$). Since none of the 95% confidence intervals included zero, the observed RMSE differences were consistently reproduced across the bootstrap samples.

A segment-level analysis of the Random Forest predictions showed that model performance varied across operational conditions. During peak periods, the MAE increased to 3.3588 minutes and the RMSE to 5.7567 minutes. In contrast, during off-peak periods, the MAE was 2.7028 minutes and the RMSE was 3.8411 minutes. Weekday and weekend performance was relatively similar, while high-traffic conditions produced a slightly higher MAE than low-traffic conditions. These results suggest that GDD prediction becomes more challenging when operational activity increases.

Further interpretive analysis was conducted using permutation feature importance for the CatBoost model, as shown in Figure 5. Model performance was metric dependent: Random Forest achieved the lowest RMSE and the highest R^2, whereas CatBoost achieved the lowest MAE. Because MAE was used as the primary optimization criterion and directly reflects the average absolute prediction error, CatBoost was selected for the feature-importance analysis. The results show that RIU_DEP was by far the most influential predictor, with a permutation-importance value substantially higher than those of the remaining variables. The next most influential features were A_TYPE, AIRLINES, and PUSHBACK_HOUR_SIN, followed by ADES, MTOW, PUSHBACK_DEVIATION_MIN, and PUSHBACK_HOUR_COS. POB, ACTIVE_GROUND_AIRCRAFT, the 15- and 30-minute traffic-window variables, and DAY_OF_WEEK showed comparatively smaller individual contributions. These findings indicate that CatBoost predictions were primarily influenced by runway-use information, aircraft and airline characteristics, and temporal departure patterns, while traffic-density variables provided smaller complementary contributions.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

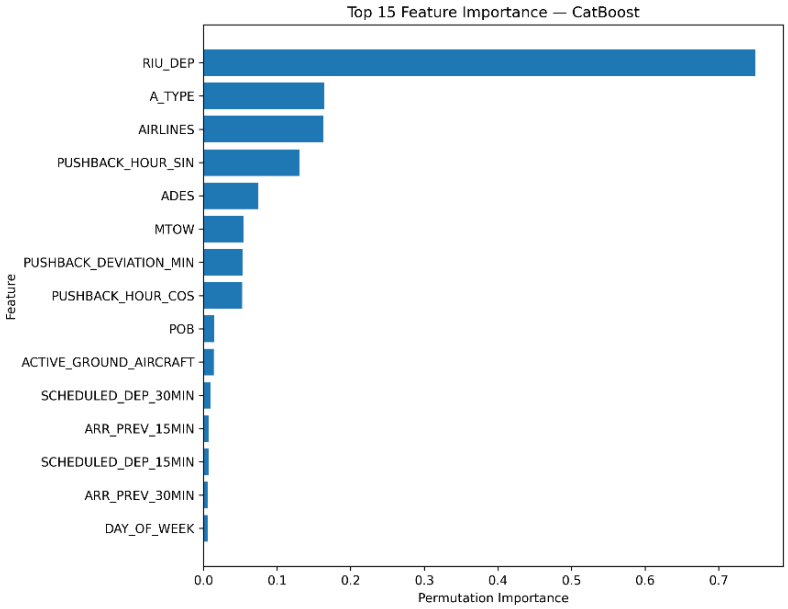


Figure 5. Features with the highest Permutation Feature Importance scores

The dominance of A_TYPE and AIRLINES indicates that flight operational characteristics are strongly associated with GDD variation. Aircraft type may reflect differences in physical characteristics and operating procedures, while AIRLINES may capture operator-specific operational practices. Temporal features such as PUSHBACK_HOUR_SIN and PUSHBACK_HOUR_COS also appear among the important predictors in Figure 5, suggesting that GDD varies with time of day and changes in airport activity. PUSHBACK_DEVIATION_MIN also contributes meaningfully, indicating that deviations between EOBT and actual pushback time are related to GDD. This is consistent with previous findings that off-block time uncertainty can affect airport surface operations [3].

Short-term traffic features, including ACTIVE_GROUND_AIRCRAFT, SCHEDULED_DEP_30MIN, SCHEDULED_DEP_15MIN, ARR_PREV_15MIN, and ARR_PREV_30MIN, also appear in Figure 5, although with lower importance than the main operational features. This suggests that traffic conditions remain relevant, but operational and temporal characteristics have a stronger influence in the regional airport context.

The traffic-density representation incorporates operational activity over both 15- and 30-min temporal windows, providing complementary information on short-term airport traffic conditions. However, the ablation analysis indicates that the contribution of these window-based features is modest and model dependent, with the combined 15+30-min configuration not consistently outperforming the individual windows. Overall, Table 3 and Figures 2–4 show that the ensemble models capture general GDD patterns but remain limited in predicting extreme durations. Random Forest achieves the lowest RMSE and highest R^2, whereas CatBoost achieves the lowest MAE. Operational and temporal variables remain the dominant predictors, while short-term traffic-density information provides a smaller complementary contribution.

The proposed GDD prediction framework has both practical and theoretical implications. Practically, it can support short-term departure planning by helping airport and ATC personnel anticipate longer ground-departure durations using information available before or at pushback. Theoretically, the findings show that GDD at a regional airport is shaped by the interaction of operational, temporal, schedule-adherence, runway-configuration, and short-term traffic conditions, indicating that the dominant factors may differ from those reported for large hub airports.

4. CONCLUSION

This study developed machine-learning models to predict Ground Departure Duration (GDD) at a regional airport using operational, temporal, schedule-adherence, short-term traffic-density, and meteorological features. Random Forest, Extra Trees, and CatBoost were evaluated using a chronological train-test framework. Among the evaluated models, Random Forest achieved the lowest

RMSE of 5.1295 minutes and the highest R^2 of 0.3230, while CatBoost obtained the lowest MAE of 2.9685 minutes. Extra Trees showed intermediate performance. Although the models captured general patterns in GDD, their relatively moderate explanatory performance and tendency to underestimate unusually long departure durations indicate that substantial variability remains unexplained, particularly for extreme operational conditions.

Permutation Feature Importance showed that operational and temporal variables contributed most strongly to prediction, particularly RIU_DEP, A_TYPE, AIRLINES, and PUSHBACK_HOUR_SIN. PUSHBACK_DEVIATION_MIN also provided meaningful predictive information, while multi-window traffic-density variables contributed additional but comparatively smaller predictive value. These findings suggest that GDD is influenced by a combination of aircraft and airline characteristics, departure timing, schedule adherence, and the short-term operational environment.

Within the context of the studied airport and observation period, the results demonstrate that tree-based ensemble models can provide useful predictive information for characterizing GDD patterns. However, the findings should not be interpreted as evidence of general applicability across airports or operational environments. The analysis was conducted using data from a single regional airport, and airport-specific operational procedures, traffic characteristics, infrastructure, and weather conditions may influence model performance. In addition, operational patterns may change over time, which could introduce temporal distribution shifts and reduce predictive reliability when the models are applied to future periods.

Therefore, the potential use of these models to support departure planning should be considered preliminary and context-specific rather than immediately operational. Further research should evaluate the proposed framework using data from different time periods and additional airports, assess model robustness under changing operational conditions, and improve the prediction of unusually long GDD events. Such temporal and external validation is necessary before broader operational implementation can be recommended.

5. REFERENCES

[1] J. Yin, M. Zhang, Y. Ma, W. Wu, H. Li, and P. Chen, "Prediction and Analysis of Airport Surface Taxi Time: Classification, Features, and Methodology," *Applied Sciences*, vol. 14, no. 3, p. 1306, Feb. 2024, doi: 10.3390/app14031306.

[2] Y. Chen, L. Quan, and J. Yu, "Aircraft Taxi Path Optimization Considering Environmental Impacts Based on a Bilevel Spatial–Temporal Optimization Model," *Energies (Basel).*, vol. 17, no. 11, p. 2692, Jun. 2024, doi: 10.3390/en17112692.

[3] S. Badrinath, H. Balakrishnan, E. Joback, and T. G. Reynolds, "Impact of Off-Block Time Uncertainty on the Control of Airport Surface Operations," *Transportation Science*, vol. 54, no. 4, pp. 920–943, Jul. 2020, doi: 10.1287/trsc.2019.0957.

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

[5] D. K. Park and J. K. Kim, "Influential factors to aircraft taxi time in airport," *J. Air Transp. Manag.*, vol. 106, p. 102321, Jan. 2023, doi: 10.1016/j.jairtraman.2022.102321.

[6] F. Wang, J. Bi, D. Xie, and X. Zhao, "A data-driven prediction model for aircraft taxi time by considering time series about gate and real-time factors," *Transportmetrica A: Transport Science*, vol. 19, no. 3, May 2023, doi: 10.1080/23249935.2022.2071353.

[7] L. Yuan, J. Liu, H. Chen, D. Fang, and W. Chen, "A CNN-GRU Hybrid Model for Predicting Airport Departure Taxiing Time," *Aerospace*, vol. 11, no. 4, p. 261, Mar. 2024, doi: 10.3390/aerospace11040261.

[8] S. Okwir, K. Amouzgar, and A. HC. Ng, "Exploring prediction accuracy for optimal taxi times in airport operations using various machine learning models," *J. Air Transp. Manag.*, vol. 122, p. 102684, Jan. 2025, doi: 10.1016/j.jairtraman.2024.102684.

[9] Y. Lim, S. Alam, N. Lilith, F. Tan, and P. L. Toon, "Graph Neural Networks with Spatiotemporal Flow Features for Aircraft Taxi-Out-Time Prediction," *Journal of Air Transportation*, vol. 33, no. 4, pp. 297–318, Oct. 2025, doi: 10.2514/1.D0369.

[10] E. Scornet and G. Hooker, "Theory of Random Forests," *Annu. Rev. Stat. Appl.*, vol. 13, no. 1, pp. 99–121, Mar. 2026, doi: 10.1146/annurev-statistics-112723-034707.
[11] S. Martiello Mastelini, F. K. Nakano, C. Vens, and A. C. P. de Leon Ferreira de Carvalho, "Online Extra Trees Regressor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 6755–6767, Oct. 2023, doi: 10.1109/TNNLS.2022.3212859.
[12] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1706.09516
[13] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.
[14] D. Alves, F. Mendonça, S. S. Mostafa, and F. Morgado-Dias, "Low Tropospheric Wind Forecasts in Aviation: The Potential of Deep Learning for Terminal Aerodrome Forecast Bulletins," *Pure Appl. Geophys.*, vol. 181, no. 7, pp. 2265–2276, Jul. 2024, doi: 10.1007/s00024-024-03522-z.
[15] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.
[16] S. Meisenbacher *et al.*, "Review of automated time series forecasting pipelines," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 6, Nov. 2022, doi: 10.1002/widm.1475.
[17] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: an empirical study on performance estimation methods," *Mach. Learn.*, vol. 109, no. 11, pp. 1997–2028, Nov. 2020, doi: 10.1007/s10994-020-05910-7.
[18] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Comput. Stat. Data Anal.*, vol. 120, pp. 70–83, Apr. 2018, doi: 10.1016/j.csda.2017.11.003.
[19] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.
[20] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.
[21] D. Debeer and C. Strobl, "Conditional permutation importance revisited," *BMC Bioinformatics*, vol. 21, no. 1, p. 307, Dec. 2020, doi: 10.1186/s12859-020-03622-2.