International Journal of Transport Development and Integration
https://www.acadlore.com/journals/IJTDI

Leveraging Real-Time GTFS and Integrated Data for High-Accuracy LRT Departure Delay Prediction Using Optimized Machine Learning

Rossi Passarella[1*], Aulya Putri Ayu[1], Mastura Diana Marieska[2], Isbatudinia[1], Nurainiyah Solehan[1], Harumi Veny[3], Romi Fadillah Rahmat[4]

[1] Department of Computer Engineering, Faculty of Computer Science, Sriwijaya University, 30662 Indralaya, Indonesia
[2] Department of Informatics, Faculty of Computer Science, Sriwijaya University, 30662 Indralaya, Indonesia
[3] College of Engineering, School of Chemical Engineering, MARA University of Technology, 40450 Shah Alam, Malaysia
[4] Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, 20155 Medan, Indonesia

* Correspondence: Rossi Passarella (passarella.rossi@unsri.ac.id)

Received: 08-28-2025 **Revised:** 09-28-2025 **Accepted:** 09-29-2025

Citation: R. Passarella, A. P. Ayu, M. D. Marieska, Isbatudinia, N. Solehan, H. Veny, and R. F. Rahmat, "Leveraging real-time GTFS and integrated data for high-accuracy LRT departure delay prediction using optimized machine learning," *Int. J. Transp. Dev. Integr.*, vol. 9, no. 4, pp. 790–803, 2025. https://doi.org/10.56578/ijtdi090408.

Abstract: Efficient light rail transit (LRT) systems are crucial for sustainable urban mobility; however, unforeseen departure delays continue to be a major hurdle, undermining operational reliability and passenger satisfaction. This study establishes a data-driven framework for forecasting departure delays by combining static GTFS schedules with real-time GTFS operational data from the Canberra LRT system. The dataset included 15,538 records with 42 attributes, spanning from 28 August 2020 to 13 August 2022. A stringent preprocessing pipeline was implemented, encompassing temporal feature engineering and feature selection based on mutual information. The Random Forest regressor with feature engineering and selection (RFR-FEFS) attained the highest predictive performance on the test set (R^2 = 0.94, MAE = 2.93, MSE = 34.32). The high accuracy indicates the model's efficacy, yet it necessitates careful evaluation of potential overfitting and its generalizability beyond the examined system. Ablation experiments were performed to assess the impact of various feature groups by omitting temporal, spatial, or operational attributes. The findings indicate that the exclusion of temporal features decreased R^2 to 0.90, the exclusion of spatial features reduced it to 0.93, and the exclusion of operational features resulted in the most significant decline to 0.23. These findings affirm that all three feature categories contribute distinctly and synergistically to model performance. This research illustrates the capability of integrating diverse GTFS data with sophisticated machine learning techniques to attain precise LRT delay forecasts. Nevertheless, the framework was validated solely on one system and time frame; future research should investigate its transferability to other cities and integrate supplementary contextual data, including meteorological conditions and incident reports, to improve robustness and practical applicability.

Keywords: Canberra; Delay prediction; Feature engineering; Feature selection; GTFS real-time; LRT; Random Forest Regressor

1 Introduction

Public transport systems are fundamental to sustainable urban development, providing vital services that alleviate traffic congestion, decrease carbon emissions, and improve urban accessibility [1]. LRT systems are increasingly favored in urban regions for their capacity, speed, reliability, and comparatively reduced environmental impact relative to personal motorized transportation [2]. The efficiency of LRT is essential for offering a reliable and appealing alternative to private vehicles, thus advancing the overarching objectives of sustainable mobility.

The operational efficiency and perceived reliability of LRT systems are significantly hindered by delays. These disruptions arise from a combination of unpredictable factors, including technical malfunctions, infrastructure problems, and dynamic operational decisions [3, 4]. Such delays affect service quality and operational performance and have a direct impact on passenger satisfaction and may lead to higher operational expenses. Therefore, accurately predicting LRT arrival and departure delays is essential for better system management, enabling timely schedule

adjustments, effective resource planning, and keeping passengers informed. Achieving high prediction accuracy is crucial for enhancing the passenger experience and optimizing system operations. Although various methods have been investigated for predicting public transport delays, traditional approaches often rely heavily on historical schedule data or simplistic models, which struggle to capture the complex and dynamic interplay of factors affecting real-time delays. Recent studies suggest that integrating dynamic data sources can significantly enhance the performance of predictive models [5]. Among these sources, GTFS data, which includes both static and real-time information, such as vehicle positions and delays, provides useful details about the operational status of transit systems. While some studies have combined GTFS with external sources like weather or incident data, the specific contribution of GTFS-only features to accurate delay prediction has not been thoroughly explored, particularly in the context of LRT systems.

The machine learning methods, an integral component of artificial intelligence [6], have demonstrated exceptional proficiency in identifying intricate patterns within extensive datasets and generating precise predictions across various domains, including transportation. Applying ML to train delay prediction allows leveraging vast historical and real-time data to build models that can adapt to changing operational conditions. The efficacy of machine learning models is significantly contingent upon the quality and relevance of input features. Given the limited availability or inconsistency of external datasets such as weather or incident reports in many urban settings and in line with the scope of this study, which excludes such external variables, developing robust models based solely on GTFS-derived features holds practical value. GTFS data is frequently unrefined and encompasses numerous attributes, many of which require transformation into significant inputs via feature engineering and selection.

To optimize GTFS data for precise LRT delay predictions through machine learning, it is crucial to utilize sophisticated methods for feature engineering and selection. Feature engineering entails converting raw data into more valuable information, such as deriving temporal components (e.g., hour, day, and month) from timestamps to enhance the comprehension of delay patterns. Consequently, choosing the pertinent features is essential for preserving only the most significant ones, thereby streamlining the model, augmenting its learning efficiency, mitigating the likelihood of overfitting, and ultimately enhancing the accuracy and interpretability of the predictions. Despite previous studies examining various facets of this issue, a comprehensive analysis of the impact of diverse GTFS-based features on LRT delay prediction—integrated with meticulous machine learning-oriented feature engineering and selection—has yet to be adequately investigated. This study restricts its analysis to GTFS-based features, intentionally omitting external data sources like weather, to evaluate the predictive capacity derived exclusively from publicly accessible transit feeds.

This study intentionally omits external datasets, unlike prior research on LRT delay prediction in Canberra, which integrated GTFS data with supplementary information such as weather and passenger details to enhance model performance [7]. Passenger data was unavailable for the analyzed period, and weather data was deliberately excluded to concentrate exclusively on the predictive capabilities of features derived from GTFS. This methodology enables a direct comparison of results within the same case study, facilitating an understanding of the accuracy of predictions based solely on GTFS data. This finding corresponds with previous research indicating that external datasets, including passenger counts or disruption reports, are frequently inaccessible or inconsistently documented in numerous urban transit systems [8].

We aimed to determine whether models constructed solely with features derived from GTFS data can accurately forecast LRT delays, excluding weather or passenger information utilized in prior research. Our objective is to comprehend how various categories of GTFS-based features affect prediction accuracy. To accomplish this, we established multiple experimental configurations subsequent to feature extraction:

- Baseline: Using all available features, then selecting the top 15 based on mutual information.
- Temporal only: Using only time-related features.
- No temporal: Omitting all time-related features.
- No operational: Excluding operational attributes.
- No spatial: Excluding location-based features.

By comparing how well the models perform under each scenario, we can pinpoint which groups of features have the greatest influence on prediction accuracy.

This study aims to address this deficiency by proposing and assessing a machine learning system for forecasting LRT delays utilizing both real-time and static GTFS data. The objective is to investigate the extent to which predictive performance is attainable solely through GTFS-derived features, excluding external sources like weather or incident data. The framework incorporates sophisticated feature engineering and selection and introduces an ablation study to evaluate the impact of various feature groups (temporal, spatial, and operational) on overall model performance. This method improves interpretability and bolsters the research contribution beyond mere predictive accuracy. The subsequent sections of this paper are organized as follows: Section 2 examines pertinent literature on transport delay prediction; Section 3 outlines the methodology, followed by Section 4, which presents the results and discussion; and ultimately, Section 5 concludes the paper.

2 Literature Review

This section offers a thorough examination of previous studies related to improving the precision of delay prediction in public transportation, particularly in railway and LRT systems. We examine innovative applications of real-time data such as GTFS and meteorological information, explore various machine learning methodologies, and assess the significance of feature engineering and selection techniques in this domain. The aim is to consolidate existing knowledge, ascertain the state-of-the-art, and emphasize current limitations that warrant the approach proposed in this study.

Advancements in information technology and the growing accessibility of detailed, real-time operational data have profoundly altered the domain of transport system analysis and forecasting. This is particularly evident in forecasting delays for trains and urban rail, where precise predictions are essential for efficient operations and passenger satisfaction. The incorporation of data beyond conventional static schedules has become a crucial element in the creation of more resilient and adaptable predictive models.

2.1 Utilization of Dynamic Data Sources in Delay Prediction

The effectiveness of delay prediction models is highly contingent on the quality and richness of the input data. While past operational data provides a starting point, it's important to include real-time information and outside factors to accurately understand the unpredictable nature of delays.

Real-time Operational Data: Real-time data streams, such as those adhering to the General Transit Feed Specification (GTFS-Real time), offer critical insights into the present condition of the transportation network, encompassing exact vehicle positions, operational status, and reported delays [9]. Research has progressively utilized GTFS data to advance from schedule-based forecasts to models that account for real network dynamics. For instance, in reference [10], a system was developed to collect and manage both static and real-time GTFS data to facilitate the prediction of train delays, demonstrating an improved approach to modeling unforeseen issues. The principal benefit lies in its ability to detect schedule deviations in real time, which facilitates more timely and pertinent forecasts than models that rely exclusively on historical data. The diversity and magnitude of real-time data pose difficulties in data processing, noise attenuation, and efficient feature extraction.

In addition to fixed schedules, various dynamic operational factors can affect transport delays, including network congestion, vehicular incidents, and infrastructure limitations. Precisely capturing these dynamic elements via real-time GTFS data allows models to more accurately represent the current condition of the transport system. Incorporating various sources of real-time operational data enhances the context for forecasting delays, facilitating the consideration of abrupt disruptions and augmenting model responsiveness. Nonetheless, obstacles persist in efficiently processing and integrating diverse real-time data streams to fully exploit their predictive capabilities for LRT systems.

2.2 Machine Learning Approaches for Transport Delay Prediction

Machine learning techniques have emerged as the primary approach for creating advanced predictive models in transportation, owing to their capacity to discern intricate, non-linear relationships within extensive datasets. Their application encompasses diverse predictive tasks, including travel time estimation, traffic flow forecasting, and delay prediction across multiple transportation modes [11].

Various ML algorithms have been utilized for predicting delays in trains and other modes of transportation. Ensemble methods such as RFR are distinguished by their robustness, capacity to manage high-dimensional data, and resistance to overfitting, rendering them particularly appropriate for intricate, heterogeneous datasets. Research, including that conducted by Guo et al. [12], illustrates the efficacy of RFR in forecasting delays in transportation scenarios, such as flight departures. Considering these benefits, RFR is chosen for this study to model LRT delay prediction, as it achieves a balance between predictive accuracy, interpretability, and computational efficiency. The efficacy of RFR is significantly contingent upon the quality and representation of input features, underscoring the necessity of proficient data preparation.

2.3 The Role of Feature Engineering and Selection

The efficacy of machine learning models is significantly affected by the quality and pertinence of the features utilized for training. Complex and heterogeneous data sources, such as integrated GTFS and weather data, may not allow raw attributes to directly reflect the underlying patterns indicative of delays.

Feature engineering is the process of converting raw data into features that more accurately represent the problem for the model. This may entail the creation of new attributes or the modification of existing ones [13]. Delay prediction typically involves the extraction of temporal features, including the hour of the day, day of the week, month, and holiday indicators, to discern periodic patterns in delays. Furthermore, it may entail the development of spatial features, including the distance from the origin and particular attributes of stations [14].

Subsequent to feature engineering, feature selection is essential, particularly when managing an extensive array of potential features sourced from various origins. Feature selection seeks to identify and preserve only the most pertinent and informative attributes, eliminating redundant or extraneous ones [15]. This process provides multiple advantages: it diminishes model complexity, enhances training efficiency, increases model interpretability, and, importantly, alleviates the risk of overfitting, thereby augmenting generalization performance [16, 17]. Despite its demonstrated efficacy in various machine learning applications, such as intrusion detection systems that evaluate intricate data, the implementation of sophisticated feature engineering and selection techniques tailored for the integrated GTFS-real-time and weather dataset in forecasting LRT delays requires further investigation to fully harness the advantages of machine learning models.

2.4 Synthesis and Research Gap

Based on the foregoing review, it is evident that leveraging dynamic data sources, applying powerful machine learning algorithms, and employing effective feature engineering techniques are individually recognized as crucial components for accurate delay prediction in transport systems. Prior research has explored aspects of these areas, demonstrating the value of real-time data and the predictive power of ML models while acknowledging the importance of feature processing. However, a notable gap exists in comprehensively exploring the synergistic effect of tightly integrating real-time GTFS operational data within a predictive framework that employs rigorous, data-driven feature engineering specifically optimized for LRT delay prediction. While some studies may look at these factors, there is less research on how careful processing of this specific mix of different and changing data affects the performance of popular machine learning regression models like RFR for predicting LRT delays. Addressing this gap is vital for developing truly robust and highly accurate prediction systems capable of handling the complex and dynamic operational environment of modern LRT networks. This study contributes by offering a structured, reproducible approach that integrates GTFS real-time data and tailored feature engineering to enhance delay prediction performance using RFR. Furthermore, it includes an ablation study to systematically evaluate the impact of temporal, spatial, and operational features on model performance, highlighting which aspects contribute most significantly to prediction accuracy.

3 Methodology

This section details the methodology employed to develop and evaluate the proposed framework for predicting LRT departure delays in Canberra, Australia. The overall research framework, encompassing data acquisition, preprocessing, model development, and evaluation, is illustrated in Figure 1. The subsequent subsections describe each of these stages in detail, ensuring clarity and reproducibility of the research process.

3.1 Data Acquisition and Description

The successful development of robust predictive models hinges upon the quality and comprehensiveness of the underlying dataset. This study leveraged and integrated multiple publicly available open data sources pertaining to the Canberra LRT system. The assembled dataset incorporates official schedule information, real-time operational data reflecting actual train movements, and external meteorological conditions, which are hypothesized to influence service punctuality. Integrating these diverse data streams is crucial for providing a holistic view of the factors contributing to train delays. The specific datasets acquired and utilized are described below.

3.1.1 Static GTFS data

Static GTFS data offers essential schedule information and immutable geographical data pertaining to transit networks. This study utilized static GTFS data sourced from the official open data portal of Transport Canberra (Information for Developers–Transport Canberra). The stop_times.txt file was utilized to obtain the scheduled arrival and departure times for each trip at every stop. Station locations and identification information were extracted from the stops.txt file. The two files were merged utilizing the stop_id field to produce a unified static schedule dataset. The dataset consisted of 21,324 rows and 9 features, detailing scheduled trip information associated with particular stop locations.

3.1.2 Real-time GTFS data

Real-time GTFS data captures dynamic updates on vehicle positions and service alerts, offering vital information about actual operational performance and deviations from the static schedule. Real-time data for the Canberra Metro Light Rail was acquired from the ACT Government Open Data Portal (Canberra Metro LRT Feed–Historical Archive). This data source consists of two primary feeds:

(a) Trip updates

This feed contains real-time information on trip progress, including estimated or actual arrival and departure times at stops, often indicating delays or early arrivals relative to the schedule. The dataset used contained 2,378,307 records with 14 features.

(b) Vehicle positions

This feed provides the real-time geographical coordinates (latitude, longitude) of active vehicles, obtained via GPS. The dataset comprised 13,148,937 records with 16 features, allowing for tracking vehicle movements. These two real-time datasets were integrated by matching the trip id, stop id, and stop sequence fields. The resulting combined real-time dataset contained 39,546 rows and 27 features, providing a detailed snapshot of actual operational timing and locations at specific moments.

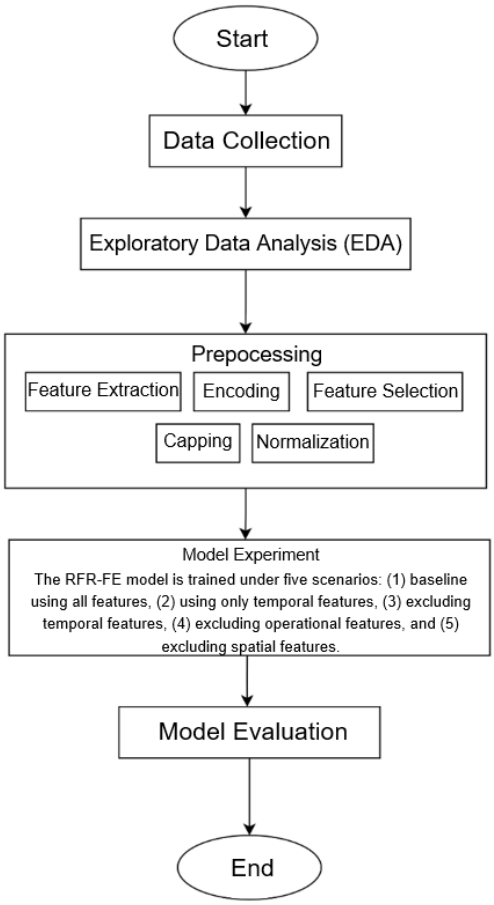

Figure 1. Research framework

3.2 Data Preprocessing

Before model training, the integrated dataset underwent a series of preprocessing steps to ensure data quality, enhance feature representation, and optimize suitability for machine learning algorithms. These steps included handling missing values, feature engineering, categorical encoding, feature selection, outlier detection and treatment, and feature scaling.

3.2.1 Handling missing values

Missing values were addressed using methods appropriate for the characteristics of each feature. For time-series data points where continuity is important, linear interpolation was applied to estimate missing values based on surrounding observations. The median value of a feature was used to fill in gaps for numerical features that could not be filled in using interpolation. For categorical features, missing values were replaced with the mode (most frequent value). Effective handling of missing data is critical to prevent biases and maintain data integrity [18].

3.2.2 Feature engineering

Feature engineering was conducted to extract more informative attributes from the raw data, focusing specifically on timestamp features [19]. The arrivalTime and departureTime timestamps were broken down into detailed temporal features, including the hour, minute, and second of the day. This approach aims to capture time-based patterns and cyclical trends in delays more effectively, such as the effects of peak hours or variations in schedule adherence throughout the day.

3.2.3 Categorical encoding

Categorical features within the dataset were converted into a numerical format suitable for machine learning algorithms [20]. The method used to convert categorical features in the dataset into a numerical format is label encoding. Label encoding transforms each category into a unique number that represents the category. This method was chosen because it is effective and efficient when the categorical features have a meaningful order, allowing the algorithm to utilize this ordinal information. Label Encoding is also more space-efficient compared to other methods like One-Hot Encoding, as it only generates one column for each categorical feature that is transformed into a number. This transformation is necessary, as most ML algorithms require numerical input.

3.2.4 Feature selection

To improve model performance, reduce dimensionality, and mitigate overfitting, a feature selection process was applied. Given the potential for nonlinear relationships between features and the target variable (departureDelay), the mutual information (MI) method was chosen. MI quantifies the dependency between two variables and is capable of detecting non-linear associations [21].

Nonetheless, MI possesses certain recognized limitations. Initially, it is susceptible to discretization decisions, indicating that the method of binning continuous variables can influence the resultant mutual information values. Secondly, mutual information does not intrinsically account for temporal dependencies, which is especially pertinent in time-series data like transit operations. Notwithstanding these constraints, MI was chosen due to its exceptional appropriateness for regression tasks. In contrast to correlation-based methods that solely identify linear relationships, MI can discern both linear and non-linear dependencies between features and the continuous target variable (departureDelay). This renders MI beneficial for modeling intricate and dynamic transportation data, where delay patterns are frequently affected by numerous non-linear interactions. Moreover, MI is capable of managing both categorical and continuous variables, which is well-suited to the heterogeneous characteristics of the GTFS dataset.

After applying MI, we ranked all the available features, including the engineered temporal ones from Section 3.2.2, based on their scores against departureDelay. We then grabbed the top 20 with the highest MI scores for the next round of modeling, and you can see the results laid out in Table 1.

Table 1. Feature selection results

No	Features	MI Scores
1	arrivalDelay	1.512629
2	actual_departure_second	1.130363
3	TripUpdateID	0.450874
4	HR	0.400808
5	trip_id	0.303691
6	stop_id	0.294136
7	scheduled_departure_second	0.259081
8	minute	0.254047
9	longitude	0.244755
10	location	0.243397
11	latitude	0.234340
12	scheduled_arrival_second	0.221010
13	odometer	0.192704
14	actual_departure_minute	0.182502
15	VehicleUpdateID	0.176679
16	stop_sequence	0.171139
17	hour	0.159469
18	scheduled_departure_time	0.147816
19	scheduled_arrival_time	0.140049
20	actual_arrival_second	0.131426

3.2.5 Outlier treatment

Outliers in the selected numerical features were handled using the Interquartile Range (IQR) capping method [22]. Values falling below Q1: -1.5 IQR or above Q3: +1.5 IQR were capped at the respective boundary values (Q1: -1.5 IQR or Q3: +1.5 IQR). This approach is effective in reducing the influence of extreme values without removing potentially valuable data points entirely.

3.2.6 Feature scaling

The numerical features were scaled to standardize their ranges and distributions. The StandardScaler was used to transform the features so that they have a mean of 0 and a standard deviation of 1. This step is crucial for algorithms sensitive to feature scales to ensure optimal performance [23].

3.3 Experimental Setup and Model Evaluation

This section outlines the experimental setup for training and evaluating the RFR model for LRT departure delay prediction. After preprocessing and feature engineering, the final dataset comprised 15,538 records with 42 features, covering the period from 28 August 2020 to 13 August 2022. The dataset was then partitioned into training and testing sets using an 80/20 chronological split. The training set consisted of the earlier 80% of the records, while the testing set contained the subsequent 20%, representing later periods. This approach was adopted to prevent temporal leakage and to ensure that the evaluation reflects realistic prediction scenarios, where future delays are predicted solely from historical patterns.

In this study, all model training and evaluation were conducted on a laptop with the following specifications: Windows 11 (64-bit) operating system, 16 GB RAM, 13th Gen Intel® Core™ i5-13500H processor (16 CPUs), and a 512 GB SSD.

The RFR model was selected due to its robustness, ability to handle high-dimensional data, and proven success in similar transportation-related prediction tasks, as discussed in Section 2. Hyperparameter tuning was carried out using Bayesian optimization combined with 5-fold cross-validation over 20 iterations. This approach strikes a balance between managing computational costs and obtaining stable, reliable estimates for the hyperparameters. Using 5-fold cross-validation helps minimize the risk of overfitting and ensures that the performance estimates are more robust and trustworthy. The evaluation metric used in this optimization process is Mean Absolute Error (MAE), expressed in its negative form according to the scoring format in BayesSearchCV. The search space for hyperparameters for each model is detailed in Table 2.

Table 2. Hyperparameter search space for models

Parameters	Value
n_estimators	100–500
max_depth	3–20
min_samples_split	2–20
min_samples_leaf	1–10
max_features	0.3–1.0

To systematically examine the contribution of different feature types, the optimized RFR model was trained and evaluated under five distinct scenarios:

- Baseline: Using all available features, then selecting the top 15 based on mutual information.
- Temporal only: Using only time-related features.
- No temporal: Omitting all time-related features.
- No operational: Excluding operational attributes.
- No spatial: Excluding location-based features.

This ablation study allows us to analyze the impact of each feature group on model performance. The tuned RFR models were then tested on a separate set of data using three common measures for regression performance.

- R-squared (R^2) Score: This statistic measures the proportion of variance in the dependent variable that can be explained by the independent variables. A higher R^2 value indicates a better fit.
- MAE: Measures the average magnitude of the errors in a set of predictions, without considering their direction. It represents the average absolute difference between predicted and actual values. A lower MAE indicates higher accuracy.
- Mean Squared Error (MSE): Measures the average of the squares of the errors. It assigns higher weight to larger errors. A lower MSE indicates higher accuracy, and it is sensitive to outliers.

These measurements together present a complete evaluation of how accurate the models are, how well they explain the data, and how errors are spread out when predicting LRT departure delays. All implementations were carried out in Python using scikit-learn for model development and evaluation and scikit-optimize for Bayesian hyperparameter tuning.

4 Result and Discussion

This section presents the results of experiments applying the RFR model to predict Canberra LRT departure delays based on five different feature scenarios. The analysis begins with a visual comparison between predicted

values and actual values. Next, a quantitative evaluation is conducted using standard regression metrics (R^2, MAE, and MSE) to assess the model's performance across each scenario. Finally, the most influential features are analyzed based on the scenario with the best performance to identify the key factors affecting the prediction results.

4.1 Qualitative Assessment of Prediction Performance

Figures 2 3 4 5 and 6 visually compare the predicted departure delays against the actual observed delays for the RFR-FE model under five different scenarios on the test dataset: (1) RFR-FEFS, (2) using only temporal features, (3) excluding temporal features, (4) excluding operational features, and (5) excluding spatial features, respectively.

Upon closer examination, Figure 2 illustrates the performance of the RFR-FEFS model by juxtaposing the predicted delay values (in orange) with the actual values (in blue) from the test dataset. The orange line closely adheres to the blue line throughout the entire range, indicating a strong correlation between the model's predictions and actual outcomes. It accurately captures both minor fluctuations and significant delays, demonstrating its ability to replicate the temporal patterns within the data. While there are some minor discrepancies with extreme outliers, the overall alignment between predictions and actual outcomes indicates that the RFR-FEFS model demonstrates substantial accuracy.

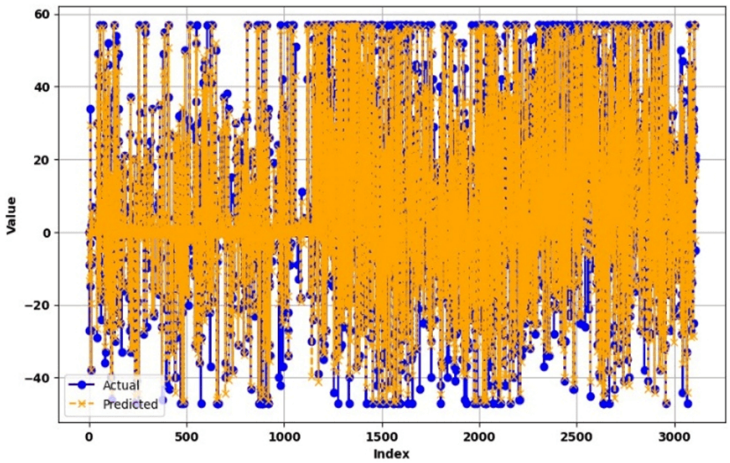

Figure 2. Comparison of actual and predicted model RFR-FEFS

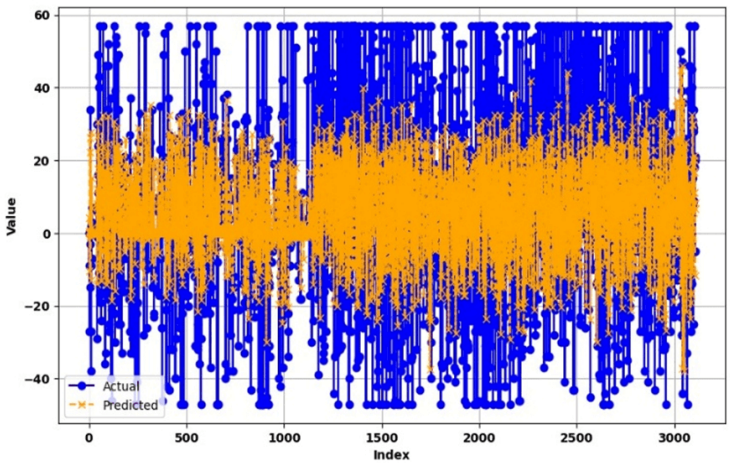

Figure 3. Comparison of actual and predicted model RFR-FE with temporal features

Figure 3 illustrates the performance of the RFR-FE model when exclusively utilizing temporal features, in sharp contrast to the feature extraction and selection scenario. The forecasted values are closely grouped around lower delay magnitudes, demonstrating significantly less variability than the actual values, which exhibit substantial fluctuations and sudden spikes. This highlights that temporal features can effectively identify general time-related delay patterns, yet they neglect the complex operational and spatial factors involved when considered in isolation. As a result, forecasts relying solely on temporal features lack the necessary precision in this regression context.

As we jump to another result, as shown in Figure 4, it illustrates that the RFR-FE model's performance lacks temporal features. The forecasted values consistently correspond with the actual delays, indicating that the model can explain a significant portion of the variability even in the absence of temporal attributes. Nonetheless, significant discrepancies remain, particularly for extreme positive and negative delays, where the predictions demonstrate diminished accuracy. This suggests that, while the model possesses robust predictive abilities, temporal features provide significant additional explanatory power in identifying periodic patterns and improving accuracy for outlier instances. In conclusion, the results indicate that temporal features enhance the model's efficacy, although their absence does not completely diminish its predictive capability.

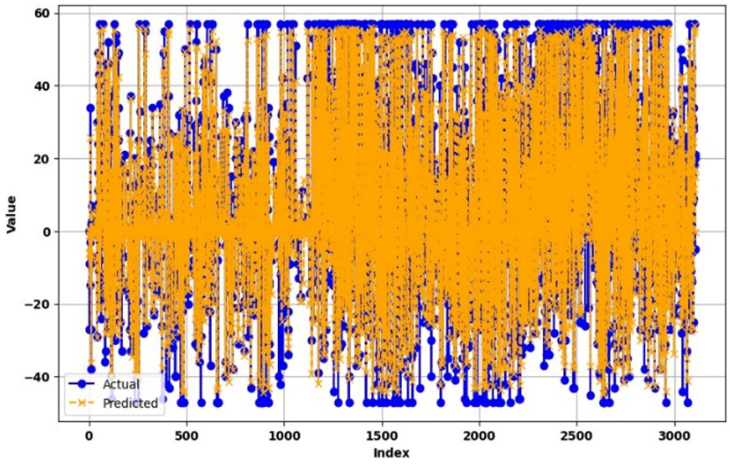

Figure 4. Comparison of actual and predicted model RFR-FE without temporal features

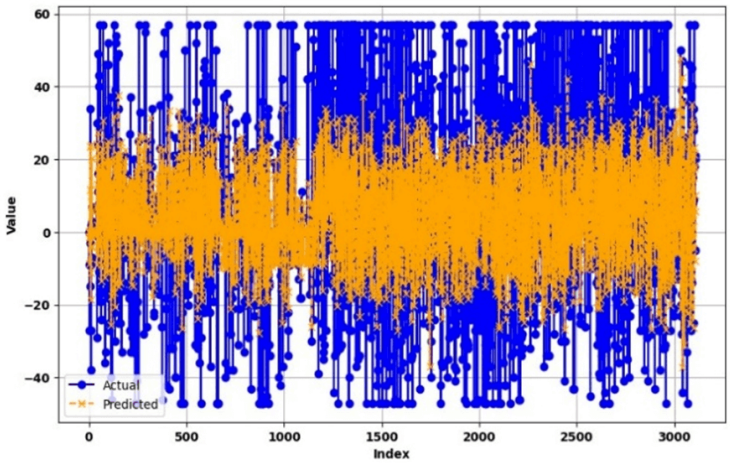

Figure 5. Comparison of actual and predicted model RFR-FE without operational features

Figure 5 presents a comparison between actual values and predicted values produced by the RFR-FE model devoid of operational features. The model's predictions cluster near zero with limited variation, whereas the actual values exhibit significant fluctuations, including numerous extreme positive and negative outliers. This condition signifies that in the absence of operational features, the model fails to identify and depict the complex relationships inherent in the actual data. The concentrated prediction values indicate that the model neglects critical information typically supplied by operational features, such as real-time conditions, vehicle status, or system interventions, which can substantially influence delays.

Figure 5 illustrates the efficacy of the RFR-FE model in the absence of operational features. In this scenario, the model's predictions significantly diverge from the actual values, predominantly concentrating on lower delay magnitudes and inadequately reflecting the substantial variability present in the real-world data. The exclusion of essential operational details, including vehicle status, odometer readings, and congestion levels, significantly limits the model's ability to accurately represent the fundamental dynamics of delays. As a result, the forecasts are demonstrated to be inaccurate and inadequately calibrated to actual conditions, especially in cases of moderate to

extreme delays. These results highlight the essential role of operational features in improving model accuracy and ensuring that forecasts accurately reflect the varied fluctuations in service operations.

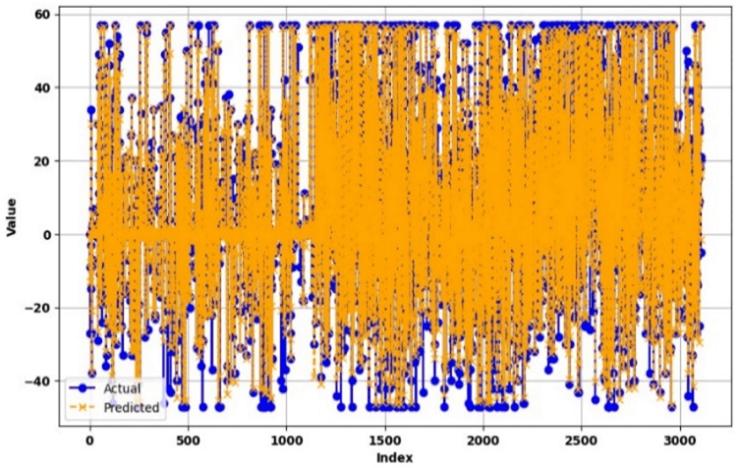

Figure 6. Comparison of actual and predicted model RFR-FE without spatial features

Figure 6 depicts the performance of the RFR-FE model in the absence of spatial features. The forecasts remain consistent with the overall pattern of actual delays and demonstrate significant variability. Nevertheless, they exhibit diminished accuracy in addressing extreme instances of both positive and negative delays. In comparison to the comprehensive model, these results seem more centralized and less influenced by geographic factors. This highlights the significant contextual insights offered by spatial information, including location, latitude, longitude, bearing, and stop sequence, which allow the model to accurately emulate real-world delay dynamics. Without these features, the model maintains its capacity to identify general patterns but loses precision in capturing location-specific variations.

4.2 Quantitative Evaluation and Model Comparison

To provide a rigorous assessment of model performance, quantitative metrics were calculated for each scenario on both the training and test datasets. Table 3 summarizes the R^2 score, MAE, and MSE across all scenarios.

Table 3. Performance evaluation across different feature scenarios

Scenario	Dataset	R^2	MAE	MSE	Computation Time (s)
RFR-FEFS	Train	0.99	0.986	3.85	728
	Test	0.94	2.934	34.32	
RFR-temporal only	Train	0.84	5.435	463.3	1065
	Test	0.20	14.46	83.43	
RFR-without temporal	Train	0.93	3.365	33.95	1086
	Test	0.90	4.574	55.44	
RFR-without operational	Train	0.91	4.021	44.99	1216
	Test	0.23	14.23	446.8	
Without spatial	Train	0.99	0.870	38.11	1313
	Test	0.93	3.449	3.449	

According to the quantitative metrics in Table 3, the RFR-FEFS scenario achieved the strongest overall performance. On the test set, it reached an R^2 of 0.94 with a relatively low MAE of 2.934 and MSE of 34.32, indicating that the model captures most of the variability in departure delays with solid accuracy. The close alignment between training (R^2 = 0.99) and test (R^2 = 0.94) further suggests that the model generalizes well and does not suffer from severe overfitting.

In contrast, the temporal-only configuration performed poorly on unseen data. While the training R^2 was 0.84, the test R^2 collapsed to only 0.20, accompanied by much larger errors (MAE = 14.46, MSE = 463.3). This large train–test gap indicates overfitting; the model learns patterns from the training data but fails to transfer them effectively to new samples. This confirms that relying solely on temporal features is insufficient for robust prediction.

The elimination of temporal features also detrimentally affected the results, albeit to a lesser extent. The test R^2 decreased to 0.90, accompanied by increased errors (MAE = 4.574, MSE = 55.44). This study affirms the

significance of temporal attributes; however, the comparatively smaller train-test gap suggests a more balanced yet still compromised model.

Excluding operational features produced the most severe deterioration. The test R^2 plummeted to 0.23 with MAE = 14.23 and MSE = 446.8, reflecting largely unreliable predictions. Here, the training R^2 reached 0.91, highlighting a clear case of overfitting where the model performed well on training data but collapsed on unseen samples.

Meanwhile, the no-spatial scenario retained fairly strong performance (test R^2 = 0.93, MAE = 3.449). This shows that while spatial features add valuable context for location-specific dynamics, temporal and operational attributes contribute more critically to prediction accuracy.

Previous research conducted by Sarhani and Voß [7] indicated that the SVR model demonstrated excellent predictive performance on the same dataset. In alignment with these findings, our study indicates that the RFR model, which is enhanced by comprehensive feature engineering that includes both extraction and selection, achieves optimal results even without incorporating weather data into the analysis. This result highlights that strong predictive accuracy can be achieved through refined feature processing alone. The correlation between previous research and our findings demonstrates the vital role of feature selection in developing robust predictive models, and it also illustrates that feature selection's effectiveness is further enhanced when combined with feature extraction.

Overall, the comparison across scenarios confirms that combining all feature groups produces the most accurate and balanced outcomes. Nonetheless, these findings are specific to the Canberra LRT dataset and should not be generalized to other light rail systems without further validation.

4.3 Feature Importance

Next, we performed a feature importance analysis on the model with the best performance, namely RFR-FEFS, to identify the individual contributions of each variable to its predictive performance. The results of this analysis are visualized in Figure 7, which represents the relative importance of each feature used by the model.

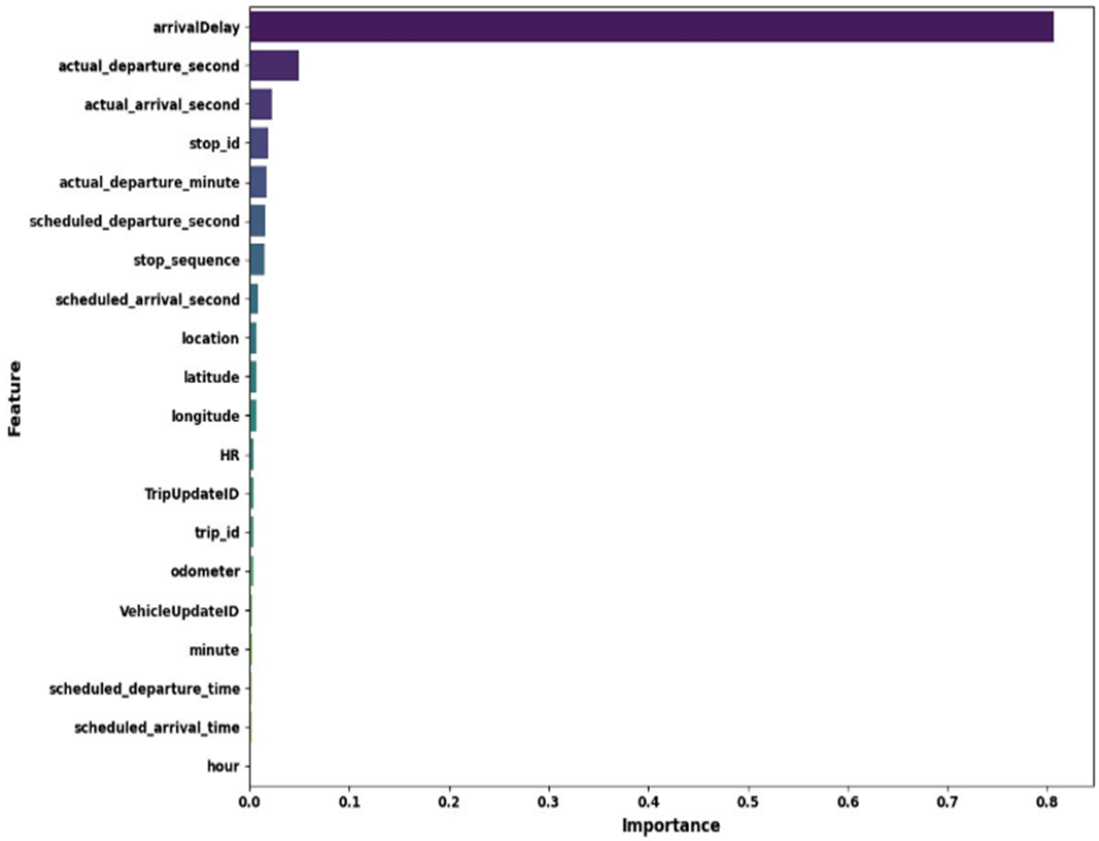

Figure 7. Feature importance in the scenario RFR-FEFS model

The feature importance analysis of the RFR-FEFS model, which performed the best, indicates that arrivalDelay is the most significant predictor, standing out by a considerable margin. This finding is logical, as a train that arrives late at a stop is very likely to depart late as well, due to the sequential relationship between arrival and departure events. Therefore, arrival delay acts as the primary variable for predicting departure delay, which clarifies why its importance score is much higher than that of any other feature.

In addition, features such as actual_departure_second and actual_arrival_second also rank relatively high, offering detailed timing information that helps capture short-term variations in delay at the second level. Other variables, including stop_id, stop_sequence, scheduled arrival and departure seconds, as well as spatial indicators like latitude, longitude, and location, contribute smaller yet meaningful signals. These features provide important contextual and operational nuances that enhance the model's ability to understand delay dynamics.

Conversely, variables such as hour, minute, odometer, and update IDs exhibit relatively low importance scores. Although their individual contributions are modest, they still offer complementary information in specific scenarios, enabling the model to recognize subtle operational contexts.

Overall, this analysis supports the conclusions drawn from the ablation study: while arrivalDelay is the single most dominant predictor, optimal predictive performance is achieved only when temporal, operational, and spatial features are integrated. The removal of any of these feature groups consistently led to a decline in model performance in previous experiments, demonstrating that each group provides unique and indispensable contributions. Therefore, constructing a reliable and generalizable predictive model for departure delays necessitates the combination of all three feature dimensions to accurately reflect the complexity of real-world transit operations.

4.4 Discussion and Implications

The extensive testing results indicate that integrating various data sources, such as static and real-time GTFS data with spatial, temporal, and operational information—coupled with methodical feature engineering and selection-substantially enhances the precision of LRT departure delay predictions. The RFR-FEFS scenario consistently surpassed all other configurations, attaining high predictive accuracy on the test set (R^2 = 0.94, MAE = 2.934, MSE = 34.32). The minor discrepancy between training and testing scores indicates that the model generalizes effectively; however, the elevated performance metrics warrant scrutiny for possible overfitting.

Compared to existing studies on delay prediction, which often rely on limited features or simpler preprocessing methods, this approach highlights the advantages of leveraging richer data and more advanced processing techniques. The Random Forest algorithm's robustness—known for capturing non-linear interactions and reducing sensitivity to noise—was instrumental in enhancing predictive stability. At the same time, the systematic feature engineering process, particularly the extraction of fine-grained temporal details (such as seconds and minutes from actual arrival and departure times), proved essential for modeling short-term dynamics closely linked to delay patterns.

Feature selection based on mutual information refined the model by identifying the most relevant predictors, effectively balancing accuracy and computational efficiency. The feature importance analysis confirmed that the variable arrivalDelay is the strongest predictor, aligning with operational intuition: a train arriving late at a stop is likely to depart late as well. However, the analysis also revealed that operational and spatial features, despite having smaller individual importance scores, provide valuable complementary context. The removal of these features resulted in significant declines in predictive accuracy, particularly when operational features were excluded, highlighting their essential role in delay estimation.

Together, these findings confirm that no single feature dimension, temporal, operational, or spatial, can be omitted without negatively affecting prediction quality. While temporal information offers the most powerful signals, spatial and operational variables enrich the model with location-specific and system-condition insights that are critical for real-world applicability.

This study demonstrates that integrating static and real-time GTFS data with rigorous feature engineering, feature selection, and ensemble-based machine learning yields highly accurate and reliable predictions of LRT departure delays. The RFR-FEFS model emerged as the best-performing configuration, validating the effectiveness of this data-driven approach. Beyond its methodological contributions, the results have practical implications: such predictive models can assist transit operators in proactive operational management, help mitigate service disruptions, and ultimately enhance the passenger experience in urban rail systems.

5 Conclusion

A proficient and dependable public transportation system, especially LRT, is critical to promoting sustainable urban mobility. This study established a data-driven framework designed to forecast departure delays by incorporating both static and real-time GTFS data. The framework was augmented through methodical feature engineering and feature selection. The Random Forest model, employing feature extraction and selection, proved to be the most effective configuration, exhibiting robust predictive performance with an R^2 of 0.94, an MAE of 2.934, and an MSE of 34.32 on the Canberra dataset. The findings validate that the integration of temporal, operational, and spatial information provides complementary insights, and the omission of any one of these feature groups leads to a substantial decrease in performance.

Beyond reaffirming previous research, this study offers new evidence that high predictive accuracy can be attained even without incorporating external data such as weather, as long as core GTFS features are thoroughly processed.

This highlights the critical importance of feature engineering and selection in shaping model performance, especially for systems where additional contextual data may not always be accessible.

However, several limitations should be acknowledged. The framework was validated solely on the Canberra LRT dataset, which raises questions about its generalizability to other networks with different infrastructure, passenger demand, and operational characteristics. Additionally, potentially influential factors such as detailed weather conditions, unplanned incidents, or special events were not included, limiting the model's ability to capture certain causes of delay. These omissions represent trade-offs: while the model demonstrates robustness using core GTFS data, it may underperform in situations where external disruptions have a dominant impact.

Future research should address these limitations by evaluating the model's transferability across different LRT or metro systems, incorporating richer contextual variables (e.g., weather, traffic disruptions, and passengers), and exploring advanced techniques. These extensions would improve the robustness of predictions and offer more information about the adaptability of data-driven delay models.

In summary, this study demonstrates that leveraging static and real-time GTFS data combined with systematic feature processing and ensemble learning produces highly accurate and reliable predictions of departure delays. While the results are promising, practical deployment should proceed cautiously, supported by validation across diverse systems and conditions. This framework lays a solid foundation for developing predictive tools that can ultimately aid proactive operations management, enhance service reliability, and improve the passenger experience in urban transit systems.

Author Contributions

Conceptualization, R.P.; methodology, R.P. and M.D.M.; software, A.P.A., N.S., and M.D.M.; validation, H.V.; formal analysis, I.I.; investigation, M.D.M.; data curation, N.S.; writing—original draft preparation, A.P.A; writing—review and editing, R.P. and H.V.; supervision, R.F.R.; All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Kopsidas, C. Milioti, K. Kepaptsoglou, and E. I. Vlachogianni, "How did the COVID-19 pandemic impact traveler behavior toward public transport? The case of Athens, Greece," *Transp. Lett.*, vol. 13, no. 5-6, pp. 344–352, 2021. https://doi.org/10.1080/19427867.2021.1901029

[2] T. Lavery and P. Kanaroglou, "Rediscovering light rail: Assessing the potential impacts of a light rail transit line on transit oriented development and transit ridership," *Transp. Lett.*, vol. 4, no. 4, pp. 211–226, 2012. https://doi.org/10.3328/TL.2012.04.04.211-226

[3] M. S. Artan and I. Sahin, "A comparative analysis of train delay prediction models for Markov chains," *Transp. Res. Procedia*, vol. 82, pp. 822–835, 2025. https://doi.org/10.1016/j.trpro.2024.12.100

[4] S. Vafaei and M. Yaghini, "Online prediction of arrival and departure times in each station for passenger trains using machine learning methods," *Transp. Eng.*, vol. 16, p. 100250, 2024. https://doi.org/10.1016/j.treng.2024.100250

[5] Z. Chen, Y. Wang, and L. Zhou, "Predicting weather-induced delays of high-speed rail and aviation in China," *Transp. Policy*, vol. 101, pp. 1–13, 2021. https://doi.org/10.1016/j.tranpol.2020.11.008

[6] P. L. Bokonda, K. Ouazzani-Touhami, and N. Souissi, "Predictive analysis using machine learning: Review of trends and methods," in *2020 International Symposium on Advanced Electrical and Communication Technologies (ISAECT 2020), Marrakech, Morocco*, 2020, pp. 1–6. https://doi.org/10.1109/ISAECT50560.2020.9523703

[7] M. Sarhani and S. Voß, "Prediction of rail transit delays with machine learning: How to exploit open data sources," *Multimodal Transp.*, vol. 3, no. 2, p. 100120, 2024. https://doi.org/10.1016/j.multra.2024.100120

[8] L. Y. Liu and H. J. Miller, "Measuring risk of missing transfers in public transit systems using high-resolution schedule and real-time bus location data," *Urban Stud.*, vol. 58, no. 15, pp. 3140–3156, 2021. https://doi.org/10.1177/0042098020919323

[9] L. Webb, G. Higgs, M. Langford, and R. Berry, "Evaluating real-time and scheduled public transport data: Challenges and opportunities," *ISPRS Int. J. Geo-Inf.*, vol. 14, no. 7, p. 243, 2025. https://doi.org/10.3390/ijgi14070243

[10] J. Q. Wu, B. Du, Z. Y. Gong, Q. Wu, J. Shen, L. P. Zhou, and C. Cai, "A GTFS data acquisition and processing framework and its application to train delay prediction," *Int. J. Transp. Sci. Technol.*, vol. 12, no. 1, pp. 201–216, 2023. https://doi.org/10.1016/j.ijtst.2022.01.005

[11] S. Tufail, H. Riggs, M. Tariq, and A. I. Sarwat, "Advancements and challenges in machine learning: A comprehensive review of models, libraries, applications, and algorithms," *Electronics*, vol. 12, no. 8, p. 1789, 2023. https://doi.org/10.3390/electronics12081789

[12] Z. Guo, B. Yu, M. Y. Hao, W. S. Wang, Y. Jiang, and F. Zong, "A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient," *Aerosp. Sci. Technol.*, vol. 116, p. 106822, 2021. https://doi.org/10.1016/j.ast.2021.106822

[13] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," *J. Inf. Intell.*, vol. 3, no. 2, pp. 113–153, 2024. https://doi.org/10.1016/j.jiixd.2024.01.002

[14] M. E. Biswas, T. Sultana, A. K. Mandal, M. Golam Morshed, and M. D. Hossain, "Spatio-temporal feature engineering and selection-based flight arrival delay prediction using deep feedforward regression network," *Electronics*, vol. 13, no. 24, p. 4910, 2024. https://doi.org/10.3390/electronics13244910

[15] W. Albattah, R. U. Khan, M. F. Alsharekh, and S. F. Khasawneh, "Feature selection techniques for big data analytics," *Electronics*, vol. 11, no. 19, p. 3177, 2022. https://doi.org/10.3390/electronics11193177

[16] M. Büyükkeçeci and M. C. Okur, "A comprehensive review of feature selection and feature selection stability in machine learning," *Gazi Univ. J. Sci.*, vol. 36, no. 4, pp. 1506–1520, 2023. https://doi.org/10.35378/gujs.993763

[17] M. A. Islam, M. Z. H. Majumder, M. S. Miah, and S. Jannaty, "Precision healthcare: A deep dive into machine learning algorithms and feature selection strategies for accurate heart disease prediction," *Comput. Biol. Med.*, vol. 176, p. 108432, 2024. https://doi.org/10.1016/j.compbiomed.2024.108432

[18] D. Ramadhani, A. M. Soleh, and E. Erfiani, "Machine learning-based univariate time series imputation method for estimating missing values in non-stationary data," *J. Mat. Stat. Komput.*, vol. 21, no. 1, pp. 307–320, 2024. https://doi.org/10.20956/j.v21i1.36468

[19] I. Ahmed, I. Kumara, V. Reshadat, A. S. M. Kayes, W. J. van den Heuvel, and D. A. Tamburri, "Travel time prediction and explanation with spatio-temporal features: A comparative study," *Electronics*, vol. 11, no. 1, p. 106, 2021. https://doi.org/10.3390/electronics11010106

[20] A. Udil, "Encoding methods for categorical data: A comparative analysis for linear models, decision trees, and support vector machines," Bachelor's thesis, Delft University of Technology, The Netherlands, 2023.

[21] L. Huang, X. Q. Zhou, L. H. Shi, and L. Gong, "Time series feature selection method based on mutual information," *Appl. Sci.*, vol. 14, no. 5, p. 1960, 2024. https://doi.org/10.3390/app14051960

[22] C. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outlier's detection and elimination framework in classification task of data mining," *Decis. Anal. J.*, vol. 6, p. 100164, 2023. https://doi.org/10.1016/j.dajour.2023.100164

[23] J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, R. V. Godoy, L. A. Ambrosio, and M. Becker, "The impact of feature scaling in machine learning: Effects on regression and classification tasks," *arXiv preprint*, 2025, Art. no. arXiv:2506.08274. https://doi.org/10.48550/arXiv.2506.08274