ARCHIVES OF TRANSPORT *ISSN (print): 0866-9546*
Volume 77, Issue 1, 2026 ***e-ISSN (online):*** *2300-8830*
DOI: *10.61089/aot2026.mnrkh112*

INTERPRETABLE MACHINE LEARNING FOR IDENTIFYING THE DETERMINANTS OF BUS DELAYS

Viktor MARTYANOV[1], Artur BUDZYŃSKI[2], Andrzej CZEREPICKI[3], Maciej KOZŁOWSKI[4]

[1, 3, 4] *Warsaw University of Technology, Faculty of Transport, Warsaw, Poland*

[2] *Silesian University of Technology, Faculty of Transport and Aviation Engineering, Katowice, Poland*

Abstract:

The study proposes an interpretable machine-learning approach for predicting bus arrival delays based on the Automatic Vehicle Location (AVL) system. The data describe vehicle movements and schedule adherence at successive stops, enabling delay estimation both along the route and across different time periods. Delays are defined as deviations between observed arrival times and scheduled times at individual stops, allowing both spatial and temporal delay patterns to be examined. The approach integrates ensemble machine-learning models with nested, temporally aware cross-validation and hyperparameter optimization. This validation strategy preserves the temporal order of observations, ensuring that model evaluation reflects realistic forecasting conditions and avoids information leakage. In particular, training and test sets are separated along the time axis, preventing future observations from influencing model training. Three ensemble models—Random Forest, XGBoost, and CatBoost—achieved comparable accuracy (RMSE $\approx$ 5.4 min; MAE $\approx$ 3.6 min; $R^2 \approx 0.44$). Model performance was evaluated at the level of individual stop arrivals, reflecting short-term delay prediction under operational conditions. Comparable performance across models indicates that the observed patterns are not specific to a single modelling technique. This consistency suggests that the identified delay patterns are robust with respect to model choice rather than driven by algorithm-specific assumptions. The results show that the proposed approach enables the identification of combined temporal and spatial feature interactions associated with bus delay magnitudes within the analysed case study. These interactions mainly involve time-of-day effects and stop sequence, illustrating how delays accumulate as vehicles progress along the route. This pattern reflects the cumulative nature of delays in mixed-traffic operations, where early disturbances propagate toward downstream stops. SHAP values were used to provide a quantitative interpretation of these interactions and to identify the most influential predictors in the analysed models. The analytical structure, designed around weekly data partitions, supports delay monitoring and operational analysis. The results indicate that the proposed approach can support the analysis and interpretation of bus delay patterns within similar operational contexts.

Keywords: *public transport reliability, travel time variability, SHAP, temporal cross-validation, urban bus operations*

To cite this article:

Martyanov, V., Budzyński, A., Czerepicki, A., Kozłowski, M. (2025). Interpretable machine learning for identifying the determinants of bus delays. Archives of Transport, 77(1), 7-26. https://doi.org/10.61089/aot2026.mnrkh112

Contact:
1) Viktor.Martyanov@pw.edu.pl [https://orcid.org/0009-0000-6538-1690]; 2) Artur.Budzynski@polsl.pl [https://orcid.org/0000-0002-5803-6749]; 3) andrzej.czerepicki@pw.edu.pl [https://orcid.org/0000-0002-8659-5695] – corresponding author; 4) maciej.kozlowski@pw.edu.pl [https://orcid.org/0000-0002-1068-8991]

1. Introduction

Urban bus services are crucial in ensuring accessibility, mobility, and sustainability within metropolitan transport systems (Toledano et al., 2025). However, one of the persistent challenges in maintaining reliable bus operations is the variability of travel times, which manifests as schedule delays. Delays reduce passenger satisfaction, increase operational costs, and decrease the overall efficiency of public transport systems (Drabicki et al., 2021; Yap & Cats, 2021). Understanding and quantifying these delays has become a central topic in analysing and managing public transit performance.

Numerous studies have examined bus delays from operational, infrastructural, and behavioural perspectives. (Chen et al., 2025) analysed temporal variations in bus delays under diverse weather and traffic conditions using Bayesian probabilistic models, revealing that external and internal factors jointly shape delay dynamics. Similarly, Zhang et al. (Q. Zhang et al., 2024) employed seemingly unrelated regression equations (SURE) to model delay propagation along routes, emphasizing spatial dependencies between consecutive stops. Park et al. (Park et al., 2020) explored real-time deviations between scheduled and actual bus arrivals, highlighting how localized disturbances in space and time can propagate and cause network-wide unreliability. From a causal standpoint, Zhang et al. (Q. Zhang et al., 2025) utilized causal graph analysis to identify direct and indirect relationships between operational variables, identifying chain reactions leading to delay accumulation.

Beyond delay modelling, researchers have also addressed robustness and resilience in public transport systems. (Ge et al., 2022) discussed the ability of public transport networks to absorb disturbances without major service degradation, while (Rezazada et al., 2024) provided a comprehensive review of the bus bunching phenomenon, framing it as a systemic manifestation of uncontrolled delay amplification. (Rosenblum et al., 2015) developed a segment-level approach for identifying congestion bottlenecks using mixed-traffic delay metrics. (Almeida et al., 2023) demonstrated how bus tracking data can effectively reveal urban congestion patterns and their operational impacts.

Despite the widespread use of mathematical and computational methods for predicting bus delays, a substantial portion of existing studies focuses primarily on improving predictive accuracy. As a result, limited attention is given to model interpretability and to explaining the operational mechanisms underlying observed delays. From a transport engineering perspective, this limits the ability to translate predictive model outputs into actionable knowledge for analysing bus line operations, particularly with respect to temporal–spatial relationships and delay propagation processes. This highlights the need for analytical approaches that combine delay prediction with interpretation in a way that supports a deeper understanding of operational dynamics.

The fundamental operational challenge addressed in this study is the stochastic nature of bus travel times in mixed-traffic conditions, which leads to service irregularity and the "bus bunching" phenomenon. On long, circumferential routes like the analyzed Line 112, small initial deviations tend to amplify non-linearly as the vehicle progresses, making it difficult for dispatchers to distinguish between random incidents and systematic structural delays. Traditional scheduling methods often fail to capture these complex spatiotemporal interactions. Therefore, the specific research problem is the lack of analytical tools that can not only predict the magnitude of delay with high accuracy but also strictly quantify the contribution of temporal and spatial determinants to this deviation, thereby supporting evidence-based timetable adjustments.

The objective of this study is to develop and validate an interpretable machine-learning framework for analysing bus arrival delays based on AVL data. The proposed framework integrates ensemble methods, temporally aware cross-validation, and model-interpretation techniques to ensure both predictive robustness and operational usefulness. The applicability of the framework is demonstrated through an empirical case study of Bus Line 112 in Warsaw.

The originality of this study lies in integrating interpretable machine learning techniques with a temporally nested validation framework and operationally oriented data preprocessing. Unlike most previous analyses of public transport reliability, the modelling process explicitly applies nested temporal cross-validation, ensuring realistic forecasting performance estimates without data leakage across time. The study also introduces a route-variant identification procedure that reconstructs the spatial

structure of services directly from AVL and timetable data, enabling fine-grained differentiation between line variants. Furthermore, the use of SHAP (SHapley Additive Explanations) allows translating model results into interpretable, operationally meaningful insights, highlighting how temporal–spatial factors jointly determine bus delays. Collectively, these elements form a reproducible analytical framework that bridges methodological rigour with practical applicability in transit performance assessment.

2. Literature review

In recent years, research on public transport has increasingly focused on understanding and predicting delays in bus operations due to their significant impact on service reliability and schedule adherence. A number of studies have developed predictive models for bus travel and arrival times using machine-learning and statistical approaches, highlighting persistent challenges in capturing real-world variability and improving operational performance. For example, Serin et al. demonstrated the application of layered machine-learning architectures for bus arrival prediction, indicating the value of data-driven models for real-time forecasting (Serin et al., 2022). Aemmer et al. proposed methods for extracting and analysing transit performance metrics from real-time feeds (GTFS-RT), illustrating how data streams can be used to quantify systematic and stochastic delays across a network Mobility Innovation Centre (Aemmer et al., 2022).

Research on bus delay and travel time prediction is dominated by predictive modelling approaches that prioritize high forecasting accuracy using historical and real-time vehicle movement data. Early studies leveraged statistical and regression-based models to estimate journey times and deviations from schedule based on Automatic Vehicle Location (AVL) data, finding that machine learning methods often outperform traditional regression techniques in terms of predictive performance (Jeong & Rilett, 2005). A wide range of studies demonstrate that machine intelligence approaches such as Random Forest, Gradient Boosting, and neural networks consistently yield superior accuracy compared with simpler baselines or historic averages, especially when evaluated using error metrics (Sun et al., 2025). For example, machine learning and AI-based techniques have become central in developing real-time bus arrival time information systems, with models trained on AVL and GTFS data to capture complex spatiotemporal dependencies (Singh & Kumar, 2022).

Despite the growing effectiveness of machine learning algorithms in predicting delays, relatively few studies focus on model interpretability and explaining the operational mechanisms behind predictions. Recent research demonstrates the potential of explainable AI techniques in this domain. For example, (Vijaya et al., 2024) applied SHAP and LIME to Random Forest and kNN models to identify key features such as time of day, weekday, and route characteristics, enabling both global and local interpretation of bus delay predictions. Similarly, (Matseliukh et al., 2025) used XGBoost combined with SHAP to analyse delay factors in urban bus operations, showing that operational cycles and spatial context dominate over weather-related influences. Additionally, (Warnakulasuriya et al., 2024) introduced an explainable bus arrival time prediction framework that incorporates topography and points of interest, highlighting the practical value of interpretable models for transport management.

While recent studies have improved predictive accuracy and, in some cases, introduced explainable AI techniques, these developments often remain fragmented. Interpretability, temporally consistent validation, and operationally grounded feature engineering are typically addressed separately rather than within a unified analytical framework. As a result, existing approaches rarely combine: interpretable modelling, leakage-free temporal validation, and transport-oriented analytical interpretation in a reproducible structure suitable for planning practice. This integrated gap motivates the present study, which brings these elements together within a single coherent framework for bus delay analysis.

3. Materials and Methods

3.1. Software

All data processing, visualization, and modelling were done in Python 3.12.7 (Anaconda) on Windows 11 (64-bit, AMD64) with an AMD64 Family 23 Model 104 Stepping 1 CPU. Analyses were conducted interactively in Jupyter Notebook 7.2.2 (Kluyver et al., 2016).

Key packages (versions in parentheses): NumPy (1.26.4) (Harris et al., 2020) and pandas (2.2.2) (McKinney, 2010) for data wrangling; scikit-learn (1.5.1) (Pedregosa et al., 2011) for feature preparation, validation, and metrics; XGBoost (T. Chen & Guestrin, 2016) (2.1.3) and CatBoost (1.2.8) (Prokhorenkova et al., 2017) for gradient-boosting models; matplotlib (3.9.2) (Hunter, 2007) and seaborn (0.13.2) (Waskom, 2021) for visualization. The Jupyter execution stack used IPython (8.27.0) and ipykernel (6.28.0).

All computations were executed locally (no GPU or cloud resources). To ensure reproducibility, all experiments were run within the same conda environment; exact package versions are listed above.

3.2. Data

3.2.1. Data acquisition

Urban buses in Warsaw are equipped with on-board motion recorders that continuously register the vehicle's geographic position, current speed, local time, and other operational parameters collected both for real-time control and analytical purposes. The recorded data are supplemented with descriptive information, including the vehicle's fleet number, serviced route number, and operating date. Data frames are generated at a fixed sampling frequency (the standard setting is 10 Hz) and stored in the recorder's local memory. Simultaneously, each current data frame is transmitted to the operator's central server.

This configuration enables continuous visualization of the current traffic situation and supports the operator in making operational decisions related to bus traffic management (e.g., rerouting, dispatching replacement vehicles, or adjusting trip assignments). The operator's central server provides access to live data transmitted from vehicles through an application programming interface (API) using the HTTP protocol. These data are publicly available at (Warsaw City Council, n.d.).

For research purposes, a dedicated software service was developed (requiring no user interaction) to periodically retrieve data from the operator's API and store them as comma-separated value (CSV) files. Using this program, a dataset was compiled representing all bus trips performed on Line 112 during October 2024.

Based on publicly available public-transport timetables, the routes of the analysed line and the coordinates of its served stops were identified. This information was necessary to determine dwell times at stops, which form the basis for subsequent analyses. Actual arrival times at each stop were then compared with the scheduled times, and the resulting delay was calculated.

The output of the preliminary data-processing stage consisted of CSV files containing the route number, trip identifier, stop number and name, stop sequence along the route, and delay time. Each CSV file corresponds to a single full-route trip. To ensure clarity regarding the model inputs, the specific variables derived from these raw data records and used in the subsequent training process are formally defined in Table 1.

3.2.2. Operational characteristics of bus line

The study focuses on Bus Line 112, a prominent daytime route operated by the Municipal Bus Transport (MZA) under the supervision of the Public Transport Authority (ZTM) in Warsaw. The line serves as a strategic circumferential connection linking distant districts: Bemowo (western Warsaw) and Targówek (eastern Warsaw). It is one of the longest bus routes in the city, traversing five districts: Bemowo, Bielany, Żoliborz, Targówek, and Bródno (Figure 1).

Table 1. Definition of model variables

Variable Type	Feature Name	Data Type	Description
Target	arrival_delay	Continuous	Difference between actual and scheduled arrival time (min) +1
Temporal	hour_rounded	Categorical	Hour of the day (0–23) derived from schedule
	weekday	Categorical	Day of the week (Monday=0 to Sunday=6)
Spatial	stop_sequence	Numerical	Order of the stop along the route trajectory
	stop_id	Categorical	Unique identifier of the bus stop +1
	direction_id	Binary	Direction of travel (0 or 1) +1
Operational	variant_simple_id	Categorical	Simplified route variant code (e.g., '1_1', '0_2')
	vehicle_number	Categorical	Unique fleet number of the bus +1
	brigade	Categorical	Operational duty number +1

By bypassing the city centre, the line enables direct inter-district travel without the need for transfers or metro usage, making it a critical component of the transport network.

The route profile is highly heterogeneous, traversing expressways, cross-city bridges connecting the left and right banks of the Vistula River, industrial zones, and high-density residential estates. This diversity makes it a valuable subject for delay analysis. The line is characterized by high passenger exchange intensity, serving distinct groups: commuters in the morning peak, students during the day, and customers visiting large shopping centres located at both terminals (CH Marki and Karolin/Nowe Bemowo).

The line operates daily from 05:00 to 23:00. The scheduled travel time for the full route is approximately 70 minutes. The service frequency (headway) is adjusted to demand, ranging from approximately 10 minutes during peak hours on workdays to 20 minutes during off-peak periods and weekends. Due to slight differences in routing at the terminals, the route length and number of stops vary depending on the direction. Detailed operational parameters are summarized inTable 2.

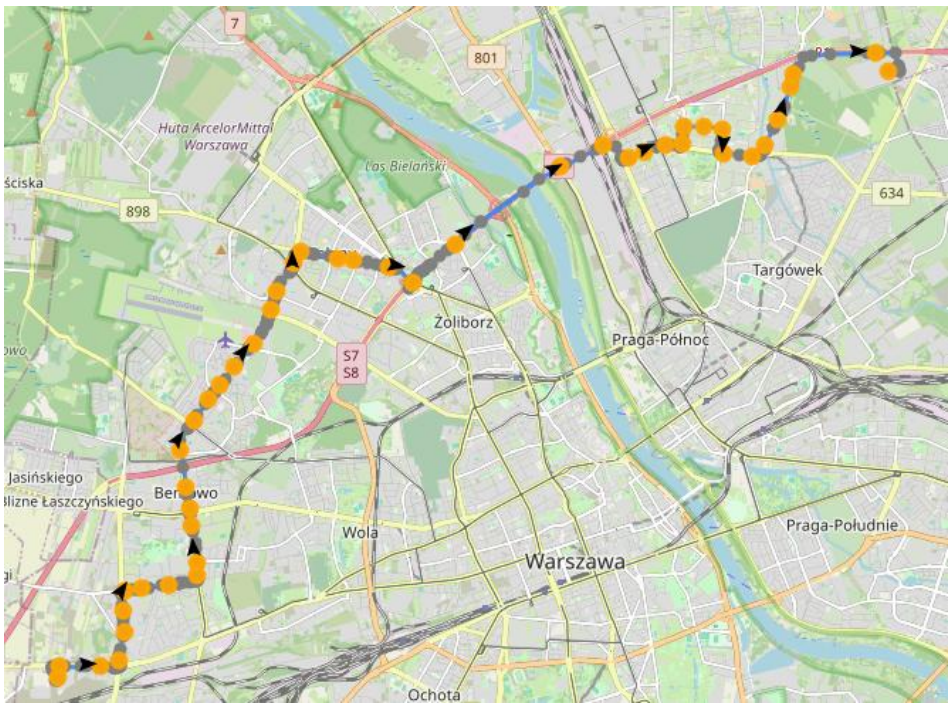

Fig. 1. Spatial layout of Bus Line 112 in Warsaw

Table 2. Operational characteristics of Bus Line 112

Parameter	**Value / Description**
Operator	MZA (Warsaw Municipal Bus Transport) for ZTM
Route Terminals	Karolin ↔ CH Marki
Route Length	23.76 km (to CH Marki) / 25.78 km (to Karolin)
Number of Stops	46 (to CH Marki) / 49 (to Karolin)
Headway (Frequency)	Peak: ~10 minOff-peak / Weekend: ~20 min
Scheduled Travel Time	Approx. 70 min
Districts Served	Bemowo, Bielany, Żoliborz, Targówek, Bródno
Key Infrastructure	S8 Expressway, Grota-Roweckiego Bridge

3.2.3. Feature engineering

Several time- and date-related features were engineered from the scheduled and actual arrival timestamps to capture temporal patterns in bus operations. From the scheduled arrival time (sched_arr_dt), the service date (service_date), hour of the day (hour), and day of the week (weekday) were extracted. These variables provide essential context for understanding systematic variations in delay behaviour, such as differences between peak and off-peak periods or weekday and weekend operations. All temporal features were derived directly from the raw timestamp values to ensure consistency across the dataset.

The process of route-variant identification was carried out using both the official timetables and the raw AVL records (Figure 2). Daily CSV logs containing stop-level delay information were merged into a single dataset and cleaned to ensure consistent data types and timestamps. Each bus run was then matched to its scheduled trip using the published timetable data, which provided reference stop sequences for all route variants. This matching procedure ensured that the spatial structure of each variant was consistent with the operational schedule rather than inferred solely from GPS data. Unique route variants were then encoded as stable alphanumeric identifiers (variant_id) and summarized by frequency, median number of stops, and terminal stop names. To simplify further analysis, each variant within a given direction was assigned a short human-readable code (variant_simple_id), ordered by the number of observed trips.

This structured representation enabled subsequent visualizations of delay profiles along the route and quantitative comparisons between the primary and secondary service variants.

Figure 3 presents the number of observed trips for each identified route variant in both directions. Five unique route variants were distinguished for each direction of Line 112. The distribution is highly unbalanced — in both directions, one dominant variant accounts for most recorded trips, while the remaining variants appear much less frequently. For direction 0, the main variant (0_1) includes approximately 1,500 trips, whereas the second variant (0_2) represents around half of that number. A similar pattern is observed in direction 1, where variant 1_1 occurs most often (about 1,400 trips), followed by variant 1_2 with roughly 700 recorded runs.

The remaining variants in both directions occur sporadically, indicating their limited operational relevance.

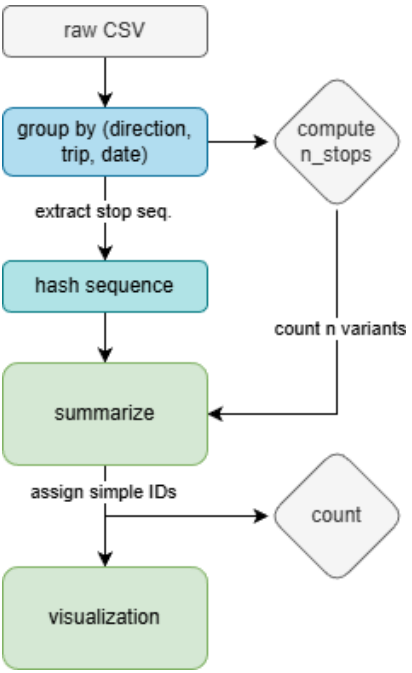

Fig. 2. Workflow for identifying unique route variants using timetable-aligned AVL data

The selection of tree-based ensemble methods (Random Forest, XGBoost, and CatBoost) was motivated by their proven effectiveness in handling tabular data involving mixed categorical and numerical features. Recent benchmarks demonstrate that gradient-boosted decision trees typically outperform deep learning architectures on structured datasets of moderate size while maintaining lower computational costs (Grinsztajn et al., 2022). Furthermore, unlike "black-box" neural networks, tree-based ensembles are directly compatible with SHAP-based explainer algorithms (TreeExplainer). This compatibility allows for the exact and efficient calculation of Shapley values, which is a core requirement for the interpretability objective of this study (S. M. Lundberg et al., 2019).

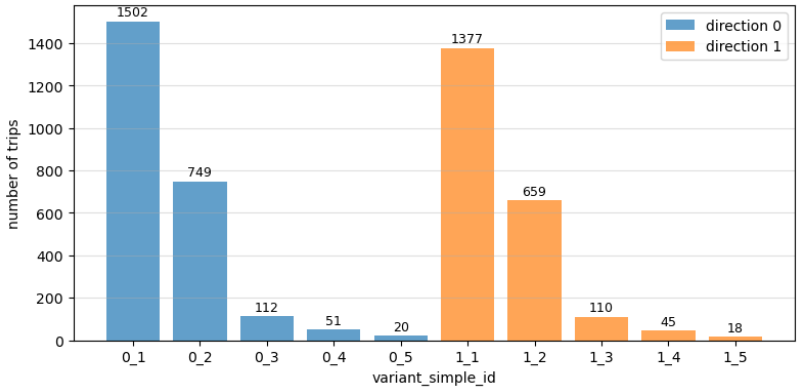

Fig. 3. Number of recorded trips per identified route variant, separated by travel direction

3.3. Models

The prediction task was defined as forecasting arrival delays for all trips and stops in the subsequent week, using historical data from previous weeks. The models were trained on past observations and evaluated on unseen data representing the next week of operation.

Three tree-based ensemble models were implemented to predict bus arrival delays: **Random Forest (RF)**, **Extreme Gradient Boosting (XGBoost)**, and **Categorical Boosting (CatBoost)**. All algorithms were used in their regression variants, as the target variable—arrival delay in min—was continuous. The choice of these models was motivated by their well-documented performance in non-linear regression tasks on tabular data, robustness to feature scaling, and their suitability for post-hoc interpretability using feature importance measures and SHAP-based analysis.

Random Forest (RF) serves as a classical ensemble baseline that aggregates predictions from multiple decorrelated decision trees trained on bootstrap samples. Each tree operates on a random subset of features, which reduces variance and prevents overfitting. The key hyperparameters tuned for this model included the number of estimators, maximum tree depth, and minimum samples per leaf (Breiman, 2001).

XGBoost extends the idea of gradient boosting by sequentially building trees that minimize a differentiable loss function, incorporating shrinkage (learning rate) and regularization to enhance generalization. It is well-suited for large datasets and supports parallel computation, making it efficient for repeated cross-validation and parameter tuning in this study.

CatBoost is a gradient boosting algorithm specifically optimized for datasets containing categorical variables. It uses ordered target statistics and symmetric trees, which help reduce prediction bias and overfitting. Unlike XGBoost, CatBoost natively handles categorical features without requiring explicit encoding, simplifying the preprocessing pipeline. Its default regularization mechanisms and efficient handling of categorical features make it particularly suitable for transport operation data, where route identifiers and weekdays play an important predictive role.

Model validation aims to estimate the generalization error of a predictive algorithm on unseen data (Varma & Simon, 2006). Conventional random k-fold cross-validation may cause information leakage for temporally ordered datasets, as future observations can unintentionally influence model training (Kaufman et al., 2012). To prevent this, temporal cross-validation (or blocked CV) was applied, ensuring that each training set contained only earlier observations relative to its test fold (Bergmeir & Benítez, 2012). Additionally, a nested cross-validation structure was adopted to avoid optimistically biased estimates after hyperparameter tuning, where the inner loop performed parameter search and the outer loop evaluated the tuned model on independent data. This combination of temporal splitting and nested validation provides realistic and reproducible performance estimates for forecasting tasks (Hyndman & Athanasopoulos, 2018).

3.4. Models interpretation

3.4.1. Feature importance

In tree-based ensemble models, feature importance quantifies the contribution of individual variables to the model's predictive performance. Traditional measures such as gain, weight, and cover are collectively so-called split-based importance metrics—evaluate how each feature improves the model's objective function during node splitting.

According to Zhou and Hooker (Zhou & Hooker, 2019), these metrics can be formally expressed through the cumulative improvement of the loss function obtained from all splits based on a given feature. Specifically, gain represents the average reduction in loss contributed by that feature, weight reflects how frequently it is used for splitting across all trees, and cover measures the number (or total weight) of samples affected by those splits. Together, these metrics describe the strength, frequency, and scope of each feature's influence within the ensemble.

However, Zhou and Hooker highlight that such split-based measures are inherently biased, as they tend to favour continuous variables or categorical attributes with many unique values. To address this limitation, they propose an unbiased estimator of feature importance based on the expected split-improvement, which provides a more consistent assessment of variable relevance across models.

3.4.2. SHAP

The SHAP (SHapley Additive exPlanations) method is used to interpret machine learning models by quantifying the contribution of individual variables to the final prediction. Its theoretical foundations originate from cooperative game theory, where each feature is considered a player contributing to the overall outcome. The Shapley values, on which SHAP is based, represent the average marginal effect of a given variable on the model output, computed across all possible combinations of the remaining features (S. Lundberg & Lee, 2017).

SHAP values indicate the extent to which each feature increases or decreases the predicted value relative to the model's expected output. The sum of SHAP values for all features equals the difference between an individual prediction and the model's mean prediction, ensuring an additive and consistent explanation of results. This framework allows both local analysis (for individual observations) and global interpretation (based on aggregated absolute |SHAP| values across the dataset).

In contrast to traditional split-based importance measures such as gain, weight, or cover, which describe only the frequency of feature usage in decision trees, SHAP directly quantifies the direction and magnitude of each variable's influence on the model outcome. In this study, the method was applied to the Random Forest model to identify the most significant determinants of bus arrival delays.

4. Results

4.1. Exploratory data analysis results

Figure 4 presents the average arrival delay by hour of day for the analysed line. The lowest mean delays occurred during early morning hours (below 3 min between 04:00 and 06:00). After 07:00, delays gradually increased, reaching approximately 5–7 min between 08:00 and 14:00. The highest average values occurred during the afternoon peak period between 15:00 and 17:00, with the maximum of about 12.5 min at 16:00. After 18:00, delays decreased steadily to approximately 2–3 min by late evening.

Figure 5 presents the average arrival delay by weekday analysed line. The highest mean delays were observed on weekdays, ranging between approximately 5 and 7.5 min. The largest average delay occurred on Wednesday (about 7.4 min), while the lowest weekday value was recorded on Monday (about 5.2 min). The mean delay noticeably reduced on weekends, reaching 5.7 min on Saturday and only 1.4 min on Sunday. Standard deviation bars indicate a high variability of delays, exceeding 10 min on working days and reaching up to 15 min on Thursday, which reflects substantial fluctuation between individual trips.

Figure 6 presents the median arrival delay for each stop along Line 112 in direction 0, calculated separately for all identified route variants. The heatmap displays the progression of median delays along the route, with color intensity indicating the delay magnitude (green – on-time or early arrivals, red – increasing delay). Stops are ordered according to their actual sequence along the route, from Nowe Bemowo (the starting terminal for the main variant, though full-length variants originate from Karolin) to CH Marki (final terminal). The figure shows that delay values generally increase as the vehicle approaches the final section of the route. At the same

time, shorter or partial variants terminate earlier, resulting in missing cells in the lower part of the plot. The visualization also highlights the relative stability of early-segment arrivals and the accumulation of delays toward the downstream stops.

Figure 7 presents the median arrivals delays by stop and route variant for direction 1, ordered according to the reference variant 1_2. The reference variant 1_2 denotes the most frequently observed full-length route variant in direction 1 and is used solely as an ordering reference for stop-level visualizations. In this direction (*CH Marki* → *Karolin*), the pattern of delays shows a gradual accumulation along the route. The lowest median values—close to zero or slightly negative—occur at the initial stops near *CH Marki*, *Geodezyjna*, and *CH Targówek*, indicating on-time or early arrivals. Delays increase progressively through the central part of the route, reaching between 4 and 6 min in the vicinity of *Metro Bródno*, *Św. Hieronima*, and *Żerań FSO*. The highest median delays, typically above 9 min, are recorded at the western terminus near *Dostawcza* and *Karolin*. Across all five route variants, the overall delay profile remains similar, with truncated variants showing slightly smaller terminal delays compared to the full-length variant 1_2.

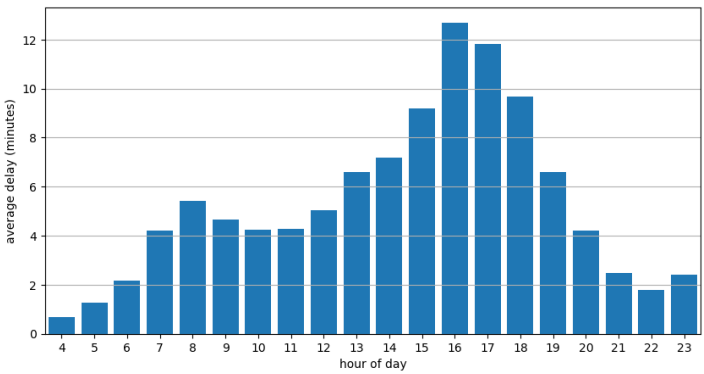

Fig. 4. Average arrival delay by hour of day for bus line 112 in October 2024

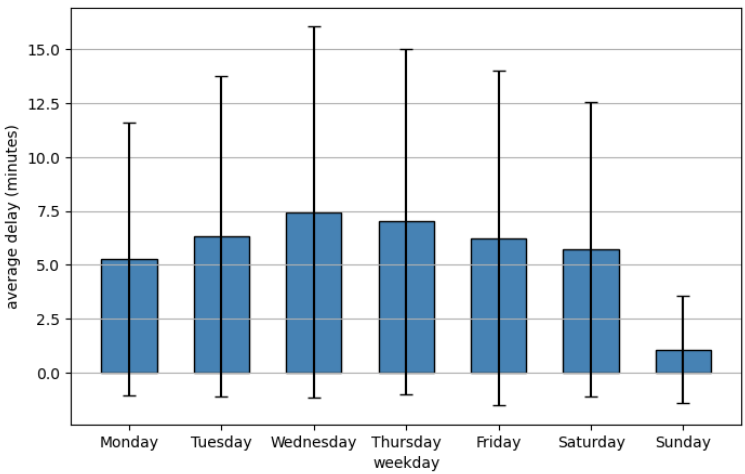

Fig. 5. Average arrival delay by weekday for bus line 112 in October 2024. Error bars represent one standard deviation

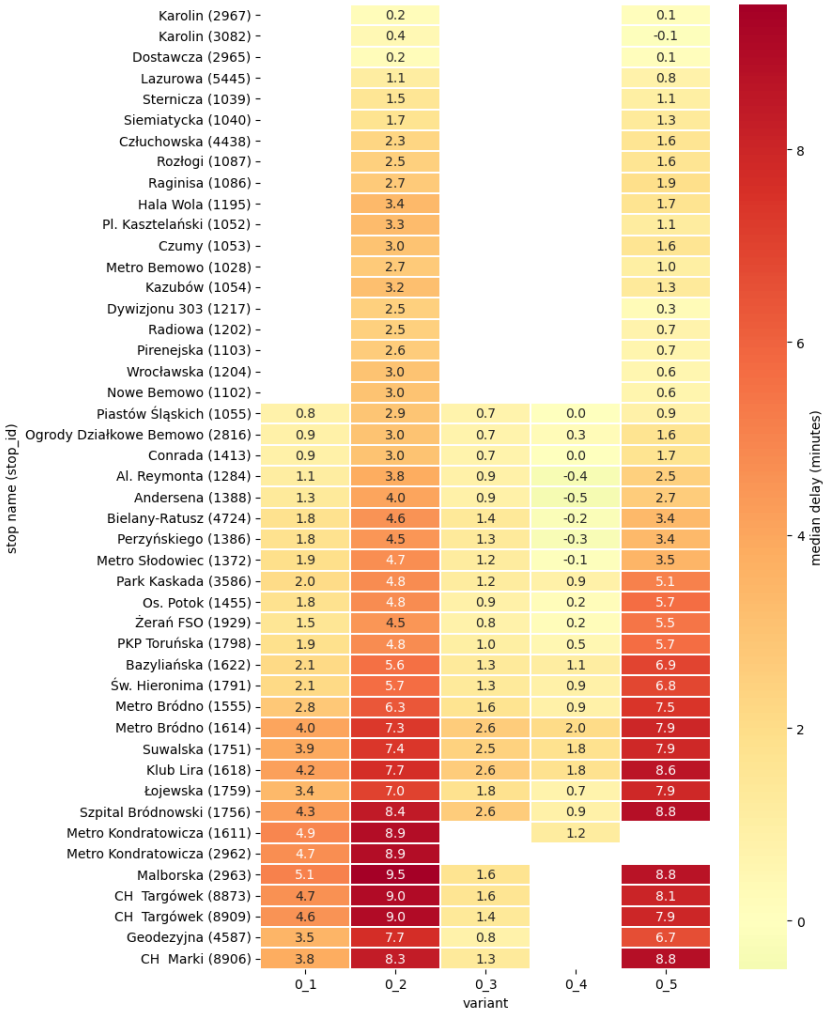

Fig. 6. Median arrivals delay by stop and route variant for Line 112 (direction 0)

4.2. Comparison of Ensemble Models for Delay Prediction

Figure 8 presents a comparison of the prediction accuracy obtained using three ensemble learning algorithms: Random Forest, XGBoost, and CatBoost.

Each bar represents the mean value of the Root Mean Squared Error (RMSE) across the outer folds of the weekly nested cross-validation, while the black lines indicate one standard deviation, illustrating the variability of model performance between weeks.

The results show that all three algorithms achieved comparable predictive performance, with RMSE values ranging from approximately 5.37 to 5.48 min.

Among them, the Random Forest model yielded the lowest mean RMSE (5.369), slightly outperforming XGBoost (5.419) and CatBoost (5.481).

Figure 9 summarizes the average Mean Absolute Error (MAE) achieved by the three evaluated models

Similarly to the RMSE results, the Random Forest algorithm provided the most accurate predictions, with a mean MAE of 3.633 min, followed by XGBoost (3.854 min) and CatBoost (3.889 min).

The black lines represent one standard deviation, reflecting the performance variability across the weekly validation folds.

Random Forest, XGBoost, and CatBoost models. Figure 10 presents the comparison of the coefficient of determination (R^2) obtained for the three ensemble models.

All algorithms explain a comparable portion of the variance in the observed bus arrival delays, with R^2 values between 0.42 and 0.45.

The Random Forest model achieved the highest average coefficient (0.445), followed closely by XGBoost (0.435) and CatBoost (0.420).

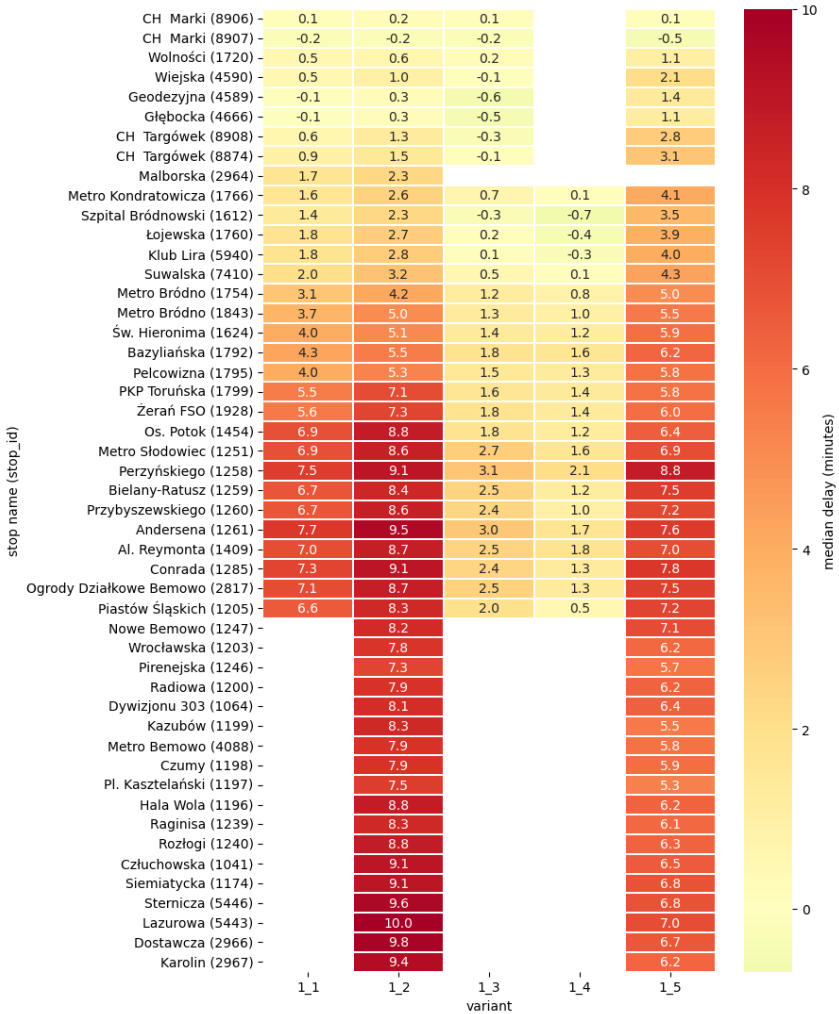

Fig. 7. Median arrivals delay by stop and route variant (direction 1)

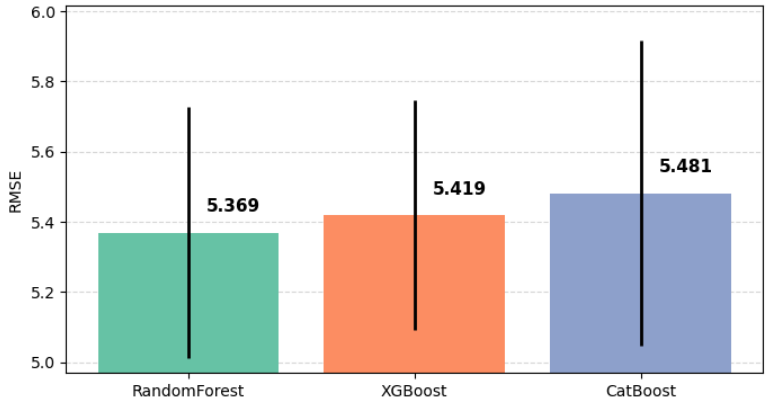

Fig. 8. Comparison of mean RMSE values across Random Forest, XGBoost, and CatBoost models

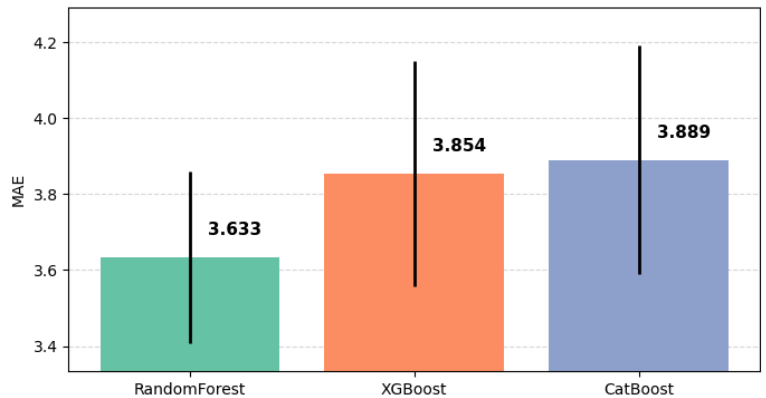

Fig. 9. Comparison of mean MAE values across

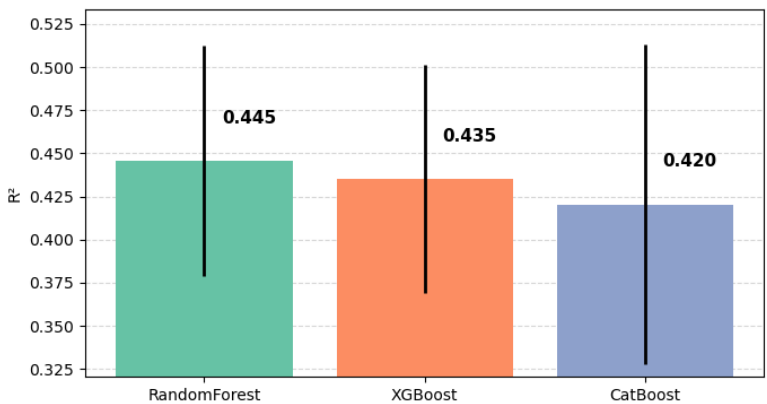

Fig. 10. Comparison of mean R^2 values across Random Forest, XGBoost, and CatBoost models

Table 3 summarises the validation protocol and the best-model settings, together with mean RMSE/MAE/R^2 computed across the outer weekly folds of the nested temporal cross-validation.

4.3. Model interpretation

4.3.1. Feature importance

Figure 11 presents the relative importance of the input variables in the Random Forest model, expressed as the normalized mean decrease in impurity (gain).

The results indicate that the most influential predictors of bus arrival delays are stop sequence, time of day (hour_rounded).

These variables collectively reflect a trip's temporal and positional context, capturing the progressive accumulation of delays along the line and the strong effect of peak-hour traffic.

Among the categorical attributes, weekday and variant_simple_id also contributed noticeably to the model's predictive power, suggesting systematic differences in delay patterns between service variants and operating days.

In contrast, operational identifiers such as vehicle_number, brigade, or stop_id exhibited minor importance, confirming that delay dynamics are primarily determined by the service schedule and spatial-temporal position rather than specific vehicles or drivers.

Figure 12 presents the alternative view of feature importance based on the split frequency (weight), indicating how often individual features were used for node splitting across all trees in the Random Forest ensemble.

The results differ from the gain-based ranking, emphasizing variables that the model most frequently referenced during partitioning rather than those that contributed the most to error reduction

In this perspective, the vehicle_number feature appears most frequently, followed by hour_rounded and brigade.

This pattern suggests that operational and temporal identifiers were commonly encountered in the decision paths of the model, even if their overall predictive contribution was less pronounced in the gain-based analysis.

Table 3. Overview of model configuration, validation, and performance (weekly nested temporal CV; metrics averaged across outer folds

Model	Validation setup	Key tuned hyperparameters (inner loop)	RMSE (min)	MAE (min)	R^2
Random Forest	Nested temporal CV (weekly blocks; no leakage)	n_estimators, max_depth, min_samples_leaf	5.369	3.633	0.445
XGBoost	Nested temporal CV (weekly blocks; no leakage)	n_estimators, max_depth, learning_rate	5.419	3.854	0.435
CatBoost	Nested temporal CV (weekly blocks; no leakage)	depth, learning_rate, l2_leaf_reg	5.481	3.889	0.42

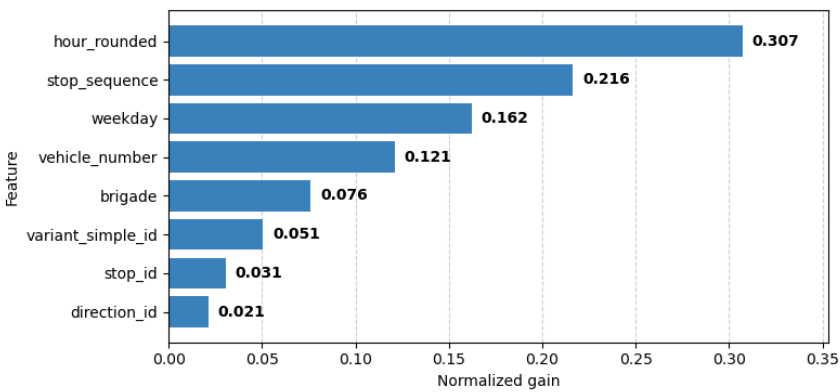

Fig. 11. Feature importance (gain) for the Random Forest model predicting bus arrival delays

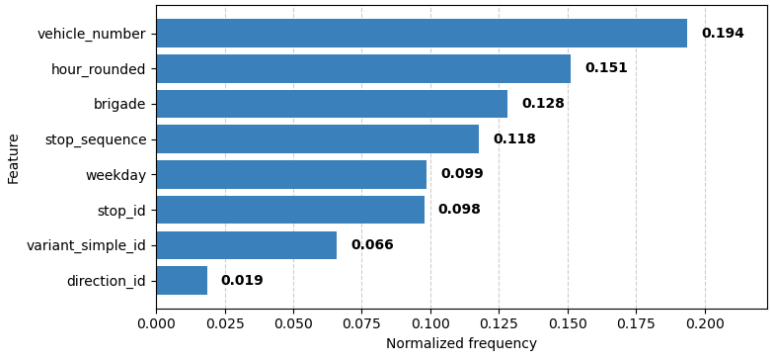

Fig. 12. Feature importance (weight) for the Random Forest model based on split frequency across all trees

Such a difference highlights the complementary nature of the two importance metrics — while gain quantifies the strength of a feature's impact, weight reflects its prevalence in the model's decision structure.

Figure 13 presents the feature importance calculated using the cover metric, which quantifies the share of training samples influenced by splits based on a given variable.

This measure highlights features that affect large portions of the dataset, representing the global scope of their influence rather than the strength or frequency of use.

The results show that hour_rounded, stop_sequence, and weekday were associated with the largest sample coverage, indicating that these attributes govern delay patterns across the majority of trips.

Operational identifiers such as vehicle_number and brigade also reached relatively high cover values, suggesting that they occasionally define broader traffic behaviour groups.

In contrast, variable direction_id had the lowest coverage, implying that their effect was limited to more localized decision regions within the model.

4.3.2. Shap

Figure 14 presents the SHAP summary plot illustrating the marginal contribution of each feature to the predicted bus arrival delay.

The horizontal spread of each distribution reflects the magnitude of impact on the model output, while the color represents the feature value (red – high, blue – low).

The analysis confirms that hour_rounded, stop_sequence, and weekday strongly influenced the predicted delay, consistent with the patterns identified in the gain- and cover-based importance rankings.

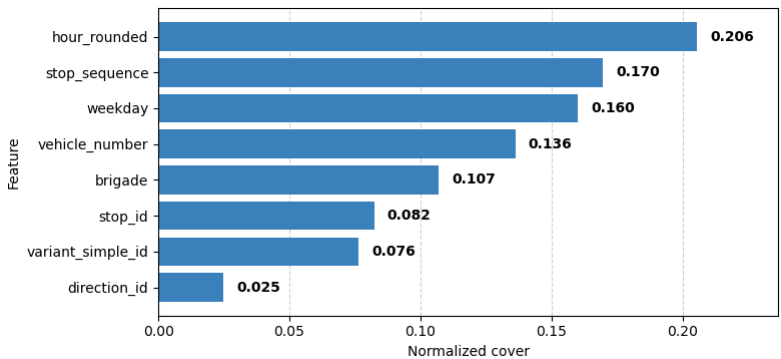

Fig. 13. Feature importance (cover) for the Random Forest model

Higher values of hour_rounded (representing afternoon peak hours) are associated with increased predicted delays, whereas early-morning observations (blue) shift the prediction downward.
Similarly, later stop_sequence positions along the route contribute positively to the delay, indicating progressive accumulation of lateness along the line.
The variable weekday differentiates workdays from weekends, capturing systematic temporal effects, while operational identifiers such as vehicle_number and brigade have a weaker, yet non-negligible, impact.
Overall, the SHAP analysis validates that the temporal and positional characteristics of service operation dominate the predictive structure of the model.

Figure 15 presents the global feature importance ranking in the Random Forest model based on the mean absolute SHAP values (mean(|SHAP|)).
The results indicate that hour_rounded (time of day) and stop_sequence (stop position along the route) exert the strongest overall influence on the predicted bus delays, confirming the dominant role of temporal and cumulative spatial effects.
The weekday variable also shows a substantial contribution, capturing the difference between weekday and weekend operating conditions.
Operational identifiers such as vehicle_number and brigade have a weaker but still noticeable impact, whereas direction_id exhibit marginal influence.
The horizontal axis represents the average absolute SHAP magnitude, which reflects the strength of each feature's contribution to the prediction rather than the direction of its effect.

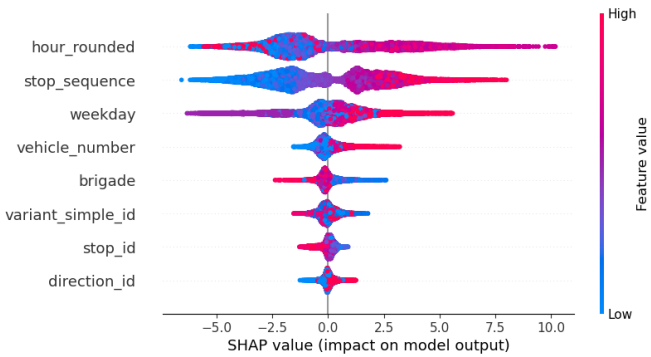

Fig. 14. SHAP summary plot for the Random Forest model showing feature contributions to predicted bus arrival delays (red = higher feature value, blue = lower feature value)

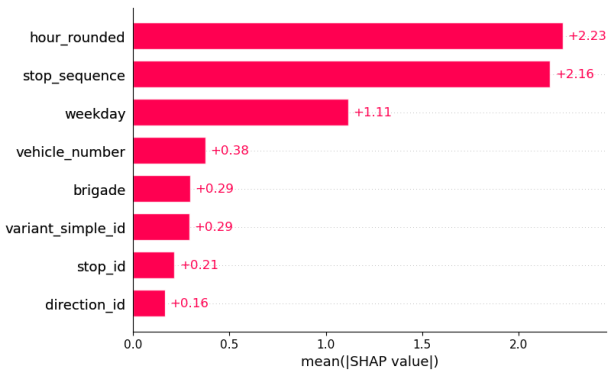

Fig. 15. Global feature importance of the Random Forest model measured by the mean absolute SHAP values (mean(|SHAP|)); higher bars correspond to stronger average influence on the predicted bus delay

5. Discussion

5.1. Interpretation of findings

The exploratory results for Line 112 in Warsaw are consistent with international evidence on temporal and spatial patterns of bus delays, while showing a stronger magnitude of effects. Similar to the findings reported for Xi'an, the hourly delay profile displays distinct morning and afternoon peaks, confirming that congestion and passenger demand are dominant determinants of schedule deviation (Y. Zhang et al., 2022). However, the amplitude of afternoon delays in Warsaw—exceeding twelve min—is higher than the 5–8 minute range typically observed in Asian cities, indicating greater sensitivity to mixed-traffic congestion (Chen et al., 2025). The spatial progression of delays along the route corresponds to the cumulative propagation described in Portland-based analyses, where minor early deviations amplify toward terminal stops (Figliozzi et al., 2021). This pattern is further consistent with spatiotemporal clustering of delays identified in Taipei, confirming that bunching and propagation are systemic rather than city-specific phenomena (Chung & Chiang, 2024). The pronounced weekday–weekend contrast observed in Warsaw mirrors temporal asymmetries reported for British networks, where reduced weekend traffic substantially improves reliability (Fonzone et al., 2015). Finally, the overall consistency of these mechanisms with global reviews of bus bunching supports the view that Warsaw's reliability problems stem from structural rather than random factors (Rezazada et al., 2024).

The modelling results obtained for Line 112—three tree-based ensembles with similar predictive performance (RMSE ≈ 5.4 min, MAE ≈ 3.6 min, R^2 ≈ 0.44)—are consistent with the range of errors reported in empirical studies that use gradient-boosting and random-forest approaches on AVL-derived urban bus data, indicating that tree-based ensembles remain a reliable baseline for short-horizon travel-time and delay prediction when feature sets are moderate in size (Zhu et al., 2022). The dominance of temporal and positional predictors in your models (hour, stop sequence, weekday, variant) mirrors established findings that congestion cycles and route progression explain most of the predictable variance in delays; this matches prior applied work showing time-of-day and stop/segment position as primary drivers of prediction performance (Antypas et al., 2024; Singh & Kumar, 2022). The finding that vehicle identifiers and crew proxies contribute little is meaningful: it suggests that, for the analysed line, delay variability is primarily driven by temporal and route-related factors rather than vehicle- or crew-specific effects noted in some other settings. From an interpretability perspective, the comparison between split-based importance metrics (gain/weight/cover) and SHAP is in line with methodological studies: SHAP provides consistent, local-aware attributions that often re-rank or temper the apparent influence of variables that split frequently or have many categories, while recent work warns that naive split-improvement measures can be biased toward variables with many potential splits (S. Lundberg & Lee, 2017; Zhou & Hooker, 2019). Finally, situating these modelling conclusions alongside empirical studies of bunching and propagation shows complementarity: models explain the predictable portion of delay driven by time/position, whereas the remaining unexplained variance reflects complex dynamics of passenger boarding patterns and local congestion that empirical and theoretical bunching studies identify as harder to predict without richer context data (Figliozzi et al., 2012; Fonzone et al., 2015).

Under the week-ahead forecasting setup, the predictions inform tactical planning rather than day-to-day dispatch. Models reliably highlight persistent problem windows across days, hours and route segments. The results support timetable slack adjustments, headway tuning, crew and vehicle allocation, and the prioritization of infrastructure measures. As a practical rule, planning changes are warranted when the predicted median delay for a weekly window materially exceeds the MAE; values near the MAE are treated as fluctuation and do not justify schedule changes.

The proposed framework can be integrated with operator tools for delay monitoring and network management. Model outputs—aggregated delay indicators and route-segment visualizations—can support prioritization of interventions, timetable adjustments, and identification of recurring bottlenecks. In practice, the workflow can complement existing analytical dashboards and provide regularly updated evidence for planning and operational decisions.

5.2. Limitations and directions for future research

Although the present analysis provides a detailed empirical overview of delay mechanisms on Warsaw's Bus Line 112, several limitations should be acknowledged. First, the spatial and temporal scope of the dataset was restricted to a single route and one month of operations, which limits the generalization of the results to the entire network or to other seasons of the year. Second, the study relied solely on Automatic Vehicle Location (AVL) records, without integration of external variables such as traffic flow, signal control, passenger demand, or weather conditions. These contextual factors are known to influence bus reliability and could refine model accuracy if incorporated. Third, potential inaccuracies in AVL timestamps and positional data may have introduced minor noise into estimating arrival delays. Finally, the descriptive analysis aimed to identify dominant spatiotemporal patterns rather than establish causal relationships among operational factors.

Future research should extend this framework both methodologically and contextually. The proposed approach could be scaled to multiple routes or the entire Warsaw bus network to identify system-wide delay clusters and propagation corridors. Expanding the temporal horizon to several months or a full year would enable assessment of seasonal effects and weather-related variability. Further work should integrate AVL data with complementary datasets capturing traffic intensity, intersection control, or passenger boarding volumes, allowing for more comprehensive explanatory and predictive modelling. From a methodological perspective, dynamic and sequence-based models such as recurrent neural networks or temporal gradient-boosting frameworks could be applied to forecast short-term delay evolution in real time. Comparative studies involving other European cities with higher levels of bus priority infrastructure would also help contextualize Warsaw's results. Ultimately, such analyses could inform the development of decision-support tools for transit authorities, aimed at monitoring delay propagation and optimizing timetable robustness.

6. Conclusions

The analysis of AVL data for the analysed bus line (Line 112) suggests that major delays exhibit a predominantly systemic character, shaped by daily temporal patterns and cumulative delay propagation along the route. The highest mean delays occurred during the afternoon peak hours (15:00–17:00), exceeding 12 min, indicating a strong influence of mixed-traffic congestion on service reliability. Spatial analysis revealed a progressive increase in delays from the initial to terminal stops, confirming the propagation effect widely reported in international literature.

The ensemble models (Random Forest, XGBoost, CatBoost) achieved comparable predictive accuracy (RMSE ≈ 5.4 min; MAE ≈ 3.6 min; $R^2 \approx 0.44$), indicating their suitability within the analysed forecasting setup for short-term bus delay forecasting. The most influential predictors were time of day, stop sequence, and weekday, highlighting the prominent role of temporal–spatial factors over vehicle- or crew-specific effects. Comparison between split-based importance metrics (gain, weight, cover) and SHAP values showed that SHAP provides a more consistent and less biased interpretation of feature influence, especially in datasets with numerous categorical variables.

For the analysed line, the findings suggest that addressing observed punctuality issues may benefit from systemic measures—such as signal priority, dedicated bus lanes, or improved schedule alignment with prevailing traffic conditions—rather than vehicle-level adjustments alone. The proposed analytical framework can be extended to support network-wide delay monitoring and visualization in future applications, thereby assisting decision-making processes in public transport management.

References

1. Aemmer, Z., Ranjbari, A., & MacKenzie, D. (2022). Measurement and classification of transit delays using GTFS-RT data. *Public Transport*, *14*(2), 263–285. https://doi.org/10.1007/s12469-022-00291-7
2. Almeida, A., Brás, S., Sargento, S., & Oliveira, I. (2023). Exploring bus tracking data to characterize urban traffic congestion. *Journal of Urban Mobility*, *4*, 100065. https://doi.org/10.1016/j.urbmob.2023.100065

3. Antypas, E., Spanos, G., Lalas, A., Votis, K., & Tzovaras, D. (2024). *A time-series approach for estimated time of arrival prediction in autonomous vehicles*. *78*, 166–173. https://doi.org/10.1016/j.trpro.2024.02.022
4. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028
5. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
6. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
7. Chen, X., Saidi, S., & Sun, L. (2025). Understanding bus delay patterns under different temporal and weather conditions: A Bayesian Gaussian mixture model. *Transportation Research Part C: Emerging Technologies*, *171*, 105000. https://doi.org/10.1016/j.trc.2025.105000
8. Chung, Y.-S., & Chiang, Y.-C. (2024). Characterizing spatiotemporal patterns of bus bunching frequency on a bus route network: A case study of Taipei city. *Asian Transport Studies*, *10*, 100139. https://doi.org/10.1016/j.eastsj.2024.100139
9. Drabicki, A. A., Islam, M. F., & Szarata, A. (2021). Investigating the Impact of Public Transport Service Disruptions upon Passenger Travel Behaviour—Results from Krakow City. *Energies*, *14*(16), 4889. https://doi.org/10.3390/en14164889
10. Figliozzi, M., Feng, W., & Laferriere, G. (2021). *A Study of Headway Maintenance for Bus Routes: Causes and Effects of "Bus Bunching" in Extensive and Congested Service Areas*. https://rosap.ntl.bts.gov/view/dot/24701/dot_24701_DS1.pdf
11. Figliozzi, M., Feng, W., Laferriere, G., & Feng, W. (2012). *A Study of Headway Maintenance for Bus Routes: Causes and Effects of "Bus Bunching" in Extensive and Congested Service Areas*. Portland State University Library. https://doi.org/10.15760/trec.107
12. Fonzone, A., Schmöcker, J.-D., & Liu, R. (2015). A model of bus bunching under reliability-based passenger arrival patterns. *Transportation Research Part C: Emerging Technologies*, *59*, 164–182. https://doi.org/10.1016/j.trc.2015.05.020
13. Ge, L., Voß, S., & Xie, L. (2022). Robustness and disturbances in public transport. *Public Transport (Heidelberg, Germany)*, *14*(1), 191–261. https://doi.org/10.1007/s12469-022-00301-8
14. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2207.08815
15. Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
16. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55
17. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.
18. Jeong, R., & Rilett, L. R. (2005). Prediction Model of Bus Arrival Time for Real-Time Applications. *Transportation Research Record: Journal of the Transportation Research Board*, *1927*(1), 195–204. https://doi.org/10.1177/0361198105192700123
19. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, *6*(4), 1–21. https://doi.org/10.1145/2382577.2382579
20. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Team, J. D. (2016). Jupyter Notebooks—A publishing format for reproducible computational workflows. *International Conference on Electronic Publishing*.

21. Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1705.07874
22. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). *Explainable AI for Trees: From Local Explanations to Global Understanding* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1905.04610
23. Matseliukh, Y., Lytvyn, V., Hu, Z., & Bublyk, M. (2025). Predictive Modelling and Factor Analysis of Public Transport Delays in Smart City Using Interpretable Machine Learning. *International Journal of Information Technology and Computer Science*, *17*(6), 1–28. https://doi.org/10.5815/ijitcs.2025.06.01
24. McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
25. Park, Y., Mount, J., Liu, L., Xiao, N., & Miller, H. J. (2020). Assessing public transit performance using real-time data: Spatiotemporal patterns of bus operation delays in Columbus, Ohio, USA. *International Journal of Geographical Information Science*, *34*(2), 367–392. https://doi.org/10.1080/13658816.2019.1608997
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. https://doi.org/10.48550/ARXIV.1201.0490
27. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). *CatBoost: Unbiased boosting with categorical features* (Version 5). arXiv. https://doi.org/10.48550/ARXIV.1706.09516
28. Rezazada, M., Nassir, N., Tanin, E., & Ceder, A. (Avi). (2024). Bus bunching: A comprehensive review from demand, supply, and decision-making perspectives. *Transport Reviews*, *44*(4), 766–790. https://doi.org/10.1080/01441647.2024.2313969
29. Rosenblum, J. L., Allen, D. W., Bennett, T. L., Warade, R. K., & Stoughton, C. M. (2015). Method for Assessing Bus Delay in Mixed Traffic to Identify Transit Priority Improvement Locations in Cambridge, Massachusetts. *Transportation Research Record: Journal of the Transportation Research Board*, *2533*(1), 60–67. https://doi.org/10.3141/2533-07
30. Serin, F., Alisan, Y., & Erturkler, M. (2022). Predicting bus travel time using machine learning methods with three-layer architecture. *Measurement*, *198*, 111403. https://doi.org/10.1016/j.measurement.2022.111403
31. Singh, N., & Kumar, K. (2022). A review of bus arrival time prediction using artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, *12*(4), e1457. https://doi.org/10.1002/widm.1457
32. Sun, Y., Spall, J., Wong, W., & Zhao, X. (2025). *Real-time Bus Travel Time Prediction and Reliability Quantification: A Hybrid Markov Model* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2503.05907
33. Toledano, J. S., Monedero, B. D., Flores - Ureba, S., & De Blas, C. S. (2025). The efficiency of urban public transport and its impact on environmental sustainability. *Sustainable Technology and Entrepreneurship*, *4*(2), 100097. https://doi.org/10.1016/j.stae.2025.100097
34. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91
35. Vijaya, A., Bhattarai, S., Angreani, L. S., & Wicaksono, H. (2024). Enhancing Transparency in Public Transportation Delay Predictions with SHAP and LIME. *2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1285–1289. https://doi.org/10.1109/IEEM62345.2024.10857000
36. Warnakulasuriya, A. K., Weerasinghe, C. D. R. M., Wickramarathna, H. K. G. V. L., Ratneswaran, S., & Thayasivam, U. (2024). Explainable Bus Arrival Time Prediction Model with Improved Features Related to Topography and Points of Interest. *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2131–2136. https://doi.org/10.1109/ITSC58415.2024.10920146
37. Warsaw City Council. (n.d.). *Warsaw Open Data Repository* [Data set]. Retrieved 29 October 2025, from https://api.um.warszawa.pl/

38. Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021
39. Yap, M., & Cats, O. (2021). Predicting disruptions and their passenger delay impacts for public transport stops. *Transportation*, *48*(4), 1703–1731. https://doi.org/10.1007/s11116-020-10109-9
40. Zhang, Q., Ma, Z., Ling, Y., Qin, Z., Zhang, P., & Zhao, Z. (2025). Causal Graph Discovery for Urban Bus Operation Delays: A Case Study in Stockholm. *Transportation Research Record: Journal of the Transportation Research Board*, *2679*(5), 256–272. https://doi.org/10.1177/03611981241306754
41. Zhang, Q., Ma, Z., Zhang, P., Ling, Y., & Jenelius, E. (2024). Real-time bus arrival delays analysis using seemingly unrelated regression model. *Transportation*. https://doi.org/10.1007/s11116-024-10507-3
42. Zhang, Y., Xu, H., Lu, Q.-C., & Fan, X. (2022). Travel Time Reliability Analysis Considering Bus Bunching: A Case Study in Xi'an, China. *Sustainability*, *14*(23), 15583. https://doi.org/10.3390/su142315583
43. Zhou, Z., & Hooker, G. (2019). *Unbiased Measurement of Feature Importance in Tree-Based Methods* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1903.05179
44. Zhu, L., Shu, S., & Zou, L. (2022). XGBoost-Based Travel Time Prediction between Bus Stations and Analysis of Influencing Factors. *Wireless Communications and Mobile Computing*, *2022*(1), 3504704. https://doi.org/10.1155/2022/3504704