scientific reports

Check for updates

OPEN

Explainable AI with EDA for V2I path loss prediction

Mongi Ben Ameur[1], Jalel Chebil[1], Mohamed Hadi Habaebi[2]✉, Jamel Bel Hadj Tahar[1], Md Rafiqul Islam[2] & Abdul Manan Sheikh[3]✉

Accurate pathloss (PL) prediction is essential for reliable Vehicle-to-Infrastructure (V2I) communication, particularly in dense urban environments characterized by mobility, multipath effects, and complex street geometries. Traditional empirical models often fail to capture these variations, while black-box machine learning (ML) methods lack transparency, limiting their suitability for safety-critical V2X applications. This paper proposes a fully explainable V2I PL prediction framework that integrates Exploratory Data Analysis (EDA), optimized Kalman filtering, and inherently interpretable ML models, including Explainable Boosting Machines (EBM), Generalized Additive Models (GAM), and Generalized Neural Additive Models (GNAM). The framework is validated using a large-scale dataset of 24 heterogeneous urban scenarios and evaluated through 5-fold cross-validation and multi-seed runs. Results show that interpretable models offer competitive accuracy compared to black-box approaches while providing robust global and local explanations of feature contributions. The study also discusses computational considerations, real-time feasibility, and ethical aspects relevant to practical V2X deployment. The proposed framework demonstrates high potential for transparent and trustworthy PL prediction in future 5G/6G V2I systems.

Keywords Path loss prediction, V2I communications, Explainable AI (ExAI), Channel modeling

Accurate PL prediction is essential for the design and optimization of modern wireless communication systems, particularly for fifth-generation (5G) and V2I networks. Reliable PL estimation ensures robust connectivity, efficient resource allocation, and high quality of service in dense urban environments characterized by mobility, multipath effects, and complex street layouts[1]. Traditional empirical and deterministic models often fail to capture such dynamic and heterogeneous conditions, limiting their applicability in real-world vehicular scenarios.

Machine learning (ML) approaches have recently demonstrated strong potential to improve PL prediction accuracy and adaptability. Deep learning methods, for instance, have been successfully applied to 5G New Radio V2X scenarios, outperforming conventional models[2], while[3] showed ML's effectiveness in dense urban vehicular environments. Yet, most ML models operate as black boxes, offering limited interpretability and reducing their suitability for safety-critical V2X applications. Explainable artificial intelligence (ExAI) techniques address this gap by enabling transparent "glass-box" modeling, providing both global and local insights into feature contributions without compromising predictive performance[4,5].

In this work, we propose a fully explainable V2I PL prediction framework that integrates EDA, optimized Kalman filtering, and inherently interpretable ML models, including EBM, GAM, and GNAM. Evaluated across 24 heterogeneous urban scenarios using 5-fold cross-validation and multi-seed experiments, the framework demonstrates that interpretable models achieve competitive accuracy while providing actionable insights into feature contributions, computational feasibility, and ethical considerations, highlighting their potential for transparent and trustworthy V2I deployment in future 5G/6G networks.

Related work

The introduction of machine learning (ML) into path loss (PL) prediction has significantly improved predictive performance but, in doing so, has introduced the ultimate black box problem. In response, more and more studies have been focused on explainable AI (ExAI) in this area. Existing studies can typically be categorized into two paradigms: post hoc explanations of unintelligible models and inherently interpretable models that are transparent by design. Most of the literature is under the purview of the assumption that ML models possess an inherent opacity, and thus interpretability needs to be achieved through external measures. A typical approach involves perturbation-based explanations, where feature values are systematically altered and the resulting

[1]NOCCS Laboratory, University of Sousse, Sousse, Tunisia. [2]Department of ECE, International Islamic University Malaysia, Kuala Lumpur, Malaysia. [3]Department of EECS, College of Engineering, A'Sharqiyah University, 400, Ibra, Oman. ✉email: habaebi@iium.edu.my; abdul.manan@asu.edu.om

impact on the model's output is examined. A well-known example of such a practice is LIME (Local Interpretable Model-agnostic Explanations)[6], which learns a simple surrogate model in the vicinity of an individual prediction to facilitate a localized explanation. Research such as[7] has used these techniques to explain feature importance in predicting path loss results. Another major thread of research in this paradigm has its roots in Shapley value theory. SHAP (SHapley Additive exPlanations)[8] directly assigns contributions to the outputs of black-box models without resorting to surrogate models and thereby provides stronger theoretical guarantees. The method has found extensive use in PL research, including in the work of[9–11]. Although such frameworks provide useful insights, they still function as approximations of the model's reasoning and can be vulnerable to perturbations, sampling methods, or feature dependencies[12]. A more compact but conceptually advanced literature tackles the black-box issue head-on by using models that are inherently interpretable, or often simply called "glass box" models. Traditional approaches commonly involve linear regression and decision trees, while more recent developments, like Explainable Boosting Machines (EBMs)[13], attain state-of-the-art accuracy levels while being fully transparent. Redondi et al.[14] offer a particularly relevant example by specifically exploring interpretable regression models in the context of millimeter-wave and V2I communications. Their work highlights the importance of explanation fidelity in applications where safety is critical, as the decisions of communication systems have immediate repercussions on reliability and user safety. Following these findings, the current work takes the second paradigm forward by bringing intrinsically interpretable machine learning models to V2I path loss prediction. Unlike previous work that has predominantly resorted to post-hoc methods, our method ensures interpretability is inherent in the model itself, so the explanation serves as a true representation of its internal reasoning. By showing that state-of-the-art predictive accuracy is attainable in the V2I scenario without compromises in transparency, this paper sets a new course for explainable path loss prediction. In the process, it shows that interpretability and performance are not opposing goals but can be achieved concurrently, a point that is especially vital for vehicular communication systems.

In spite of being inherently interpretable and demonstrating a promise for V2I path loss prediction, there is a lack of studies addressing EBM, GAM, and GNAM ML models' performance under highly dynamic vehicular conditions, their scalability across diverse urban scenarios, their real-time computational feasibility, and their safety-critical implications for practical deployment in 5G/6G vehicular networks.

Dataset description & preprocessing

In this study, an open-source dataset from a vehicle-to infrastructure (V2I) measurement campaign carried out in Bologna, Italy, at 5.9 GHz is employed in this study[15]. The transmitter antennas (Tx) were installed at heights of 6.5 m and 10.5 m, while the receiver antenna (Rx) was mounted on a vehicle rooftop at approximately 2.5 m. Since all measurements were carried out in urban environments, the height of the Rx was assumed constant throughout the campaign. The dataset provides large-scale channel information in terms of the received signal strength indicator (RSSI). Path loss is derived from the measured RSSI[16], with corrections applied for antenna patterns and cable losses, and is calculated as

$$PL = P_{TX} - P_{RX} + L_{cable} - G_{TX} - G_{RX} \tag{1}$$

where PL is the path loss in dB, P_{TX} is the transmission power in dBm, and P_{RX} is the received power in dBm, which in this context is the measured RSSI value, L_{cable} represents the cable loss in dB, and G_{TX} and G_{RX} are the Tx and Rx antenna gains in dBi, respectively.

Preprocessing is an essential machine learning pipeline step, which converts raw data into a structured, consistent format for model training. Effective preprocessing significantly improves predictive performance, reduces bias, and enhances model generalization, whereas poor preprocessing can severely degrade results. In this work, several preprocessing methods were used, with each methodological decision being guided by exploratory data analysis (EDA) and visual inspection. The complete preprocessing workflow is shown in Fig. 1. It displays the sequential steps taken on the data, from its raw form to normalization, standardization, outlier removal, and filtering, and ending with feature ranking and selection. The dataset is classified into seven case studies capturing LOS segments, NLOS from buildings or vegetation, bridges and elevation changes, roundabouts, traffic density variations, and dynamic heavy-vehicle blockages, enabling models to learn from heterogeneous propagation conditions and generalize effectively across realistic 5G/6G V2I environments.

Normalization and standardization

Feature scaling is a critical preprocessing step in machine learning based path loss prediction. Without proper scaling, features with larger numerical ranges can dominate the learning process, leading to biased predictions. In this study, two standard techniques were applied:

Normalization (min–max scaling)

$$\acute{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Where x represents an individual feature in the dataset. $x_{min}$and x_{max} are the minimum and maximum values of that feature across the dataset. x' is the normalized feature, which now lies between 0 and 1. A value of 0 corresponds to the minimum observed value, and 1 corresponds to the maximum. All intermediate values are scaled proportionally.

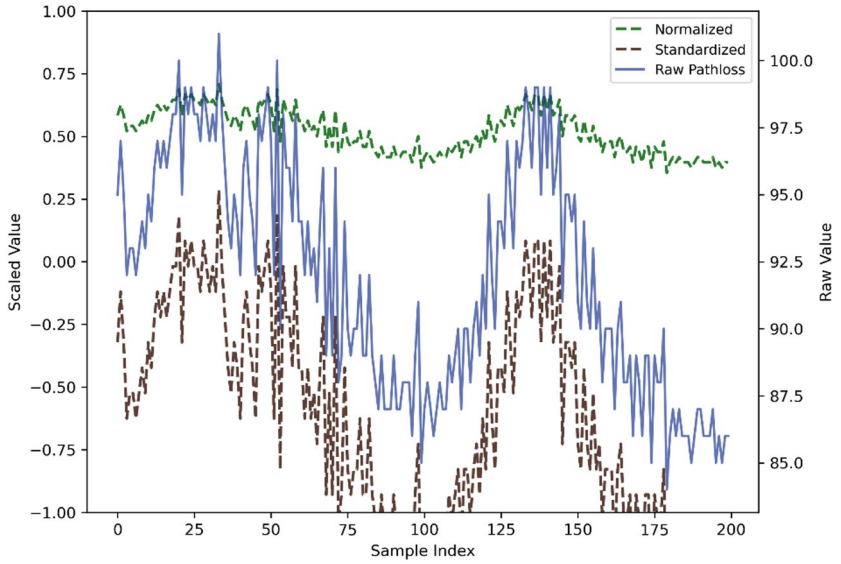

Fig. 1. Effect of normalization and standardization on raw measurements.

Standardization (Z-score scaling)

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

where μ and σ are the mean and standard deviation of the feature, respectively. z is the standardized feature, which represents the number of standard deviations the original value x is from the mean. Positive values indicate x is above the mean, negative values indicate x is below the mean, and values close to 0 indicate proximity to the average. The effect of feature scaling is illustrated in Fig. 1, where normalized and standardized curves transform raw measurements into consistent ranges, allowing fair comparison across different features[16,17]. The data preprocessing workflow applied in this study proceeds as follows: first, raw data were normalized to rescale all features to a uniform range, followed by standardization to center the features around zero mean and adjust them to unit variance. After scaling, outlier removal was performed to eliminate extreme values that could potentially bias the models. Finally, feature ranking and selection were carried out using a baseline machine learning model to identify the most relevant input features for accurate path loss prediction.

Outlier removal

Outliers, which may arise from measurement errors, transmission faults, or rare extreme conditions, can distort feature distributions and bias machine learning models. To mitigate these effects, the interquartile range (IQR) method was applied[18]:

$$IQR = Q_3 - Q_1 \tag{4}$$

$$outliers\ if\ x < Q_1 - 1.5 \times\ IQR \quad or\ x > Q_3 + 1.5 \times\ IQR \tag{5}$$

where Q_1 and Q_3 represent the first and third quartiles, respectively.

Removing outliers is crucial to prevent extreme values from biasing parameter estimation or degrading model performance. In this study, detected outliers were excluded prior to further analysis, enhancing the robustness and accuracy of subsequent modeling. Figure 2 shows the pathloss distribution across V2I scenarios, with outliers clearly highlighted beyond the whiskers of the boxplots.

Data filtering

Data filtering is a crucial preprocessing step in V2I pathloss analysis, as measurements are often contaminated by noise, multipath effects, and sensor errors. Proper filtering enhances signal quality, reduces the impact of outliers, and preserves the underlying propagation dynamics, ensuring more accurate modeling and machine learning predictions. In this work, a classical Kalman filter is applied, followed by a systematic optimization of its parameters to maximize predictive performance.

Kalman filtering for noise reduction

To reduce the impact of noise on V2I pathloss measurements, a classical discrete-time Kalman filter was applied.

The system is modeled as:

$$x_k = Ax_{K-1} + w_K \tag{6}$$

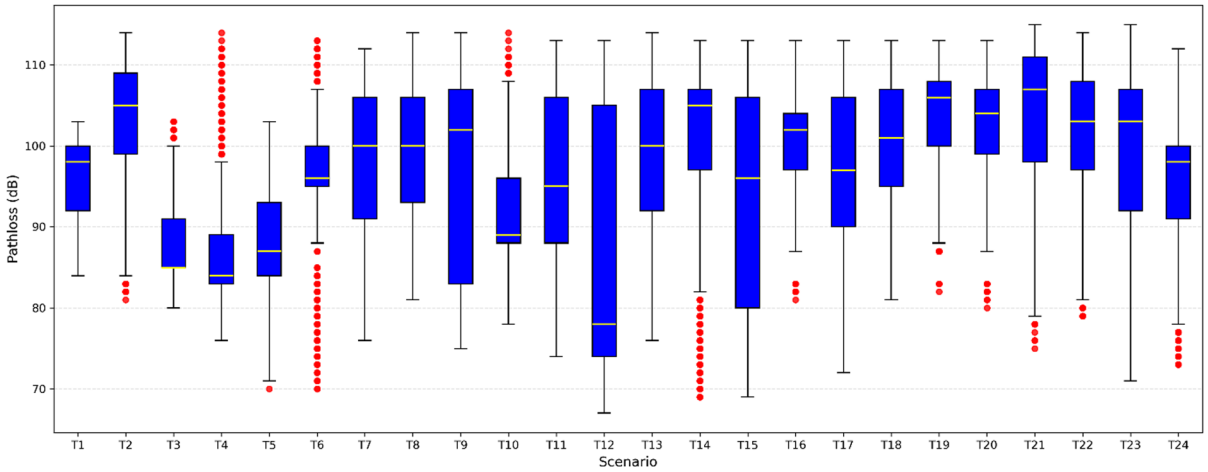

Fig. 2. Pathloss distribution per V2I scenario with outliers highlighted.

$$z_k = Hx_K + v_K \tag{7}$$

where x_k is the true Pathloss state, z_k is the noisy measurement, A is the state transition matrix, H is the observation matrix, and w_K~N(0,Q), v_K~N(0,R), are the process and measurement noises.

The Kalman filter equations are:

$$\widehat{x}_{k|k-1} = A\widehat{x}_{k-1|k-1} \tag{8}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q \tag{9}$$

$$K_k = P_{k|k-1}H^T(HP_{k|k-1}H^T + R)^{-1} \tag{10}$$

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + K_k(z_k - H\widehat{x}_{k|k-1}) \tag{11}$$

$$P_{k|k} = (I - K_kH)P_{k|k-1} \tag{12}$$

where:

- $\widehat{x}_{k|k-1}$: predicted state estimate,
- $\widehat{x}_{k|k}$: filtered (updated) state estimate,
- P : estimation error covariance,
- K_k : Kalman gain,
- Q, R: process and measurement noise covariances.

This formulation ensures efficient smoothing of Pathloss measurements while preserving propagation dynamics linked to distance and environment.

In V2I datasets, noise and multipath effects introduce outliers that can bias statistical modeling. To mitigate these distortions, the measurements were first normalized and standardized, and outliers were removed using a rolling mean combined with a standard deviation threshold. The resulting clean signal was then filtered using the Kalman algorithm described above. The effect of each preprocessing step is illustrated in Fig. 3, which summarizes the full pipeline (Raw → Normalized → Standardized → Outliers Removed → Kalman Filtering).

Kalman parameter optimization for enhanced Pathloss filtering

The process and measurement noise covariances, Q and R, critically affect Kalman filter performance. A grid search was conducted over R ∈ [0.1, 1.0] and Q ∈ [0.001, 0.01], evaluating each (Q, R) pair using the R^2 score of a downstream pathloss prediction model. The optimal parameters (Q*, R*) were selected to maximize predictive accuracy. The 3D R^2 surface, showing the optimal region, is presented in Fig. 4, demonstrating that proper tuning of Q and R effectively suppresses noise while preserving the key V2I propagation characteristics.

Feature ranking and selection

The correlation heat maps (Figs.) further support the selection decisions of characteristics. Feature importance was calculated as the contribution to error reduction on average across models[18,19]:

$$I_j = \frac{1}{M}\sum {}^{M}_{m=1}\Delta\, E_j\,(m) \tag{13}$$

Where:

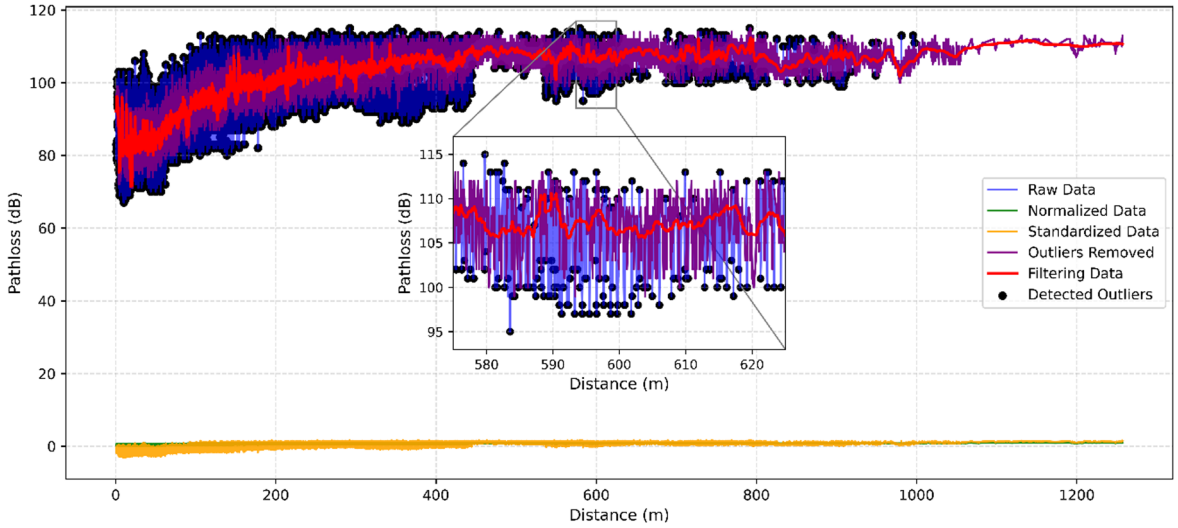

Fig. 3. Pathloss processing steps in V2I measurements.

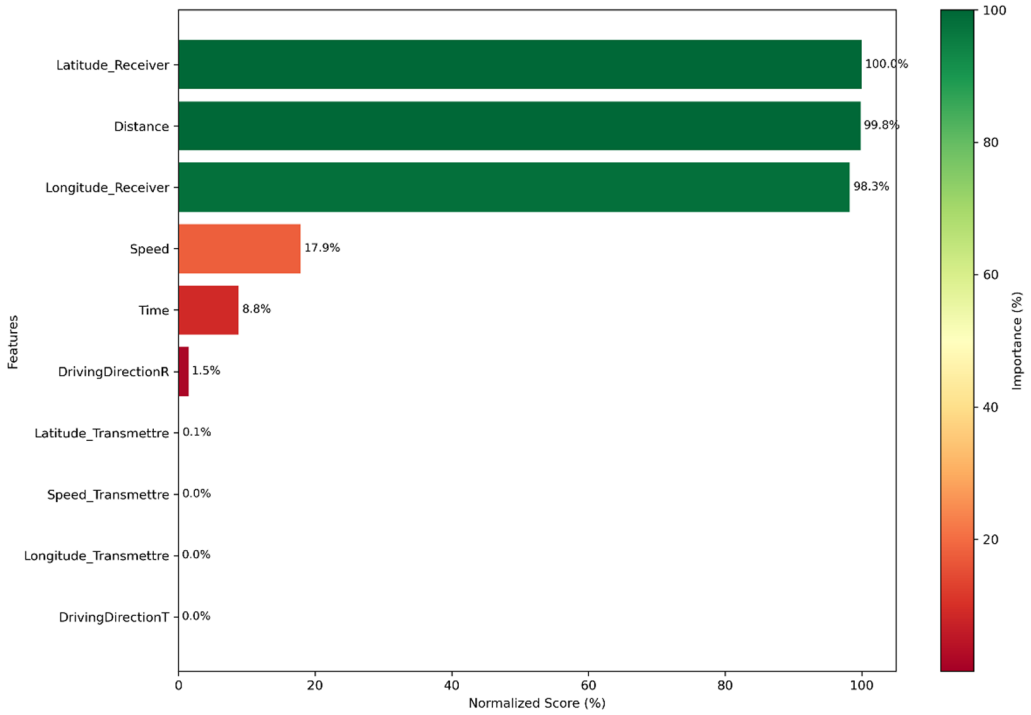

Fig. 4. Feature ranking based on average contribution to error reduction.

- $\Delta E_j\,(m)$ represents the error reduction due to feature.
- *xj* in model *m*,
- *M* is the total number of models.

The most influential predictors are shown in Fig. 4. As illustrated, Distance became the most influential feature with a normalized importance value of 100%, closely followed by Latitude Receiver at 96.3% and Longitude Receiver at 95.8%. These three spatial features clearly capture the geometric dependence inherent in the V2I path loss, thus explaining a large part of the model's predictive ability. In contrast, Speed and Time show moderate contributions, with values of 25.3% and 7%, respectively, pointing to their secondary but not negligible role in propagation. Driving directions and transmitter-side coordinates were found to have minimal or no effect on the result, which justifies an explanation for their omission in simplified models. EDA supports these observations by investigating the one-to-one relationships between the features and the path loss values[20]. The correlation heatmap shown in Fig. 5 reveals a strong positive correlation between path loss and distance ($r = 0.79$), validating

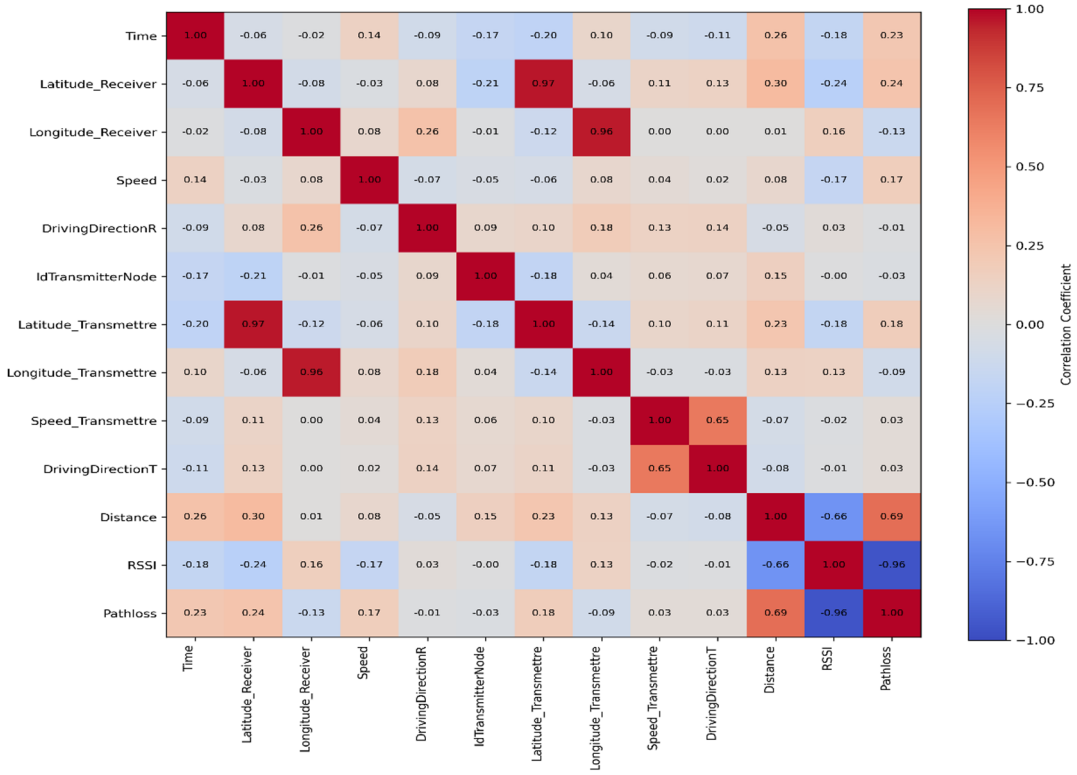

Fig. 5. Correlation heatmap of raw features.

the physical propagation principle that signal attenuation is directly related to distance. By contrast, Latitude Receiver and Longitude Receiver exhibit comparatively weak direct correlations with path loss (0.14 and-0.17, respectively), but their high mutual correlation with distance indicates their indirect contribution to path loss prediction. Thus, the integration of feature importance scores and correlation analysis allows for model interpretability as well as feature selection. This not only improves predictability but also provides insight into V2I communication behavior under realistic environments. EDA has the extra advantage of correlating the features with the required output. In our case, the characteristics were individually correlated with the path loss values. This resulted in producing scores to allow for feature ranking and, hence, selection of the most alternative features. Further, this correlation allows us to explain the machine learning algorithms' behavior in the context of modeling V2I communication path loss. Hence, we can predict and understand the interaction between the different features and the accuracy of the predicted path loss.

Proposed framework and system model

This section outlines the comprehensive framework designed to predict path loss in V2I communications. The framework integrates statistical standards, ML techniques, and ExAI approaches that are guided by EDA and reinforced by advanced data preprocessing techniques. Figure 6 illustrates the overall flow of the suggested system that highlights the interplays of data preparations, model types, explainability factors, and evaluation metrics.

The proposed framework begins from raw V2I measurements acquired in urban vehicular networks. This is subsequently converted into a well-organized, high-quality format by means of EDA driven preprocessing and is then routed into three complementary streams of modeling:

- Statistical Models – traditional benchmarks for propagation.
- Black-Box ML Models – extremely accurate but black-box ensembles.
- Glass-Box ML Models (ExAI) – interpretable and precise techniques for reliable deployment.

Methodology for V2I pathloss modeling

This section outlines the methodology used to develop robust V2I pathloss models. The approach includes dataset partitioning, cross-validation, test-set evaluation, and systematic hyperparameter optimization to ensure accurate and generalizable predictions across varying V2I conditions.

Dataset split

For each V2I scenario, the dataset was partitioned into a training set (80%) and a test set (20%), following standard machine learning practices. This separation allows models to be trained on the majority of the data while reserving an independent portion for testing. Evaluating the models on unseen data provides a realistic

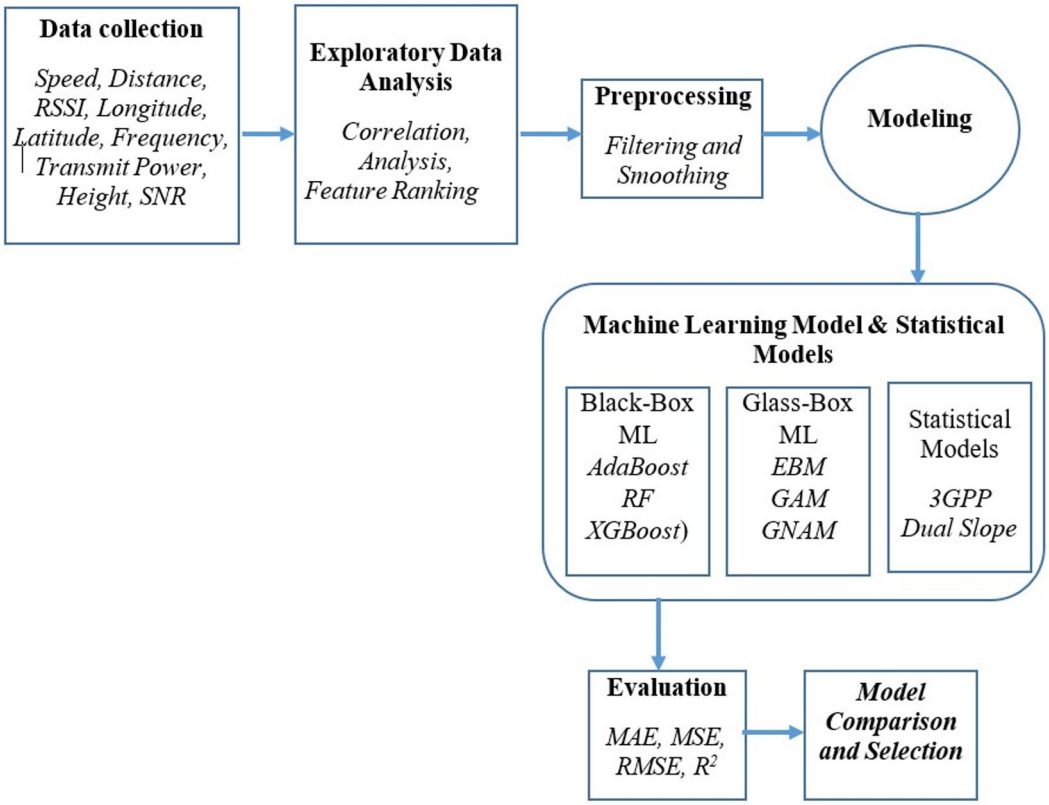

Fig. 6. Framework for EDA and ExAI pathloss prediction for V2I channels.

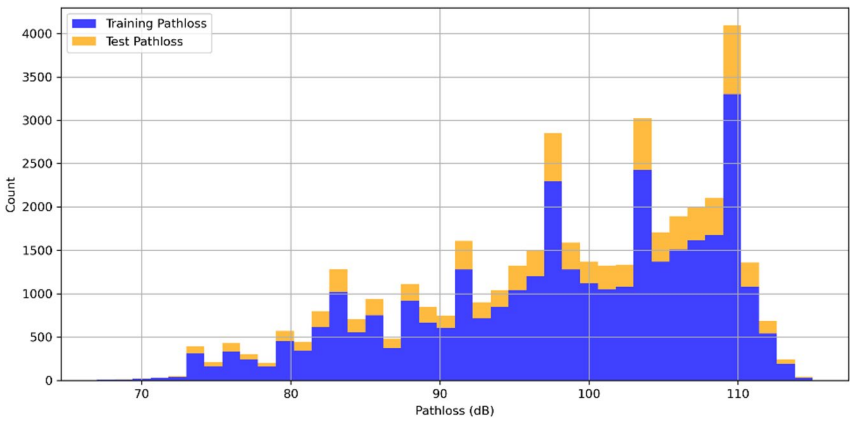

Fig. 7. Distribution of training and test sets along pathloss.

estimate of their generalization capability, which is crucial for accurately predicting pathloss across varying distances in real-world conditions. This distribution is illustrated in Fig. 7.

5-Fold cross-validation

To ensure model robustness and limit overfitting, 5-fold cross-validation was applied. The dataset was split into five folds, with the model trained on four folds and validated on the remaining fold in each iteration. Performance metrics were averaged across all folds to provide a reliable estimate of predictive accuracy and reduce bias. Figure 8 illustrates this 5-fold cross-validation procedure.

Statistical models

The following subsections present two representative cases: the 3GPP model for cellular systems and the dual slope model with distance-dependent exponents.

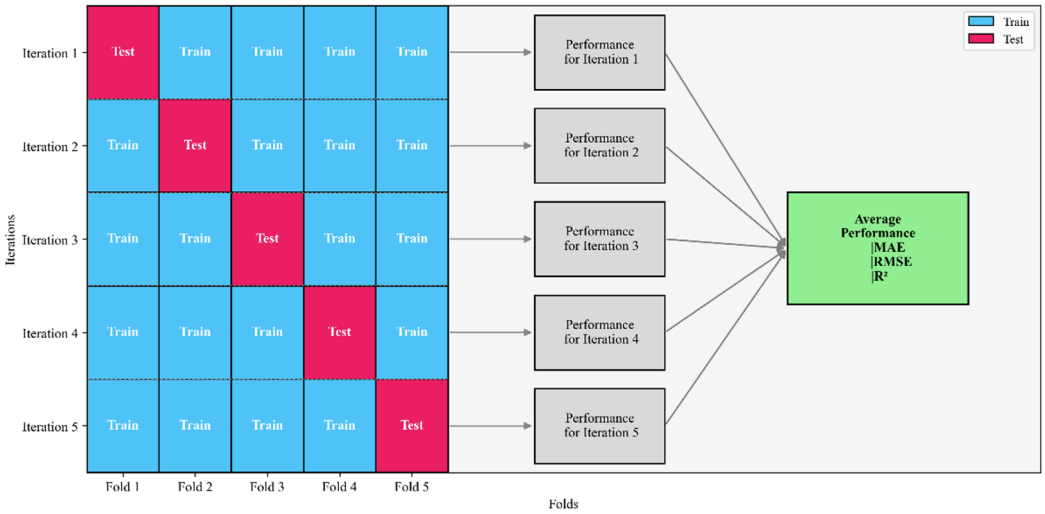

Fig. 8. Schematic of 5-fold cross-validation.

3GPP model

In the technical report 3GPP TR 38.901[21], path loss models are presented for various environments over carrier frequencies ranging from 0.5 GHz to 100 GHz. For outdoor users in the urban macrocell (UMa) scenario, where the transmitter–receiver separation is smaller than the breakpoint distance, the average path loss for LOS and NLOS conditions is specified as:

$$PL_{LoS} = 28 + 22log_{10}\left(d_{3D}\right) + 20log_{10}\left(f_c\right) + \sigma_{\ LoS}^{3GPP} \tag{14}$$

$$PL_{NLoS} = max(PL_{LoS},\ PL\prime_{\ NLoS}) \tag{15}$$

With

$$PL\prime_{\ NLoS} = 13.54 + 39.09log_{10}\left(d_{3D}\right) + 20log_{10}\left(f_c\right) - 0.6\left(h_{UT} - 1.5\right) + \sigma_{\ NLoS}^{3GPP} \tag{16}$$

Here, PL_{LoS} and PL_{NLoS} denote the mean path loss in LOS and NLOS conditions (dB); d3D is the three-dimensional Tx–Rx separation distance (m); fc is the carrier frequency (GHz); h_{UT} is the user terminal height (m); and $\sigma_{\ LoS}^{3GPP}$, $\sigma_{\ NLoS}^{3GPP}$ represent the shadow fading standard deviations (dB) for LOS and NLOS conditions, respectively.

where PL_{LoS} is the line-of-sight probability, as detailed in "Table 7.4.2-1.2" of "TR 38.901"[21].

Dual slope path loss (benchmark) model

The dual slope model is an empirical approach that accounts for distinct propagation characteristics at different distances. It is expressed as follows:

Equation 2: Dual Slope Path Loss Model

$$PL_d = \begin{cases} PL_0 + n_1log_{10}\left(d/d_0\right), & d < d_{break} \\ PL_0 + n_1log_{10}\left(d_{break}/d_0\right) + n_2log_{10}\left(d/d_{break}\right), & d \geq d_{break} \end{cases} \tag{17}$$

where:

PL_d (dB) represents the predicted path loss at distance d,
PL_0 (dB) denotes the reference path loss at d_0,
d_{break} (m) is the breakpoint distance,
n_1 and n_2 are the path loss exponents before and after d_{break}.
The breakpoint distance d_{break} (m) is given by:

$$d_{break} = \frac{4h_{TX}\ h_{RX} - \frac{\lambda}{2}}{4\lambda} \tag{18}$$

Where:

- h_{TX} (m) and h_{RX} (m) denote the transmitter and receiver antenna heights, respectively.
- λ (m)is the wavelength.

The parameters PL_0, d_{break}, n_1, and n_2 were individually estimated for each case study using a least-squares curve-fitting approach, minimizing the error between the model's predictions and the measured path loss data.

The effectiveness of the dual slope model in different environments has been demonstrated in several studies[22–24].

Black-Box machine learning models

Machine learning (ML) models have been extensively applied to predictive tasks. Ensemble methods such as Random Forest (RF) and boosting algorithms (AdaBoost, XGBoost) are dominant due to their accuracy. However, their opaque nature requires post-hoc explanation methods like SHAP.

Extreme gradient boosting (XGBoost)

XGBoost is a scalable and efficient implementation of gradient-boosted decision trees, developed to maximize predictive accuracy while controlling model complexity. Its learning objective is expressed as:

$$L(\phi) = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i) + \sum_{K=1}^{K} \Omega(f_k) \tag{19}$$

where the regularization term is:

$$\Omega\ (f_k) = \gamma\, T\ +\ \left(\frac{1}{2}\right) \lambda\ \|w\|^2 \tag{20}$$

with T represents the number of leaves, w the leaf weights, and γ and λ act as regularization parameters. The model is optimized through a second-order Taylor expansion of the loss function, which leverages both the gradient (g_i) and the Hessian (h_i) to ensure faster convergence and improved numerical stability compared to first-order approaches. XGBoost leverages features such as vehicle speed, distance, heading, antenna height, and transmission power to accurately predict path loss in V2I channels, providing a highly robust model for complex urban propagation environments.

Random forest (RF)

Random Forest (RF) is a learning ensemble approach building many decision trees with bootstrap data sampling and random feature subsets for predicting V2I loss propagation, as depicted in Fig. 9. Individual trees within the forest are trained on various data subsets and feature sets with minimized inter-tree correlation and increased generalizability of the model. The overall prediction is computed by taking the average outputs of individual trees, minimizing variance and overfitting. Here, we implemented 200 trees with a depth of 15 to extract intricate nonlinear patterns found in propagation data without any feature transformation. The randomness inherent to the algorithm resists data noise and outliers typically found in real-world V2I measurements. The important parameters, such as tree numbers and depth, strike a balance between accurate prediction and efficient computation to achieve productive deployment.

Adaptive boosting (AdaBoost)

AdaBoost is a type of ensemble learning algorithm that builds a powerful predictive model by sequentially aggregating numerous weak learners, which are mostly shallow decision trees. Its usage to predict Vehicle-to-Infrastructure (V2I) path loss follows an iterative and adaptive process. A weak learner is initially trained on the

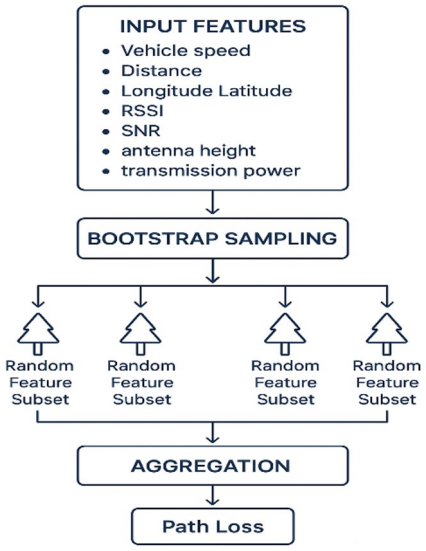

Fig. 9. Random Forest architecture for V2I path loss prediction.

available dataset to provide a baseline model. In each subsequent iteration, the algorithm increases the weights of instances that were misclassified by the previous learners, thereby forcing the next learner to focus on the more difficult cases. This adaptive process is mathematically formalized, and the contribution of each weak learner to the final ensemble is determined by its weight.

$$\alpha_t = \eta \cdot log((1 - \epsilon_t) / \epsilon_t) \tag{21}$$

where η is the learning rate and ε_t is the prediction error of the t^{th} weak learner. The final ensemble prediction is given by:

$$F(x) = \sum_{t=1}^{T} \propto h_t(x) \tag{22}$$

where $h_t(x)$ denotes the prediction of the t^{th} learner.

In the context of V2I path loss estimation, AdaBoost is particularly effective at capturing complex, nonlinear propagation phenomena. By iteratively emphasizing hard-to-predict instances, it identifies and models localized signal variations and intricate spatial dependencies that often challenge single-model approaches.

ExAI for path loss prediction

Accurate estimation of path loss is central to wireless network planning, particularly for complex cases such as smart campuses, vehicular roads, and urban microcells. We begin our study with an EDA to investigate the spatial behaviors and signal patterns in the measurement dataset. A Kalman filter is applied as a preprocessing step to remove noise and regularize the signal. Second, we compare certain black-box machine learning algorithms (XGBoost, Random Forest, and AdaBoost) with well-established statistical models, the 3GPP empirical model and the Dual Slope model. While these black-box models offer superior predictive performance, the lack of interpretability reduces transparency and trust in real-world applications. In an attempt to go beyond this limitation, we explore intrinsically interpretable approaches in the ExAI framework. The following subsections present three glass-box models (Generalized Additive Models (GAM), Generalized Neural Additive Models (GNAM), and Explainable Boosting Machines (EBM)), each offering transparent mechanisms for modeling path loss and enabling explicit inspection of feature contributions and interactions.

Generalized additive models (GAM)

The GAM is a semi-parametric extension of the Generalized Linear Model (GLM) where non-linear relationships between the target variable and input features can be freely modelled without losing interpretability. GAM can be defined more formally as the expectation value of a response variable equaling a sum over a collection of smooths operating on each input feature:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{n} f_j(x_j) + \epsilon \tag{23}$$

where $g(.)$ is the link function, β_0 is the intercept, $f_j(x_j)$ denotes a smooth function (typically a spline) for feature x_j, and $\varepsilon \sim \aleph(0,\sigma^2)$ is the residual error and it follows a normal distribution with mean zero and variance σ^2. Each function f_j is estimated independently, allowing the model to capture complex feature-specific effects without assuming linearity. In this research, GAM was executed by employing the library pyGAM with standard hyperparameters, and the training was done by reducing the residual error as a regularizer was used to limit the smoothness in each spline. GAM's additive structure guarantees that each feature can be independently interpreted by visualization as its contribution towards prediction; however, GAM by its nature doesn't model feature interactions natively, thus potentially reducing its expressiveness in the presence of strong feature dependencies. Furthermore, its performance can suffer when moving towards a highly non-linear or a high-dimensional space if spline parameters are not optimally set.

Generalized neural additive models (GNAM)

The GNAM expands on the GAM by replacing each feature-specific spline function $f(x_j)$with an assigned multilayer perceptron (MLP). In this definition, the model can yet further enhance its manifold expressibility without losing its structure of being additive so that interpretability retains its place. Technically, GNAM can be expressed as:

$$g(E[y]) = \beta_0 + \sum_{j=1}^{n} f_j^{MLP}(x_j) + \epsilon \tag{24}$$

where f_j^{MLP} denotes the output of a neural network trained specifically on feature x_j, and $\varepsilon \sim \aleph(0,\sigma^2)$ is the residual error. Each MLP typically consists of one hidden layer with non-linear activation, enabling the model to learn complex feature-specific patterns. In this study, the GNAM was implemented using the NAMpy framework to predict V2I path loss across multiple propagation scenarios.

Explainable boosting machine (EBM)

The EBM is an advanced glass-box model that combines the interpretability of additive models with the predictive power of gradient-boosted decision treess. EBM models each feature separately with a shallow-encoded chain of decision trees learned from residuals and optionally includes pairwise interaction terms to capture feature dependencies. The model is mathematically formulated as:

$$g\left(E\left[y\right]\right) = \beta_0 + \sum_{j=1}^{n} f_j^{DT}\left(x_j\right) + \sum_{i<j} f_{ij}^{DT}\left(x_i,\ x_j\right) + \epsilon \tag{25}$$

where $f_j^{DT}\left(x_j\right)$represents the contribution of feature x_j modeled via boosted trees, and $f_j^{DT}\left(x_i, x_j\right)$ captures the interaction between features x_i and x_j. Training proceeds in a round-robin fashion, fitting one feature at a time to the residuals of the previous iteration, which mitigates collinearity and ensures stable learning. Interaction terms are selected using the Feature Association Strength Test algorithm, which ranks feature pairs based on their residual impact. In this study, the EBM was implemented using a fixed random seed to ensure reproducibility and consistent results across multiple runs. The model was trained on standardized V2I features, including receiver coordinates, link distance, and normalized path loss, allowing it to learn both linear and non-linear feature effects.

Hyperparameter configuration

Hyperparameters were optimized to balance performance and complexity as shown in Table 1.

Performance measures

The performance of the proposed algorithms was evaluated using three standard metrics: coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). R^2 assesses the goodness of fit by quantifying the linear correlation between predicted and measured values ($0 \le R2 \le 1$), with higher values indicating better prediction accuracy. RMSE represents the standard deviation of prediction errors, while MAE reflects their average absolute magnitude. Lower RMSE and MAE values denote superior predictive performance. The three metrics are defined as follows:

Mean Absolute Error (MAE) Represents the average absolute differences between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{26}$$

Root Mean Squared Error (RMSE) Provides the standard deviation of residuals, indicating model precision.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \tag{27}$$

R-squared (R^2) Measures the proportion of variance explained by the model, indicating overall goodness of fit.

$$R2\ Score = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{28}$$

Where $\bar{y}$ is the mean of the observed path loss values. Lower MAE and RMSE values and higher R^2 indicate better model accuracy and predictive capability.

Results and discussion

This section describes an all-encompassing framework for V2I path loss prediction by integrating statistical models with black-box and glass-box machine learning frameworks. The framework determines EDA to have a central role in uncovering fundamental patterns and anomalies, uses glass-box models as explainable baseline metrics to evaluate performance and tuning purposes, and incorporates filtering methods, to dampen fluctuations in input data. Compared with other propagation environments such as vegetation-obstructed environments and NLOS environments, both environments represent LOS environments. Balancing predictive accuracy with methodological rigor to generate valid and generalizable schemes for predicting path loss is highlighted as presented. Under constant LOS conditions, both black-box models (RF, XGB) and glass-box models (EBM, GAM, GNAM) regularly performed better than statistical baselines (3GPP, DS), obtaining small error and strong R^2 metrics even for unfiltered data. This indicates the natural robustness of ML schemes to measurement noise in relatively deterministic channels. Usage of Kalman filtering also improved performance, mostly for glass-box models. Compared to NLOS conditions, these had more variability due to obstructions together with multipaths. Despite this increased complexity, all ML models stayed ahead of statistical models, consistent with their general robustness across differing propagation conditions.

Model	Configuration
RF	200 trees, max depth = 15
AdaBoost/XGBoost	200 estimators, learning rate = 0.1
XGBoost	max depth = 6 to avoid overfitting
GAM	settings (Spline Type = 0.4, Basis Functions = 25 and Spline Order = 3)
GNAM	20, 25, 15, 10 splines per feature
EBM	Varied seed for reproducibility

Table 1. Models' hyperparameters configuration.

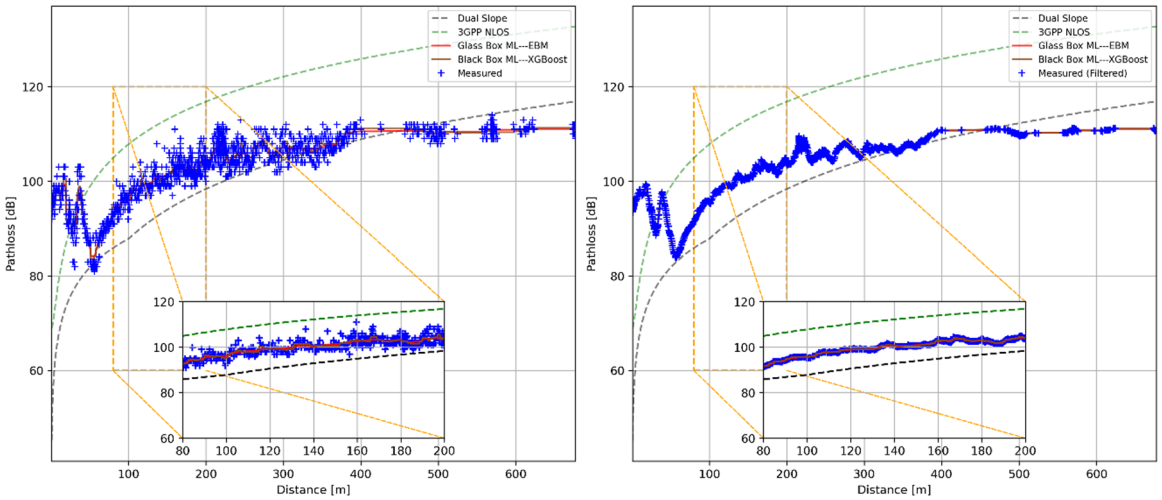

Fig. 10. Path loss predictions in Scenario T2 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (EBM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T2 Raw Data	RMSE	14.629	8.899	1.944	1.545	1.487	1.516	1.889	2.915
	MAE	15.727	13.977	2.443	2.031	1.980	2.006	2.557	3.425
	R^2	−4.136	−3.054	0.876	0.914	0.918	0.916	0.864	0.776
T2 Filtered Data	RMSE	14.649	8.715	0.982	0.453	0.339	0.314	1.139	2.165
	MAE	15.655	13.797	1.338	0.624	0.460	0.412	1.669	2.537
	R^2	−4.516	−3.301	0.960	0.991	0.995	0.996	0.937	0.849

Table 2. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T2) before and after data filtering and statistical model benchmarking.

Model accuracy evaluation

Case study 1: transmission power

This scenario looks at Case Study T2, which assesses path loss modeling for transmission power conditions. The scenario offers a predictable environment whereby signal propagation is typically predictable, although not exempt from variability over short-range and environmental dynamics. noise. Figure 10 presents comparisons between statistical model (3GPP NLOS, Dual Slope) estimates, optimum black-box (XGBoost) estimates, and optimum glass-box (EBM) model estimates with measured data both before and after filtering. For raw measurement (left panel), statistical models are not very accurate: the 3GPP model is always overestimating path loss and unable to depict fine details at near distances. However, machine learning models, such as EBM, are more consistent with measured data in spite of noise. After Kalman filtering (right panel), propagation patterns are smoother and measurement noises are attenuated. Both the XGBoost and EBM approaches are ideal matches for filtered datasets, especially in the near-range region (80–200 m) highlighted by the inset. Table 2 confirms these findings. Before filtering, EBM and XGBoost have RMSE measures of 0.905 and 1.069 and resulting R^2 scores of 0.981 and 0.974, substantially better than statistical models. After filtering, both models have an RMSE measure of 0.215 and an R^2 measure of 0.999, with very good prediction ability. In summary, for LOS environments, machine learning models, and explainable ones like EBM in particular, and their estimates of path loss are highly reliable and transparent. Also, by incorporating Kalman filtering, both reliability and explainability are enhanced across environments with fine propagation effects.

Case study 2: NLOS conditions

This case study applies Scenario T6 to demonstrate the influence of Exploratory Data Analysis (EDA) and Kalman filtering on path loss modeling. Scenario T6 describes a realistic and highly complex non-line-of-sight (NLOS) urban environment. Path loss estimates calculated from statistical models (3GPP NLOS, Dual Slope), as well as from the best-performing black-box model (XGBoost) and top-performing glass-box model (GAM), are compared with empirical data both prior to and subsequent to Kalman filtering (Fig. 11). Statistical models display relatively modest precision in initial measurements; namely, the 3GPP model tends to overestimate path loss, whereas the Dual Slope model does not effectively reflect variability in the near-range region. Conversely,

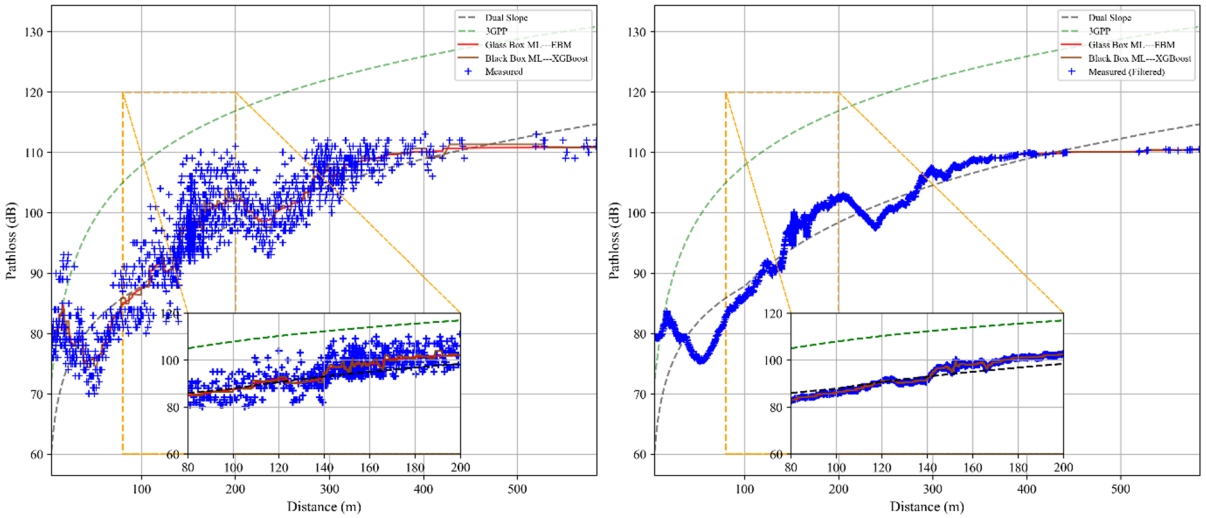

Fig. 11. Path loss predictions in Scenario T6 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (GAM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T6 Raw Data	RMSE	16.563	3.271	2.649	2.080	2.027	2.064	2.251	3.314
	MAE	17.057	4.482	3.334	2.921	2.860	2.871	3.063	3.756
	R^2	−3.916	0.659	0.812	0.856	0.861	0.860	0.841	0.719
T6 Filtered Data	RMSE	16.723	2.700	0.796	0.352	0.204	0.158	0.950	2.013
	MAE	17.077	3.466	1.017	0.460	0.277	0.215	1.198	1.891
	R^2	−4.654	0.767	0.980	0.996	0.999	0.999	0.972	0.850

Table 3. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T6) before and after data filtering and statistical model benchmarking.

machine learning models display strong performance: XGBoost follows strong nonlinear propagation behaviors closely, while GAM offers transparent and explainable visualization. Application of preprocessing through Kalman filtering effectively diminishes measurement noise substantially, boosting propagation trend intelligibility and boosting model conformity to filtered data, especially in terms of strong near-field variability (short-range area from 80 to 200 m). Qualitative metrics listed in Table 3 quantify the RMSE, MAE, and R^2 quantities before and after filtering. XGBoost and GAM qualify as top-performing black-box and best-performing glass-box models, respectively, displaying appreciable loss function reductions due to preprocessing. These findings confirm that comprehensive application of data smoothing through Kalman filtering and explainable machine learning models provides accurate, trustworthy, and transparent path loss estimates in highly demanding NLOS urban cases.

Case study 3: bridges/terrain elevation

This case study focuses on Scenario T7, where an arched bridge causes a rapid LOS–NLOS transition, creating elevation-induced obstructions that generate strong fluctuations in the measured path loss. Statistical models are unable to capture this behavior and therefore yield high prediction errors, whereas machine learning models, particularly RF, show strong robustness. After Kalman filtering, the RF model achieves near-optimal performance, confirming the effectiveness of combining data filtering and ML techniques for accurate path loss prediction in topographically irregular V2I environments. As illustrated in Fig. 12, the raw measured data exhibit pronounced variability around the bridge location, reflecting abrupt elevation-driven shadowing effects and rapid LOS–NLOS transitions. In this context, conventional statistical models (3GPP NLOS and Dual Slope) capture only the global attenuation trend and fail to follow local fluctuations. In contrast, the RF model closely tracks the nonlinear propagation behavior, while the GAM provides a smoother yet interpretable approximation of the path loss evolution. The application of Kalman filtering significantly reduces measurement noise and enhances trend consistency, leading to improved alignment between predictions and empirical data. The quantitative results reported in Table 4 corroborate these observations, showing substantial reductions in RMSE and MAE and a marked increase in R^2, with the RF model reaching near-optimal accuracy after filtering.

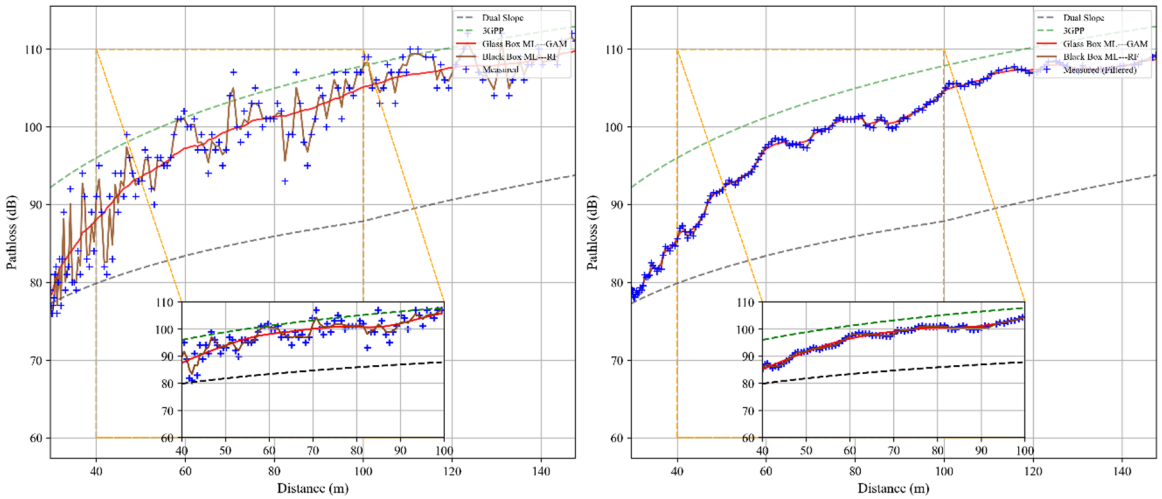

Fig. 12. Path loss predictions in Scenario T7 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (RF), and the best-performing glass-box model (GAM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T7 Raw Data	RMSE	5.599	12.674	2.706	2.663	2.789	2.721	2.617	3.680
	MAE	7.198	14.004	3.446	3.450	3.695	3.446	3.381	4.074
	R^2	0.493	−1.031	0.879	0.877	0.860	0.880	0.884	0.762
T7 Filtered Data	RMSE	6.368	11.610	0.555	0.396	0.507	0.485	0.644	1.707
	MAE	7.437	12.845	0.710	0.498	0.662	0.613	0.786	1.479
	R^2	0.439	−0.698	0.995	0.997	0.995	0.996	0.994	0.872

Table 4. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T7) before and after data filtering and statistical model benchmarking.

Case study 4: trees and vegetation

For this case study, Scenario T11 was chosen to demonstrate the effect of EDA and Kalman filtering on path loss modeling. The scenario depicts signal propagation in vegetative urban settings where trees and thick vegetation cause additional attenuation and multipaths such that traditional path loss models are inadequate. Figure 13 contrasts classical models (3GPP NLOS and Dual Slope) with machine learning models (EBM and XGBoost), both prior to (left) and after (right) Kalman filtering. The figure indicates how filtering affects model compliance with measured data. Pre-filtering: The raw measurements (blue crosses) contain extensive variability, especially at fine scales. The classical models are poorly precise and struggle to reflect these fine-scale fluctuations. On the other hand, machine learning models like EBM and XGBoost yield more precise approximations and effectively reflect nonlinear relationships in the data. Post-filtering: Noise substantially decreases after Kalman filtering such that smoother propagation patterns emerge. The filtered measurements (blue dots) closely resemble EBM estimates such that it was able to recover fine-grained dynamics in vegetative settings. Table 5 provides quantitative summaries about model performance. XGBoost and EBM represent the best-performing black-box and glass-box models, respectively. EBM attained nearly perfect agreement with filtered data ($R^2 = 0.999$). All things considered, Scenario T11 emphasizes that classical path loss models are inadequate in vegetative conditions and demonstrates how Kalman filtering is essential in enhancing prediction precision. A combination of interpretable machine learning and filtering provides a robust framework for wireless channel modeling under compound NLOS propagation effects in vegetation in cities.

Case study 5: roundabouts

For this case study, Scenario T16 was selected to investigate the effects of antenna height, transmission power, and urban obstacles on path loss modeling around roundabouts, which are highly dynamic environments. Frequent changes in vehicle trajectories induce rapid LOS-to-NLOS transitions, resulting in significant fluctuations in the received signal. Traditional statistical models, such as 3GPP NLOS and Dual Slope, are unable to capture these rapid variations, often producing large prediction errors. In contrast, the combination of Kalman filtering and interpretable machine learning models, particularly Explainable Boosting Machines (EBM), provides a robust and accurate framework for path loss prediction. This approach effectively mitigates measurement noise,

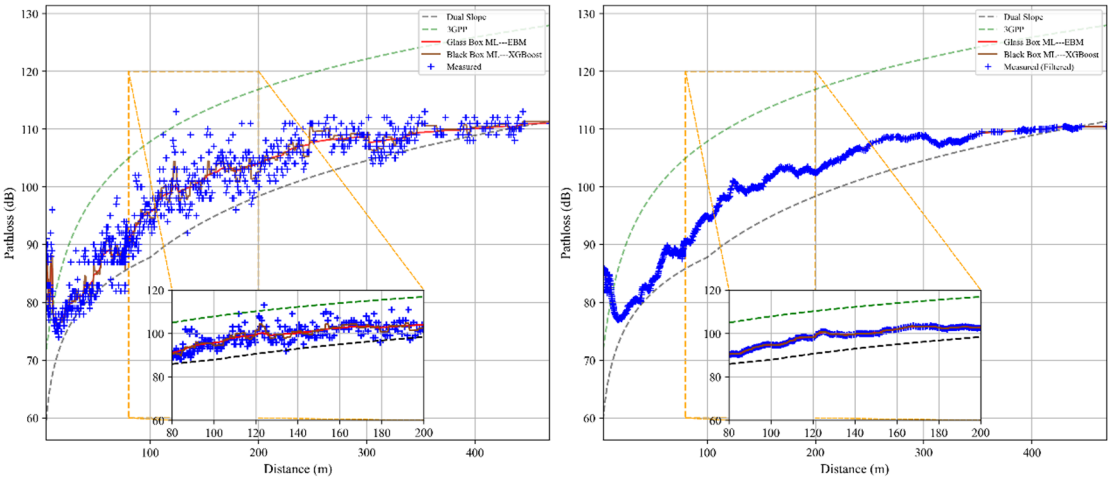

Fig. 13. Path loss predictions in Scenario T11 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (EBM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T11 Raw Data	RMSE	12.761	5.798	2.430	2.082	2.091	2.063	2.324	3.551
	MAE	13.289	7.607	3.146	2.868	2.861	2.779	3.081	4.062
	R^2	−0.480	0.511	0.917	0.931	0.931	0.935	0.920	0.701
T11 Filtered Data	RMSE	13.257	5.278	0.767	0.215	0.172	0.166	0.706	1.933
	MAE	13.527	6.908	1.009	0.290	0.233	0.212	1.000	1.981
	R^2	−0.691	0.557	0.991	0.999	0.999	1.000	0.991	0.772

Table 5. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T11) before and after data filtering and statistical model benchmarking.

captures nonlinear propagation effects, and accounts for environmental complexity, antenna configuration, and transmission power. As illustrated in Fig. 14, raw measurements show substantial variability, whereas filtered data align more closely with machine learning model outputs. The comparative performance of statistical, black-box, and interpretable models is summarized in Table 6, highlighting the superior predictive performance of this approach in dynamic urban scenarios where traditional models fail.

Case study 6: traffic

For this case study, Scenario T22 was selected to examine the impact of traffic dynamics, antenna configurations, and urban obstacles on path loss behavior in urban streets with heterogeneous vehicle density. This highly dynamic setting induces frequent LOS-to-NLOS transitions and pronounced signal fluctuations, posing significant challenges for traditional statistical propagation models such as 3GPP NLOS and Dual Slope, which fail to capture the fine-grained variability. Machine learning models, in contrast, exhibit strong predictive performance and robustness under these complex conditions. The integration of Kalman filtering effectively mitigates measurement noise, producing smoother propagation trends and enhancing the correspondence between model predictions and observed measurements. As illustrated in Fig. 15, raw measurements show substantial variability, whereas filtered data reveal clearer patterns that align closely with both black-box and glass-box machine learning models. The comparative performance of statistical, black-box, and interpretable models is summarized in Table 7, demonstrating that interpretable approaches can achieve high accuracy while preserving model transparency, thereby providing a reliable framework for modeling V2I propagation in dynamic urban environments.

Case study 7: heavy vehicles

For this case study, Scenario T24 was selected to examine the impact of heavy vehicles on path loss behavior in urban V2I environments. The presence of large vehicles introduces significant obstructions and rapid variations in the propagation channel, generating frequent LOS-to-NLOS transitions and pronounced fluctuations in received signal strength. Traditional statistical models, such as 3GPP NLOS and Dual Slope, are unable to accurately capture these complex dynamics, resulting in substantial prediction errors. In contrast, machine learning models, particularly ensemble-based black-box approaches like XGBoost and interpretable glass-

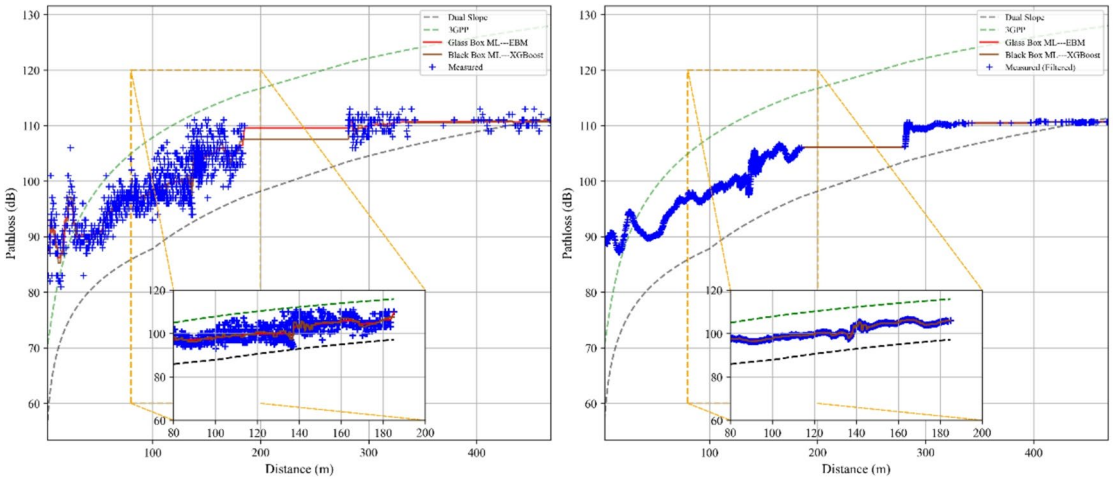

Fig. 14. Path loss predictions in Scenario T16 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (EBM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T16 Raw Data	RMSE	16.563	3.271	2.649	2.080	2.027	2.064	2.251	3.314
	MAE	17.057	4.482	3.334	2.921	2.860	2.871	3.063	3.756
	R^2	−3.916	0.659	0.812	0.856	0.861	0.860	0.841	0.719
T16 Filtered Data	RMSE	16.723	2.700	0.796	0.352	0.204	0.158	0.950	2.013
	MAE	17.077	3.466	1.017	0.460	0.277	0.215	1.198	1.891
	R^2	−4.654	0.767	0.980	0.996	0.999	0.999	0.972	0.850

Table 6. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T16) before and after data filtering and statistical model benchmarking.

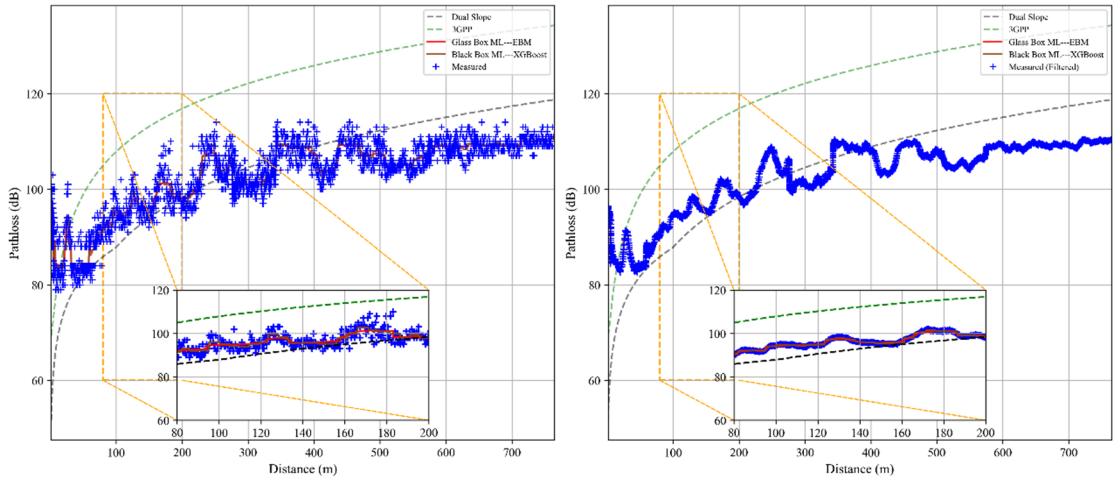

Fig. 15. Path loss predictions in Scenario T22 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (EBM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T22 Raw Data	RMSE	19.129	6.864	2.403	1.676	1.654	1.689	2.288	3.351
	MAE	19.849	10.606	2.958	2.314	2.284	2.304	3.005	3.698
	R^2	−5.119	−0.753	0.864	0.916	0.919	0.917	0.860	0.810
T22 Filtered Data	RMSE	19.192	6.628	1.579	0.515	0.408	0.362	1.480	2.543
	MAE	19.810	10.442	1.935	0.744	0.576	0.483	1.963	2.656
	R^2	−5.634	−0.845	0.935	0.991	0.994	0.996	0.935	0.885

Table 7. Average 5-Fold cross validation performance comparison of black box ML, and glass box ML models for case study (scenarios T22) before and after data filtering and statistical model benchmarking.

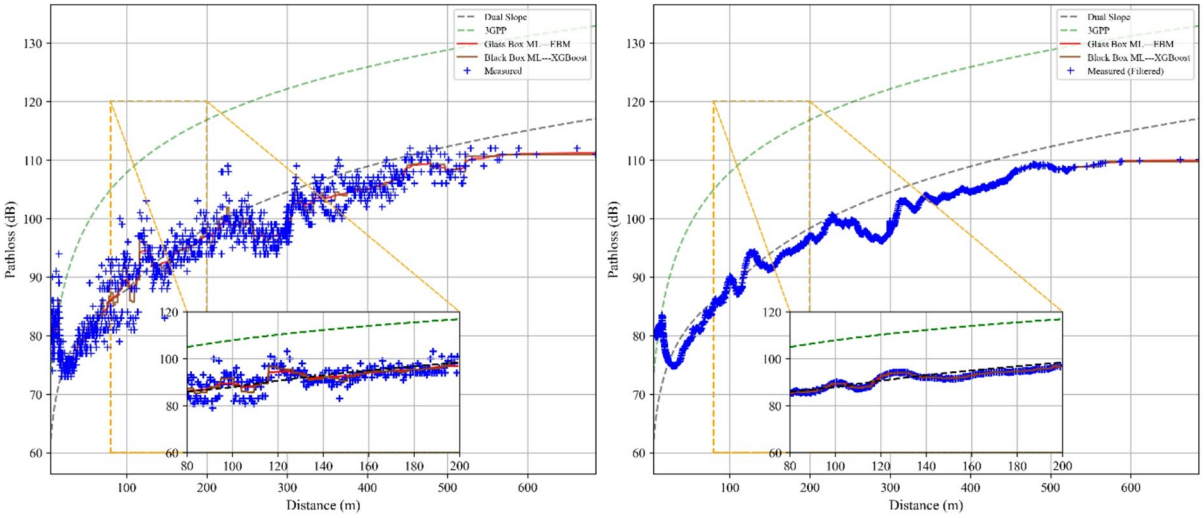

Fig. 16. Path loss predictions in Scenario T24 before and after Kalman filtering. Statistical models (Dual Slope, 3GPP NLOS), the best-performing black-box model (XGBoost), and the best-performing glass-box model (EBM) are compared against measured data.

Scenario	Metric	Statistical models		Black-box ML model			Glass-box ML model		
		3GPP	DS	AB	RF	XGB	EBM	GAM	GNAM
T24 Raw Data	RMSE	19.802	4.942	1.917	1.556	1.544	1.551	1.709	2.572
	MAE	20.600	5.943	2.616	2.300	2.256	2.253	2.484	3.477
	R^2	−4.231	0.564	0.916	0.935	0.937	0.937	0.924	0.875
T24 Filtered Data	RMSE	19.905	4.704	0.883	0.282	0.231	0.197	0.737	1.600
	MAE	20.673	5.618	1.072	0.362	0.318	0.268	1.100	2.093
	R^2	−4.598	0.586	0.985	0.998	0.999	0.999	0.984	0.935

Table 8. Average 5-Fold cross-validation performance comparison of black box ML, and glass box ML models for case study (scenarios T24) before and after data filtering and statistical model benchmarking.

box models such as EBM, demonstrate strong robustness and predictive performance under these challenging conditions. The application of Kalman filtering effectively mitigates measurement noise, smoothing propagation patterns and enhancing the alignment between model predictions and observed data. As illustrated in Fig. 16, raw measurements (left panel) exhibit high variability due to heavy vehicle-induced signal blockages, whereas filtered data (right panel) reveal more consistent propagation trends that closely follow machine learning model predictions. The comparative performance of statistical, black-box, and glass-box models is summarized in Table 8, highlighting that interpretable models can achieve near-optimal accuracy while maintaining transparency, providing a reliable framework for path loss modeling in environments dominated by heavy vehicles.

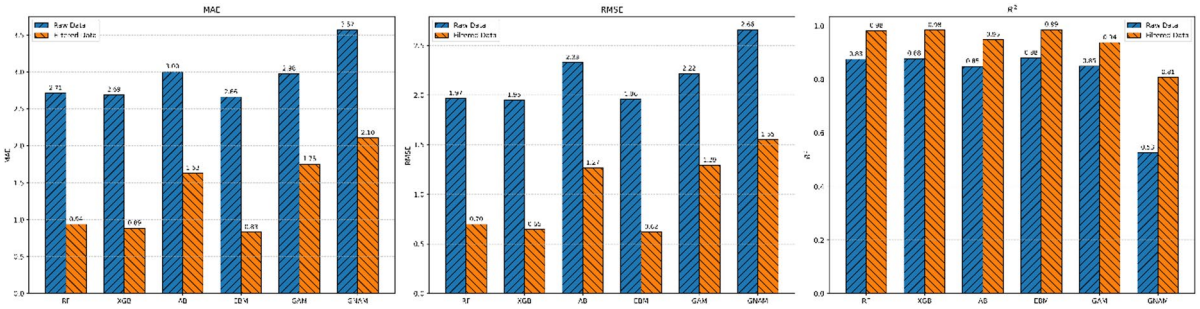

Fig. 17. Performance Metrics (MAE, RMSE, R^2) Across All 24 Scenarios: Raw vs. Kalman Filtered Pathloss.

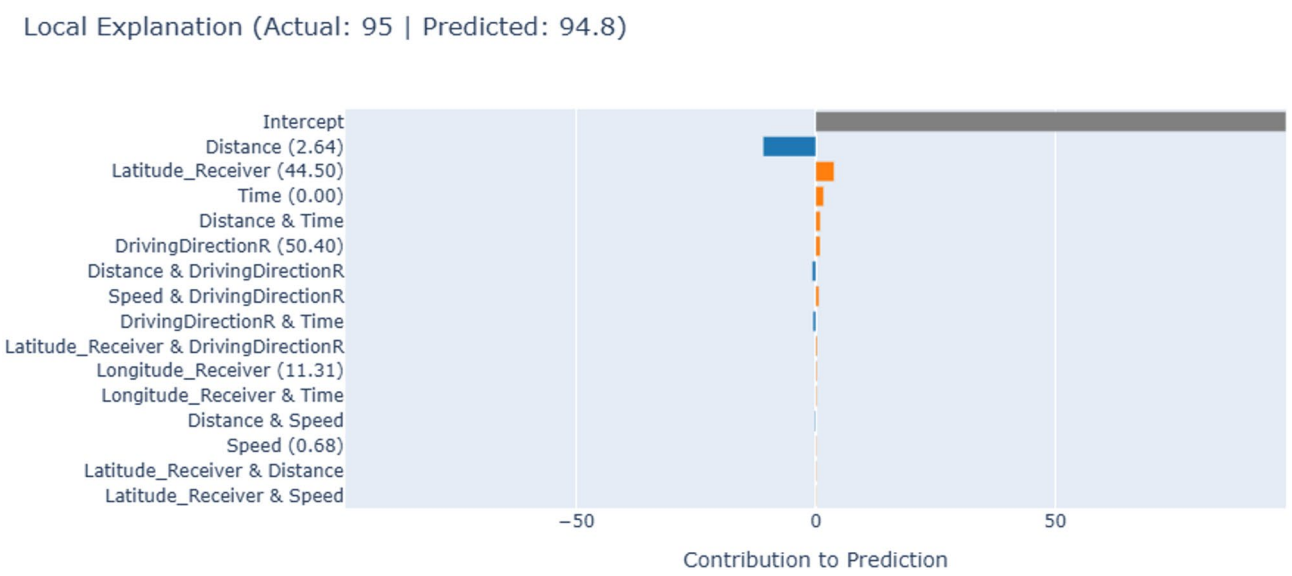

Fig. 18. Local feature contributions of the EBM model for a representative receiver point.

Further supplementary results on the performance of additional scenarios for the different case studies are presented in the Appendix.

Model global evaluation on all dataset scenarios

Following cross-validation, the trained models were evaluated on an independent test set using both raw and Kalman-filtered pathloss data for all 24 scenarios in the dataset, with MAE, RMSE, and R^2 computed to quantify predictive accuracy and assess model fit. As illustrated in Fig. 17, Kalman filtering consistently reduces prediction errors and enhances model performance, confirming that smoothing the pathloss data not only improves the generalization of the models but also ensures more reliable and realistic predictions for practical V2I communication Scenarios. These results complement the observations from the cross-validation phase and provide a robust indication of how the models are expected to perform under real-world conditions.

Interpretation of model predictions

Local explanations

Local explanations reveal how the EBM predicts pathloss at a specific receiver point by showing each feature's contribution, as shown in Eq. (25). In Fig. 18, contributions are visualized as horizontal bars: positive in orange, negative in blue, and intercept (β_0) in gray. The baseline pathloss (β_0) is ~95 units. The Distance × Driving direction receiver interaction dominates with ≈ 50.40 units, followed by Latitude receiver (≈ 44.50) and Longitude receiver (≈ 11.31). Distance alone adds ≈ 2.64 units, while other features, including speed, time, and minor interactions, have negligible effects.

This decomposition highlights that interactions between spatial and directional features drive the prediction, offering interpretable insights for scenario analysis, anomaly detection, and assessing feature importance.

Global explanations

Feature significance The significance of each input feature globally was measured using the mean absolute contribution over the entire training data. This quantified value represents how strongly both positive and negative influences affect corresponding predictions; hence, it captures the overall impact of features on model behavior.

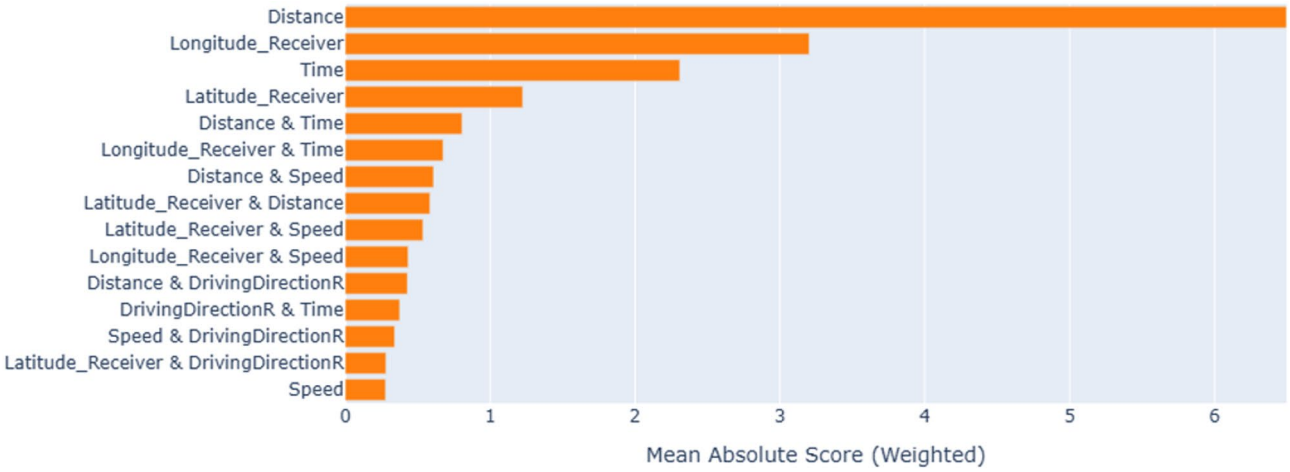

Fig. 19. Global feature importance from the EBM model for pathloss prediction.

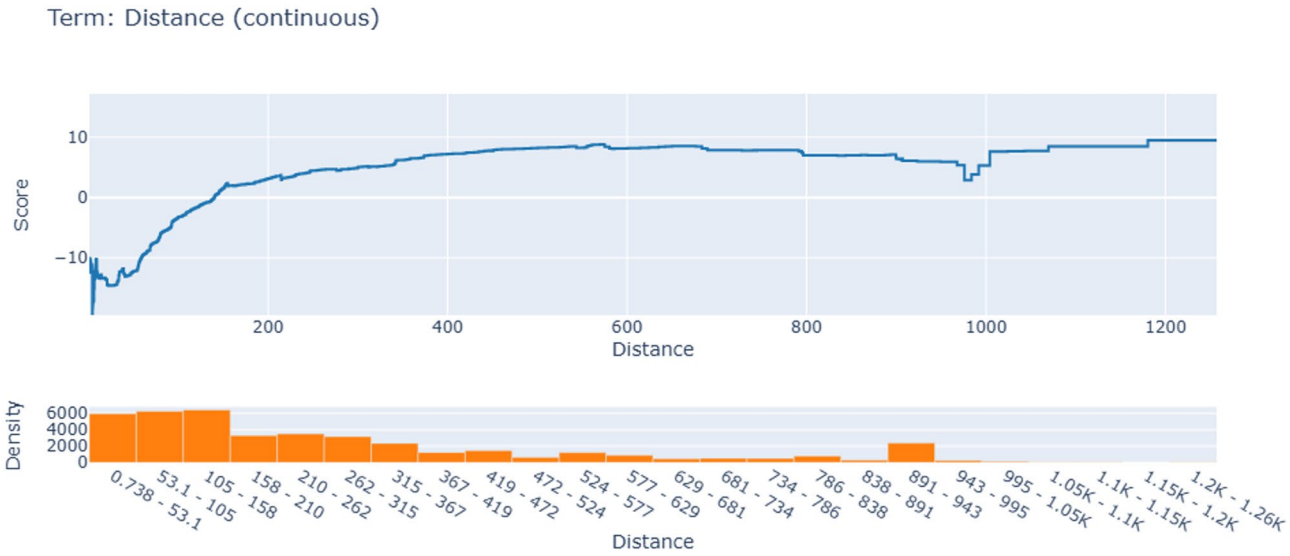

Fig. 20. Independent distance features and their marginal effects on EBM model for pathloss prediction.

In Fig. 19 above, Distance stands out as the most influential factor, followed by Longitude_Receiver and then Time. These features, in particular, are known to have a significant influence globally.

The geographic factor, on the other hand, referred to as Latitude_Receiver, has been found to be of moderate significance. Other interaction factors, including Distance & Time, Longitude_Receiver & Time, as well as Distance & Speed, also play a role in making the predictions.Other spatial-temporal interactions involve progressively decreasing contributions, which imply a limited role in modeling. On the other hand, spatial features like speed and driving direction interactions are found to lack significance.In sum, feature importances reinforce that spatial variables, especially those related to distance, are dominant in determining a model result, with temporal and interaction variables as supplementary factors.

Feature marginal contribution The marginal contribution function describes how each feature value affects the model prediction. In continuous features, EBM learns a shape function, which translates feature values into their corresponding contributions. Figure 20 above shows the marginal contribution of distance to the predicted path loss. The upper graph represents the contribution learned by EBM, showing how this value affects predictions with increasing distance. The histogram shown below illustrates how distance values are distributed in the data set, showing where in the graph data support for a particular value would be most substantiated.

Ethical, privacy, and practical deployment considerations

The use of vehicular data derived from Global Navigation Satellite Systems (GNSS) raises privacy concerns, which are mitigated through strict anonymization and the removal of all identifiable information. The interpretable models considered in this study, EBM, GAM, and GNAM, offer low-latency, resource-efficient inference suitable

Case Study	Transmission power	NLOS conditions	Bridges/terrain elevation	Trees and vegetation	Roundabouts	Traffic	Heavy vehicles
P test value	0.00019	0.00014	0.00081	0.00036	0.00026	0.00014	0.00014

Table 9. Statistical significance of ML models across all scenarios in case studies (Friedman Chi-Square Test).

for real-time deployment on RSUs and OBUs. Their transparent decision mechanisms facilitate auditing and accountability, enabling trustworthy and safe integration within V2I communication systems.

Ethical and privacy considerations are also paramount in V2I communication systems due to the continuous collection of sensitive vehicular data, including location, speed, and mobility patterns, which may expose users to privacy risks if improperly handled. Ensuring data anonymization, secure transmission, and compliance with data protection regulations is essential to prevent unauthorized access and misuse. Furthermore, the use of explainable (glass-box) models such as EBM, Amand GNAM models in safety-critical V2I applications enhances ethical accountability by enabling transparency, traceability, and human interpretability of model decisions. Such explainability feature provides trust, supports regulatory compliance, and allows system operators to understand, validate, and correct model behaviour, hence reducing the risk of unintended or unsafe actions in real-time vehicular environments.

Explainable models, such as EBM, GAM, and GNAM, can be effectively integrated into vehicular communication protocols and network planning tools by providing transparent, feature-level insights into path-loss predictions. These models enable network designers and V2I systems to identify critical factors influencing signal propagation, such as vehicle speed, distance, and environmental conditions. This, in process, allows network designers to make informed decisions on resource allocation, RSU placement, and transmission power adjustment. By incorporating interpretable outputs directly into protocol logic or planning dashboards, explainable models support adaptive routing, dynamic link selection, and predictive maintenance strategies, enhancing overall network reliability while maintaining transparency and accountability in safety-critical vehicular communication systems.

Statistical significance

A statistical significance test was carried out using the Friedman chi-square test to evaluate $P<0.05$ on the obtained 5-fold cross-validation RMSE values from the black box and glass box ML models for filtered datasets. The test results achieved in Table 9 indicate that the models' results are statistically significant.

Computational feasibility and energy consumption in real time V2I environment

In real-time V2I environments, path-loss prediction models must satisfy strict constraints on latency, computational complexity, and energy consumption, in addition to achieving high prediction accuracy. Black-box machine learning models such as RF, AB, and XGB can capture complex nonlinear propagation effects and often provide strong predictive performance. However, their reliance on large ensembles or deep architectures leads to higher computational overhead, increased inference latency, and elevated energy consumption, which can limit their feasibility for deployment on resource-constrained on-board units and roadside infrastructure.

In contrast, glass-box models, including EBM, GAM and GNAM offer a favorable balance between accuracy and efficiency. Their additive and interpretable structure enables rapid inference with reduced computational and energy requirements, making them well-suited for real-time V2I applications. Although black-box models may yield marginal accuracy gains, glass-box models often achieve competitive performance while ensuring transparency, scalability, and practical deployability, which are critical for reliable and energy-efficient V2I communication systems.

Discussion of findings

This study demonstrates that reliable V2I path loss prediction requires a rigorous, multi-stage methodology. EDA served as a diagnostic step, exposing structural anomalies, guiding preprocessing decisions, and providing insights into data distributions. The detection of stochastic fluctuations motivated the adoption of Kalman filtering as a targeted corrective measure. This filtering reduced the RMSE by more than 50%, effectively revealing deterministic propagation trends that were otherwise hidden in the raw measurements. As a result, models received cleaner, more informative input, significantly enhancing their learning capacity. Among the evaluated approaches, glass-box models benefited the most, achieving near optimal accuracy under NLOS conditions, with RMSE values consistently below 0.3 dB. Black-box models also improved after preprocessing but plateaued at higher error levels, highlighting their weaker alignment with physical propagation principles. These findings underscore a synergistic relationship between robust preprocessing and interpretable modeling: both model families outperformed traditional statistical baselines, yet only glass-box models successfully combined predictive performance with transparency. In conclusion, the results confirm that EDA, advanced filtering, and interpretable modeling collectively form the foundation for high-fidelity, generalizable wireless channel prediction, providing a reliable framework for practical V2I deployment and future enhancements[25–35].

Conclusion

This study presents a framework for accurate and interpretable V2I path loss prediction by combining EDA, glass-box machine learning models (EBM, GAM, and GNAM), and robust data filtering with optimized Kalman filtering. The approach delivers accuracy comparable to black-box models while providing clear insights into key factors for safety-critical V2X applications. By enhancing data stability and reliability under dynamic urban

conditions, it balances predictive performance with interpretability, crucial for designing and validating robust V2I communication systems. Future work will extend this framework to diverse environments, incorporate temporal dynamics, and explore hybrid physics-informed ML models to better capture rapidly varying vehicular channels[36–41].

Data availability

The dataset supporting this study is available from the corresponding author upon request due to the complexity of the logging structure; further details are available at [https://uwicore.umh.es/V2I-measurement-campaign].

Received: 9 October 2025; Accepted: 1 January 2026
Published online: 09 January 2026

References

1. Huang, J., Wang, C. X., Bai, L., Sun, J. & Yang, Y. A big data enabled channel model for 5G wireless communication systems. *IEEE Trans. Wireless Commun.* **17**(12), 8129–8141. https://doi.org/10.48550/arXiv.2002.12561 (2018).
2. Sung, S., Choi, W., Kim, H. & Jung, J. I. Deep learning-based path loss prediction for fifth-generation new radio vehicle communications. *IEEE Access.* **11**, 74494–74504. https://doi.org/10.1109/ACCESS.2023.3297215 (2023).
3. Ameur, M. B., Chebil, J., Habaebi, M. H. & Tahar, J. B. H. Machine learning for improved path loss prediction in urban vehicle-to-infrastructure communication systems. *Front. Artif. Intell.* https://doi.org/10.3389/frai.2025.1597981 (2025).
4. Gizzini, A. K., Medjahdi, Y., Ghandour, A. J. & Al-Dweik, A. Towards explainable AI for channel Estimation in wireless communications. *IEEE Trans. Wireless Commun.* https://doi.org/10.48550/arXiv.2307.00952 (2023).
5. Ayoub, O., Di Cicco, N., Ezzeddine, F. & Bruschetta, F. Explainable artificial intelligence in communication networks: A use case for failure identification in microwave networks. *Comput. Netw.* **217**, 109341. https://doi.org/10.1016/j.comnet.2022.109466 (2022).
6. Ribeiro, M. T., Singh, S., Guestrin, C. & Aug. Why should I trust you? Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)* 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
7. Yazıcı, M. & Gures, E. Interpretable machine learning-based path loss prediction in wireless communications. *IEEE Access.* **9**, 150300–150313. https://doi.org/10.3390/app9091908 (2021).
8. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems (NeurIPS)* 4765–4774 (2017). https://doi.org/10.48550/arXiv.1705.07874
9. Juang, C. H. Explainable AI for wireless channel modeling: SHAP-based feature contribution analysis. *IEEE Wirel. Commun. Lett.* **10**(12), 2660–2664. https://doi.org/10.1109/TMLCN.2025.3596548 (2021).
10. Vasudevan, M. & Yuksel, M. Machine learning for radio propagation modeling: A comprehensive survey. *IEEE Open. J. Commun. Soc.* **5**, 5123–5153. https://doi.org/10.1109/OJCOMS.2024.3446457 (2024).
11. Sun, Y., Zhang, J. & Letaief, K. B. Explainable machine learning for wireless communications: an application to path loss prediction. *IEEE Trans. Wireless Commun.* **21**(7), 5101–5115. https://doi.org/10.3390/app9091908 (2022).
12. Usama, M., Qayyum, A., Qadir, J. & Al-Fuqaha, A. Black-box adversarial machine learning attack on network traffic classification. *15th International Wireless Communications & Mobile Computing Conference (IWCMC)* 84–89 (2019). https://doi.org/10.1109/IWCMC.2019.8766505
13. Nori, H., Jenkins, S., Koch, P. & Caruana, R. Interpretml: A unified framework for machine learning interpretability. *J. Mach. Learn. Res.* **21**(130), 1–8. https://doi.org/10.1016/j.cscm.2024.e03350 (2020).
14. Redondi, A. E. C., Innamorati, C., Gallucci, S., Fiocchi, S. & Matera, F. A survey on future Millimeter-Wave communication applications. *IEEE Access.* **12**, 133165–133182. https://doi.org/10.1109/ACCESS.2024.3438625 (2024).
15. Gozalvez, J., Sepulcre, M. & Bauza, R. IEEE 802.11p vehicle-to-infrastructure communications in urban environments. *IEEE Commun. Mag.* **50**(5), 176–183. https://doi.org/10.1109/MCOM.2012.6194400 (2012).
16. Molisch, A. F. *Wireless Communications*, 2nd ed. (Wiley-IEEE Press, 2011). https://ieeexplore.ieee.org/servlet/opac?bknumber=5635423
17. Rappaport, T. S., Heath, R. W., Daniels, R. C. & Murdock, J. N. *Millimeter Wave Wireless Communications*. Pearson Education, (2015).
18. Fan, Y. et al. Measurements and characterization for the vehicle-to-infrastructure channel in urban and highway scenarios at 5.92 ghz. *China Commun.* **19**(4), 28–43. https://doi.org/10.23919/JCC.2022.04.003 (2022).
19. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
20. Ziad, M., Abdullah, A. & Mohamed, A. H. Advances and challenges in feature selection methods: A comprehensive review. *J. Artif. Intell. Metaheuristics*. https://doi.org/10.54216/JAIM.070105 (2024).
21. 3GPP. Study on channel model for frequencies from 0.5 to 100 GHz, 3rd Generation Partnership Project (3GPP), Technical Report TR 38.901, version 16.1.0 (2019).
22. Khalili, H., Frey, H. & Wimmer, M. A. Balancing prediction accuracy and explanation power of path loss modeling in a university campus environment via explainable AI. *Future Internet*. **17**(4), 155. https://doi.org/10.3390/fi17040155 (2025).
23. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223, 2019. https://doi.org/10.48550/arXiv.1909.09223
24. Yazici, I., Ozkan, E. & Gures, E. Enhancing path loss prediction through explainable machine learning approach. in *10th International Conference on Wireless Networks and Mobile Communications (WINCOM)* 1–6 (IEEE, 2023). https://doi.org/10.3390/fi17040155
25. Breiman, L. Random forests. *Mach. Learn.* 5–32. https://doi.org/10.1023/A:1010950718922 (2001).
26. Solomatine, D. P. & Shrestha, D. L. AdaBoost.RT: a boosting algorithm for regression problems, in *IEEE International Joint Conference on Neural Networks (IJCNN)*, vol. 2,1163–1168 (2004). https://doi.org/10.1109/IJCNN.2004.1380102
27. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785
28. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* https://doi.org/10.48550/arXiv.1705.07874 (2017).
29. Guo, W. Explainable artificial intelligence for 6G: improving trust between human and machine. *IEEE Commun. Mag.* **58**(6), 39–45. https://doi.org/10.1155/2024/8845070 (2020).
30. Agarwal, R. et al. Neural Additive Models: Interpretable Machine Learning with Neural Nets, arXiv preprint arXiv:2004.13912, (2021). https://doi.org/10.48550/arXiv.2004.13912
31. Zhang, X. & Andrews, J. G. Downlink cellular network analysis with multi-slope path loss models. *ArXiv Preprint*. https://doi.org/10.1109/TCOMM.2015.2413412 (2014).
32. Molina-Masegosa, R. et al. Dual-Slope path loss model for integrating vehicular sensing applications in urban and suburban environments. *Sensors* **24**(13), 4334. https://doi.org/10.3390/s24134334 (2024).

33. Elmezughi, M. K., Salih, O., Afullo, T. J. & Duffy, K. J. Path loss modeling based on neural networks and ensemble method for future wireless networks. *Heliyon* **9**(9). https://doi.org/10.1016/j.heliyon.2023.e19685 (2023). e19685.
34. Venkatraman, S., Talarico, S. & Cabric, D. Machine learning for wireless communications in the internet of things: A survey. *IEEE Commun. Surv. Tutorials*. **19**(3), 1617–1656. https://doi.org/10.1016/j.adhoc.2019.101913 (2017).
35. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
36. Kalman, R. E. A new approach to linear filtering and prediction problems. *Trans. ASME–Journal Basic. Eng.* **82**, 35–45 (1960).
37. Iqbal, S. et al. AD-CAM: enhancing interpretability of convolutional neural networks with a lightweight Framework- from black box to glass box. *IEEE J. Selec. Topics Signal Process.* **17**(6), 1201–1215. https://doi.org/10.1109/JBHI.2023.3329231 (2023).
38. Liao, W. et al. Explainable fault diagnosis of Oil-Immersed transformers: A Glass-Box model. *IEEE Trans. Power Deliv* **39**(1), 1–10. https://doi.org/10.1109/TIM.2024.3350131 (2024).
39. Liao, W. et al. Explainable modeling for wind power forecasting: A Glass-Box approach with high Accuracy, arxiv Preprint arxiv:2310.18629, (2023). https://doi.org/10.48550/arXiv.2310.18629
40. Alghamdi, T. A. & Javaid, N. A survey of preprocessing methods used for analysis of big data originated from smart grids. *IEEE Access.*, **10**, 29088–29107, https://doi.org/10.1109/ACCESS.2022.3157941 (2022).
41. Bianchi, G. & Tinnirello, I. Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network. in *IEEE International Conference on Communications (ICC)*, vol. 5, 3012–3016 (2003). https://doi.org/10.1109/INFCOM.2003.1208922

Acknowledgements
The authors express their thanks and appreciation to Dr. Mate Boban for providing access to the experimental data set.

Author contributions
MBA, MHH, JC, JBHT: Conceptualization, Methodology, Writing – Original Draft.MBA, MHH, JC: Data Curation, Software, Visualization.JC, JBHT, MRI, AMS: Supervision, Writing – Review & Editing, Project Administration.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-34987-8.

Correspondence and requests for materials should be addressed to M.H.H. or A.M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.