[1] **Awal Halifa ***

[2] **Emmanuel Affum Ampomah**

[3] **Kwame Oteng Gyasi**

[4] **Kwame Opuni-Boachie Obour Agyekum**

[5] **Kingsford Sarkodie Obeng Kwakye**

[6] **Piyush Kumar Shukla**

# A Comparative Analysis of Machine Learning Ensemble Methods for Accurate Path Loss Prediction


***Abstract:*** - This paper presents an ensemble path loss prediction model for wireless communication networks, leveraging machine learning. The model integrates three regression algorithms: Random Forest, Gradient Boosting, and Support Vector Regression. It is trained and tested using field data from diverse city environments. The model is optimized using feature engineering, hyperparameter tuning, and ensemble pruning techniques. Evaluation metrics, including Root Mean Square Error, Mean Square Error, and Mean Absolute Percentage Error, gauge the effectiveness of RF, GBR, SVR, and the ensemble methods. Notably, the bagging and blending ensemble models yield impressively low Mean Absolute Percentage Error values of 3.09% and 1.94%, respectively. Compared to existing empirical models, the proposed ensemble model achieved higher accuracy and generalization ability in path loss prediction, offering potential applications for network design and optimization.

***Keywords:*** Path Loss Prediction, Machine Learning, Ensemble Methods, Wireless Communication Networks, Model Optimization, Feature Selection.

## I. INTRODUCTION

To optimize communication systems, understanding electromagnetic wave (EM) propagation is crucial, as signal strength diminishes with increasing distance between transmitting and receiving antennas. Three propagation mechanisms: scattering, diffraction, and reflection—contribute to radio wave behavior [1]. Predicting path loss is complex due to the dynamic propagation environment, where path loss signifies the attenuation of radio waves during travel [2]. A precise path loss model is essential for coverage planning, base station site selection, and system performance enhancement.

Various path loss models have been explored, influenced by environmental factors impacting wireless signal propagation[3]. Empirical and deterministic methods, historically employed, have limitations in adapting to diverse scenarios [4]. Empirical models offer statistical correlations but may lack accuracy outside specific circumstances [5]. Deterministic models, while precise, demand extensive computational resources and site-specific information [6].

Machine learning, a data-driven approach, has emerged as a promising alternative to overcome these limitations. Support Vector Regression (SVR), Random Forest (RF), Artificial Neural Network (ANN), Gradient Boosting Regression (GBR), and K-Nearest Neighbor (KNN) are supervised learning regression models demonstrating potential in path loss prediction [7]. Machine learning surpasses empirical and deterministic models due to its ability to analyze vast datasets, offering enhanced accuracy [8].

### *A. Problem statement*

Traditional empirical path loss models are derived from statistical evaluations of measured data in specific environments, making them simple to use. However, their predictive accuracy often diminishes when applied to diverse or generalized deployment conditions. While these models can statistically represent the relationship between path loss and propagation parameters, they struggle to precisely predict power levels at specific locations. Alternatively, deterministic models deliver higher accuracy by using numerical analysis methods and radio wave propagation principles. Despite their precision, deterministic models require extensive computational resources and rely heavily on site-specific geometries and material parameters, making them impractical for large-scale or rapidly changing scenarios.

[1, 2, 3, 4, 5, 6] Department of Electrical and Electronics Engineering, Tamale Technical University, Tamale, Ghana
* Corresponding Author Email: ahalifa@tatu.edu.gh

In recent years, machine learning techniques have shown significant potential as alternatives for path loss prediction. These methods utilize large datasets and advanced algorithms to address the shortcomings of traditional models. Machine learning approaches such as Random Forest (RF), Gradient Boosting Regression (GBR), and Support Vector Regression (SVR) are capable of analyzing various input features—such as frequency, antenna height, distance, and environmental factors—to generate more accurate and adaptable predictions.

This study focuses on the application and optimization of existing machine learning techniques to create an ensemble model for path loss prediction. By improving prediction accuracy and ensuring adaptability to a variety of deployment environments, this research seeks to address the limitations of traditional empirical and deterministic models. The goal is to evaluate the ensemble model's performance against conventional methods and to provide meaningful insights for network design, coverage optimization, and system planning.

### *B. Specific Objectives*

1. To investigate the integration of predictive models for enhanced path loss prediction in wireless communication networks.
2. To develop and evaluate model optimization strategies for enhanced path loss prediction in wireless communication networks.
3. To evaluate and validate the performance of the proposed path loss prediction model against established empirical models.

## II. MACHINE LEARNING BASED PATHLOSS PREDICTION

The following literature review focuses on the various study domains on feature variables and feature selection strategies used to develop path loss prediction models. Several researchers have developed path loss prediction systems based on machine learning for a variety of area circumstances. One of them [9] uses an Artificial Neural Network (ANN) model to examine different indoor building styles. While [8] and [10] investigate various area types such as rural, suburban, and urban areas, [11] investigates path loss prediction in suburban regions using a range of machine learning models. Other studies that have been conducted in various locations with a unique measuring field include [12], which examines route loss prediction in an enclosed space such as an Airplane cabin. Path loss prediction of pathloss with the aid of ensemble machine learning methods has only been the subject of a few studies. This suggests that there is still opportunity for advancement in the study of path loss prediction with ensemble method approach.

In [13], the concept of path loss and how to predict it using machine learning was investigated. They also looked into the concept of path loss and defined the underlying premise of ML-based path loss predictors. We may use machine learning approaches to construct an adequate estimation function for path loss prediction if we have the results (path loss measurement) and the relevant input features, such as antenna separation distance and frequency. This function, which can be either a white box (in decision-tree-based models) or a black box (in SVR-based or ANN-based models), maps input features to path loss values [14].

The gathered information relates to measurement samples, each of which contains the path loss value and the associated input parameters. System-sensitive and environment-based parameters are the two types of input features. The propagation environment has little effect on characteristics such as carrier frequency, receiver as well as transmitter heights and positions, and other system-sensitive parameters. Additional system-dependent aspects, such as the antenna separation distance and the angle between the line-of-sight path and the horizontal plane, can be determined using the aforementioned parameters [15].

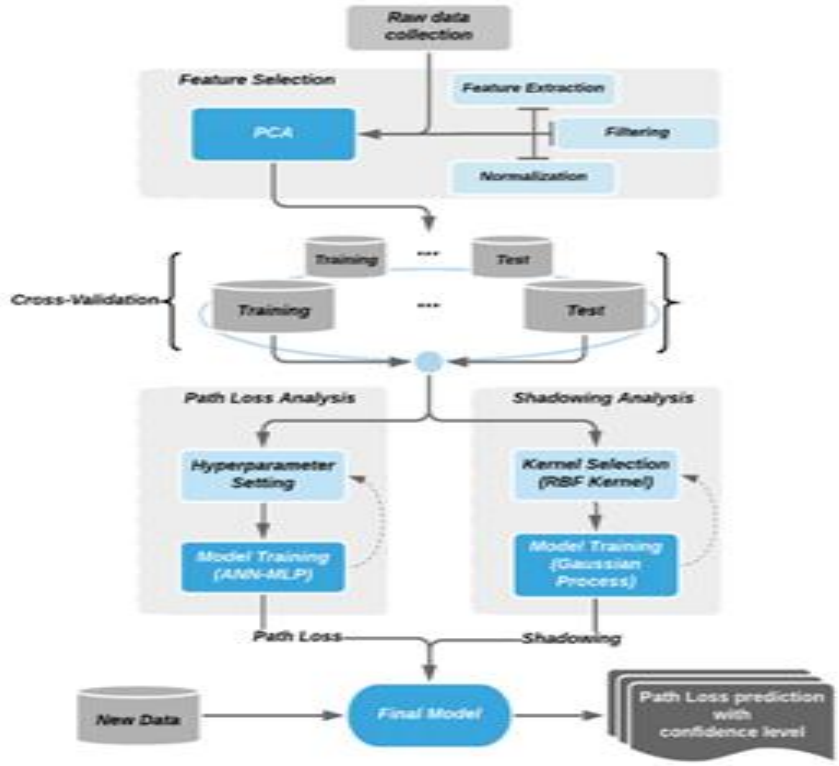


**Fig. 1.** Machine learning-based path loss analysis approach [13]

In order to develop the ability to generalize (to generate predictions based on previously unknown inputs), the models can be trained with sets of provided path loss values and matching inputs. We may use machine learning approaches to find a reasonable estimation function for path loss prediction once we know the output and the relevant input features such as antenna separation distance and frequency. This function maps input features to path loss values and can be either white (in decision-tree-based models) or black (in SVR-based or ANN-based models). The technique for machine learning-based path loss predictors is depicted in the picture below and is described in detail below.

*A. Feature Variables and Selection*

When creating path loss prediction models, a wide range of input feature kinds and quantities are used. The distance between the transmitter (TX) and receiver (RX) is the only input information used in the research in [13],[16]. The investigation conducted by [17] incorporates the frequency feature as an additional input, supplementing the TX-RX distance feature. In a similar vein, [4] introduces two extra features, along with the TX-RX distance feature, incorporating onboard GPS sensors. [18] includes the TX-RX distance feature as part of the input parameters, alongside PCC downlink throughput and PDCP downlink throughput. In the exploration carried out by [19], which focuses on the interior

of an aircraft cabin, the user's position, determined by longitude and latitude, serves as an input parameter, among others. Furthermore, both [20] and [21] examine outdoor locations using longitude and latitude in their respective studies. In certain research endeavors, environmental attributes are utilized as input features alongside system parameters. Reflecting the particular characteristics of the research domain, some studies opt for a more intricate blend of criteria to generate results of utmost accuracy. For instance, in the examination conducted by [22] and [23], a total of six input factors—longitude, latitude, elevation, altitude, clutter height, and TX-RX distance—are employed. This illustrates that the nature and quantity of features can be further refined to align with the specific focus of the research field.

*B. Feature Scaling*

In real life, the machine learning data may have hundreds of features. Poor predictor quality might result from either keeping irrelevant features or excluding important features. Finding the best subset with the fewest characteristics that contribute most to learning accuracy is the aim of feature selection [24].

The size of the input space can affect the performance of some machine-learning-based algorithms including RF, SVR, and GBR. As a result, the normalization procedure should be complete before the training starts. That

means, the values of all input characteristics and path loss should be modified to fall within the range of -1 to 1 or 0 to 1. This work uses the same normalization technique as [25], with the same results. It can be stated as

$$x_N = \frac{2(x-x_{min})}{x_{max}-x_{min}} - 1 \quad (1)$$

where x represents the value undergoing normalization, $x_{min}$ and $x_{max}$are the minimum and maximum values of the data range, respectively, and $x_N$ is the value after normalization. By applying anti-normalization in accordance with the normalization procedure, the expected values can be produced. By contrast, decision-tree-based approaches do not demand the feature scaling.

*C. Configuring Hyperparameters and Training the Model*

Hyperparameters, which are predefined values prior to the commencement of the learning process, encompass elements such as the number of hidden layers and neurons in an artificial neural network (ANN), the coefficients for regularization and parameters within the kernel function for support vector regression (SVR), and factors like ensemble size and tree depth in decision-tree-based approaches. To optimize the efficiency and performance of path loss prediction, it is crucial to carefully select an ideal set of hyperparameters. Grid search, random search, and Bayesian optimization stand out as prominent methods for hyperparameter optimization. In this investigation, the grid search methodology was employed to ascertain the final values of the hyperparameters. This approach entails a comprehensive search, evaluating all potential parameter values before identifying the most effective ones.

Model parameters, on the other hand, are inherent to training samples, evolving as part of the model training process. It is noteworthy that diverse learning methodologies involve distinct model parameters, with elements like weights and biases being autonomously acquired during the model training phase.

The research paper [26], discussed how they may train new models for predicting path loss within buildings or indoors. They argued that training is the most important and crucial role in the modelling problem. In reality, a well-trained model must be able to extrapolate or interpolate with high accuracy based on existing knowledge learned during learning from a new input in order to predict the expected output. The literature has a number of proposed neural network learning techniques, which can be categorized into supervised and unsupervised learning [27]. Unsupervised learning is used to cluster data, and this method separates the data into groups based on certain characteristics. With supervised learning techniques like the gradient descent approach, the input parameters and output values are known, and the neural network can offer an inferred function that can be used to map fresh samples. To lower the mean squared error (MSE) between the input and desired output of the neural model, the bias and weight of each neuron can be changed [27].

When creating a precise neural network model, the most influential parameters possible must be taken into account. The multi-wall model is where the inputs for the model that we developed and presented in this study come from. The elements of (L f) consist of the transmitter-receiver distance (d), frequency (f), and the attenuation caused by walls and floors (L w). The model is designed with a singular hidden layer. This hidden layer's number of neurons is set to 75% of the input layer's number of neurons [27]. This number may be altered, and the results of doing so will be discussed in the section that follows. In the output layer, there is only one output that represents the measured signal route loss.

*D. Model Evaluation and Prediction*

Typically, samples from the test dataset—which are absent from the model training process—are used to gauge how well machine learning-based route loss models perform. The evaluation criteria include complexity, generalization ability, and prediction accuracy. Performance metrics like maximum prediction error (MaxPE), mean absolute error (MAE), error standard deviation (ESD), root mean square error (RMSE), and mean absolute percentage error (MAPE) are commonly employed [28].

When deploying the model in scenarios with extra frequency bands or different environmental types, its reusability is determined by its generalization property. Gathering more data from diverse settings, including various terrains, frequencies, and vegetative cover conditions, has the potential to enhance the model's generalization capability.

Usually, processing speed and memory usage are used to gauge how difficult a computation is. Key factors influencing the processing duration of a machine learning model encompass, for example, the number of iterations and the speed at which convergence occurs throughout the training phase.

The machine learning algorithm can be chosen, the hyperparameters can be changed, and the prediction model can be further enhanced based on the evaluated outcomes. Following the construction of the ideal model, path loss values can be produced using fresh inputs.

In this study, we looked at the feature selection, hyperparameter tuning and optimization, and model selection and training available in wireless communication networks. It became clear, nonetheless, that specialized and improved path loss prediction features are needed due to the particular difficulties and complexity of heterogeneous smart city environments. Also, a more robust pathloss prediction model is needed to handle the limited features and make more accurate prediction.

## III. METHODOLOGY

This chapter goes through the resources and procedures employed. The procedure for gathering data, the suggested model for predicting pathloss, and a comparison of various models are all provided. The use of supervised learning techniques such as Random Forest, SVR, and Gradient Boosting Regressor to predict path loss is introduced, and the performance of these methods is examined using measured data. Additionally, a theory describing how signals travel through empty space is offered.

### *A. Random Forest*

The Random Forest (RF) ensemble learning method, as described in [29], is comprised of multiple regression trees. To enhance the prediction performance of each tree and address its weak robustness, a voting mechanism is employed. Breiman and Cutler introduced this distinctive non-parametric supervised machine learning technique known as Random Forest.

"Bootstrap aggregation" is the origin of the bagging technique known as RF. The primary concept behind bagging is to take a dataset, bag a weak learner like a decision tree on it, then create several bootstrap duplicates of the dataset and develop decision trees on them. To choose various training samples for each tree, Bootstrap aggregating is used. After training the trees using these samples, the ultimate outcome is determined by averaging the performance of each individual tree.

The RF remains a valuable instrument for reducing dimensionality or eliminating redundancy in datasets. While datasets with high-dimensional input features provide more information, the presence of redundant and unnecessary components can diminish prediction accuracy. In this investigation, the RF approach was applied to process the observed signal dataset, extracting relevant features while eliminating unnecessary and unimportant ones. The RF algorithms involve two or more primary hyperparameters that need specification before deploying them for regression analysis or data training [30]. One of these hyperparameters is the number of trees. The mathematical explanation of the RF input-output function model is detailed below.

$$RF(x_n, y_n) = \{f(x_n, \theta_m, y_n)\} \quad (2)$$

Here, $\theta_m$ represents the number of trees while $x_n, y_n$ denote the input and target output data, respectively. In this instance, a set of 200 trees was employed on the target measured signal datasets to accomplish the task of identifying the most valuable and informative subset of characteristics.

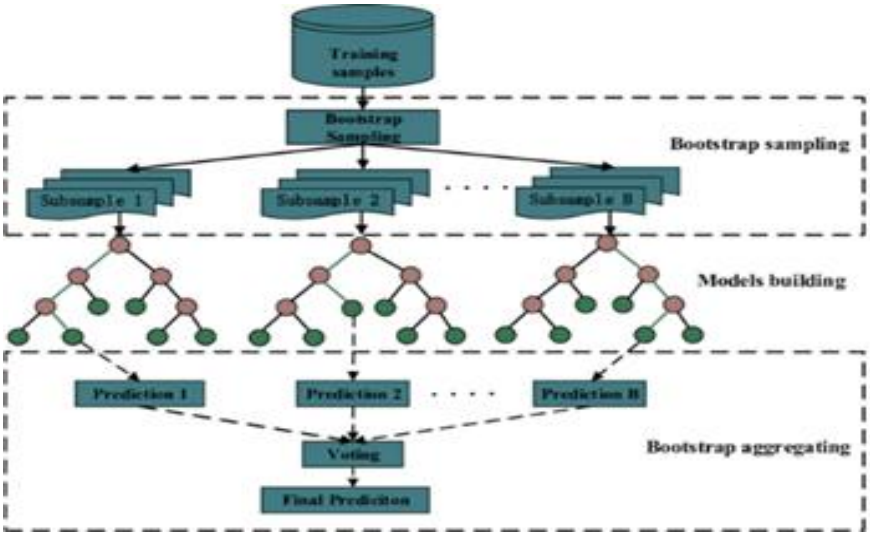


**Fig. 2.** Harnessing the combined predictions from every tree in the forest, the final outcome is determined through a majority vote [35].

*B. Gradient Boosting Regressor*

The Gradient Boosting Regressor (GBR) is a member of the boosting algorithms, falling under the category of ensemble learning techniques [31]. Its approach involves amalgamating the strengths of diverse regression trees to construct a predictive model. Originally introduced by Jerome Friedman [32], GBR is designed to address the limitations of specific weak learners, aiming to produce a robust and accurate regression model.

In GBR, a series of regression trees are trained in an iterative fashion, with each new tree aiming to fix the mistakes of the preceding ones. The model's ability to anticipate outcomes is enhanced over time by this iterative process.

Instead of using a voting mechanism, GBR concentrates on optimizing the residuals of the previous tree in the series. This is how it differs from Random Forest. Due to this quality,

GBR is very good at identifying complicated links and optimizing forecasts.

Giving additional importance to observations that earlier trees failed to accurately anticipate is the fundamental idea of boosting. Consequently, GBR can focus on the sections of the dataset where the model exhibits suboptimal performance. Each subsequent tree is trained with the specific aim of minimizing the accumulated residual errors within the ensemble. GBR does exceptionally well at adapting to the subtleties and patterns found in the data.

GBR's versatility in handling multiple data kinds and issue domains is one of its advantages. It can deliver excellent predicted accuracy and is ideally suited for datasets with diverse feature sets. To prevent overfitting, GBR might need more meticulous hyperparameter tuning than Random Forest. The gradient-based loss function is minimized by the GBR method through a series of processes, including initializing the target predictions, computing negative gradients, and creating new regression trees. Combining all of the various trees' projections yields the ultimate conclusion.

Based on the observed signal dataset, the Gradient Boosting Regressor was used in this work to forecast path loss. The goal of GBR is to offer precise path loss estimates while managing the difficulties present in real-world wireless communication settings.

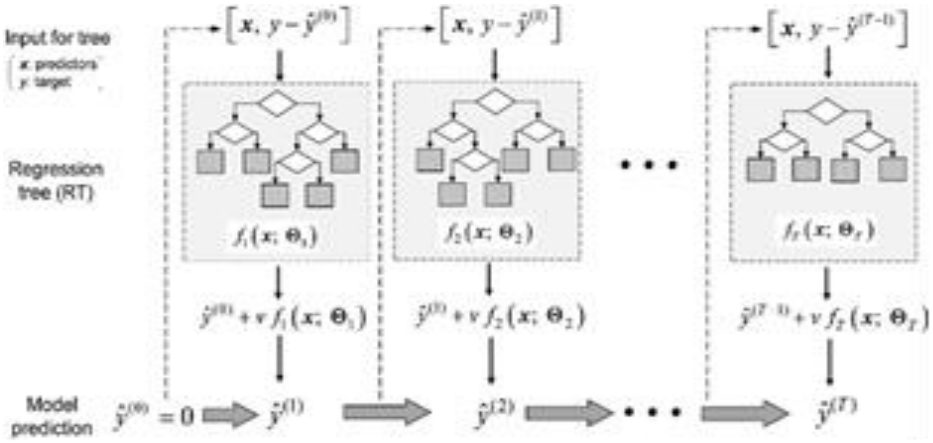


**Fig. 3.** Structure of the gradient boosting regressor [38].

*C. Support Vector Regressor (SVR)*

The support vector machine (SVM), rooted in statistical learning theory, represents a form of machine learning. SVM's core concept involves the linear separation of a dataset by nonlinearly transforming it from a finite-dimensional space to a higher-dimensional one. SVR, an extension of SVM tailored for regression challenges, enables path loss prediction [33]. The primary objective of SVR is to locate a hyperplane within the high-dimensional feature space and ensure that sample points align with it. The fundamental goal is expressed through the utilization of the following linear function to define the hyperplane in the feature space.

$$f(\boldsymbol{x}) = w^T \varphi(\boldsymbol{x}) + b \tag{3}$$

where x is an input feature vector, w is the normal vector that controls the orientation of the hyperplane, $\varphi(\cdot)$ is the nonlinear mapping function, and b is the displacement item.

The optimal hyperplane is formulated as a constrained optimization problem, as outlined in [38].

$$\min_{\boldsymbol{w}, b, \xi, \xi^*} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C \sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

$$s.t.f(x_i) - y_i \leq \varepsilon + \xi_i \quad (5)$$

$$y_i - f(x_i) \leq \varepsilon + \xi_i^* \quad (6)$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \ldots, N \quad (7)$$

Here, $C$ represents the regularization coefficient, $\varepsilon$ denotes the insensitive loss, signifying that a predicted value is deemed accurate if the difference between the predicted value and the actual value is less than $\varepsilon$. The variables $\xi_i, \xi_i^*$ are slack variables, introducing flexibility in the insensitivity range on both sides of the hyperplane, allowing for slight variations.

Then, by presenting Lagrange multipliers and solving its dual problem, the approximate function can be expressed as

$$f(x) = \sum_{i=1}^{N}(-\alpha_i + \alpha_i^*)K(x_i, x) + b \quad (8)$$

where $\alpha_i, \alpha_i^*$ are Lagrange multipliers, and $K(\cdot, \cdot)$ is a kernel function, which is used to perform the nonlinear mapping from the low-dimensional space to the high-dimensional space.

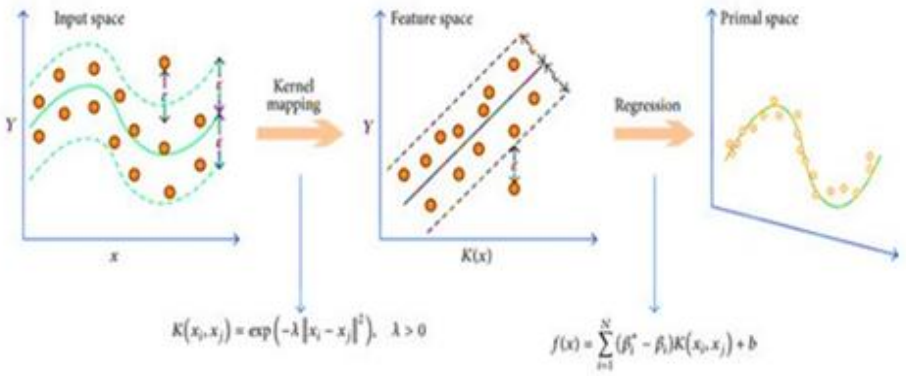


**Fig. 4.** A schematic diagram of SVR architecture [41]

The effectiveness of the SVR-based predictor relies on the choice of the kernel function. Presently, commonly used kernel functions include the sigmoid kernel, linear kernel, polynomial kernel, Gaussian radial basis function, and various combinations thereof. In this study, the selected kernel function is a Gaussian kernel with an adjustable parameter, and its definition is as follows:

$$K(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \gamma > 0 \quad (9)$$

The Gaussian kernel is often employed as a kernel function, particularly effective for tasks with restricted feature dimensions and no prior knowledge [34]. In this study, parameters like the regularization coefficient, insensitive loss, and kernel function parameter were determined using the same methodology as detailed in [35].

### *D. The proposed Pathloss Prediction Modeling Approach*

The proposed model, combines the strengths of the Random Forest, Gradient Boosting, and Support Vector Regression algorithms in accordance with the goals of the study. The model's creation, application, and evaluation are divided into several phases, as listed below:

1. Data Collection and Preprocessing: A thorough field measurement campaign is carried out to collect Received Signal Strength (RSS) values and route loss data from various metropolitan contexts. The gathered data includes a range of terrain characteristics, antenna heights, separations, and frequencies, ensuring its representativeness. To extract pertinent features for path loss prediction, feature engineering is used in the data pretreatment processes to handle missing values, identify outliers, and handle outliers.
2. Ensemble Model Development: The key idea behind the suggested method is to combine Gradient Boosting, Random Forest, and Support Vector Regression into a single predictive model. The appropriate hyperparameters and input features produced from the preprocessed data are used to train and optimize each particular model.
3. Model Integration and Weighted Averaging: The ensemble integration of the individual RF, GB, and SVR models is achieved through a weighted averaging mechanism. The weights are assigned depending on the performance and applicability of each model, which is combined with the predicted outputs from each model.

With the help of this integration, the predictive accuracy is intended to be improved by utilizing the complementing capabilities of several algorithms.

4. Cross-Validation and Model Tuning: The model is fine-tuned to optimize its hyperparameters and undergoes cross-validation to evaluate its generalization performance. The ensemble model is made stable and calibrated by this repeated process, enabling it to make precise predictions in a range of scenarios.
5. Performance Evaluation and Comparison: A thorough performance study is carried out in order to fully assess the suggested model. The assessment of the model involves employing diverse evaluation criteria, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and correlation coefficients. This comprehensive evaluation aims to gauge the accuracy, robustness, and generalization capabilities of the model. Using the same dataset, the model is rigorously contrasted with other path loss prediction methods, including Cost-Hata.

In order to provide a comprehensive solution for precise and reliable path loss prediction in heterogeneous radio network design, the Random Forest, Gradient Boosting, and Support Vector Regression, ensemble path loss prediction model is proposed. The model's effectiveness and superiority are proved by thorough validation and comparison with existing models, advancing path loss prediction in wireless communication networks for smart cities.

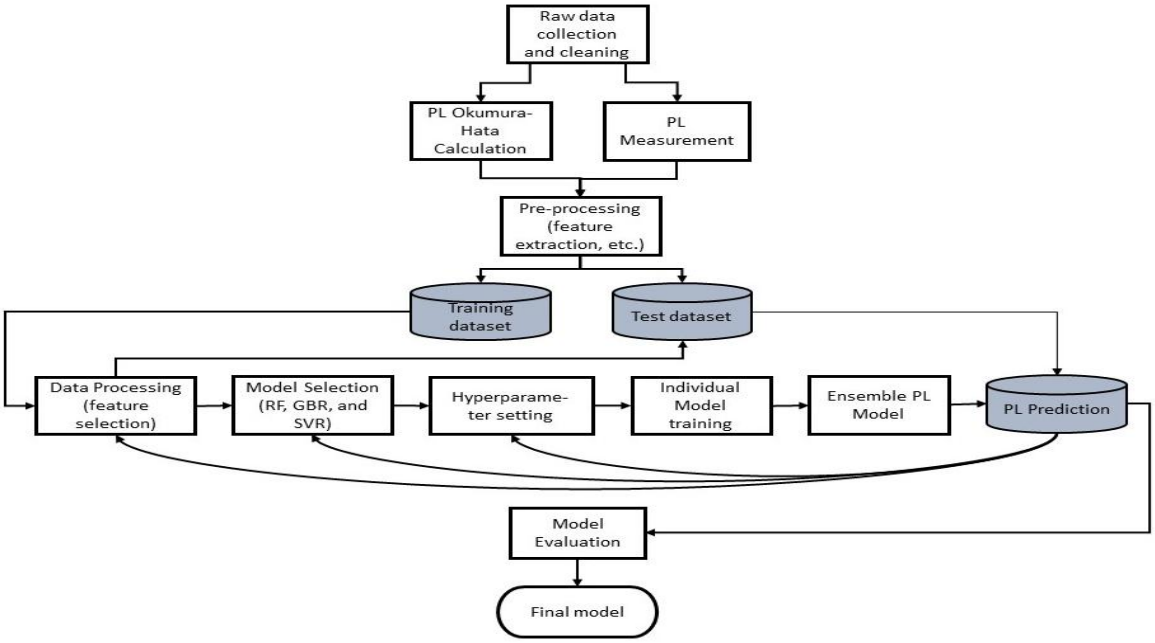


**Fig. 5.** The proposed Pathloss Prediction Model Architecture

*E. Model Algorithms*

***ALGORITHM 1:*** *TRAIN BAGGING MODEL*

***Inputs:*** *D (training data), r (number of models)*

***Output:*** *M (array of trained models) as empty array*

1. *Load the dataset (D)*
2. *Perform one-hot encoding for categorical variables*
3. *Split the dataset into training and testing sets (X_train, X_test, y_train, y_test)*
4. *Train individual models given input data X_train, y_train:*
5. *Train Random Forest model (M1) with hyperparameter tuning*
6. *Train Gradient Boosting model (M2) with hyperparameter tuning*
7. *Train Support Vector Machine (M3) with hyperparameter tuning*
8. *Append trained models M1, M2, and M3 to M*
9. *Return array of trained models M*

***ALGORITHM 2:*** *TESTING BAGGING MODEL*

***Inputs:*** *D (training data), r (number of models), M (array of trained models)*

***Output:*** *Y (bagging prediction)*

1. *Set parameters: M (array of trained models)*
2. *Initialize Y = 0*
3. *for each model m in M do*
4. *Determine prediction y from model m given input data X_test*
5. *Y = Y + y*
6. *end for*
7. $Y = \frac{Y}{r}$
8. *Return Y*

***ALGORITHM 3:*** *TRAIN BLENDING MODEL*

***Inputs:*** *D (training data)*

***Output:*** *M (array of trained models) as empty array*

1. *Load the dataset (D)*
2. *Perform one-hot encoding for categorical variables*
3. *Split the dataset into training and testing sets (X_train, X_test, y_train, y_test)*
4. *Train individual models given input data X_train, y_train:*
5. *Train Random Forest model (M1) with hyperparameter tuning*
6. *Train Gradient Boosting model (M2) with hyperparameter tuning*
7. *Train Support Vector Machine (M3) with hyperparameter tuning*
8. *Train a combiner model **C**, with validation data y_test. The combiner model is a regression model that takes the prediction from M1, M2, and M3 as inputs. The validation input data is used as inputs to M1, M2, and M3. The outputs from the three models constitute the input to **C**.*
9. *Return trained models M1, M2, M3, and combiner model C*

***ALGORITHM 4:*** *TESTING BLENDING MODEL*

***Inputs:*** *D (training data), M (array of trained models)*

***Output:*** *Y (blending prediction)*

1. *Set parameters: M (array of trained models), C (combiner model), M1 (RF model), M2 (GB model), M3 (SVM model)*
2. *Initialize Y = 0*
3. *Determine prediction m1 from model M1 given input data X_test*
4. *Determine prediction m2 from model M2 given input data X_test*
5. *Determine prediction m3 from model M3 given input data X_test*
6. *Determine prediction Y from model **C** given inputs m1, m2, and m3*
7. *Return Y*

### *F. Data Collection and Processing Campaign*

Numerous measurement campaigns were conducted in urban, and rural parts of Bekwai in the Ashanti Region, Ghana. Two city drives were used to gather experimental data. Field measurements were conducted with the assistance of a Transmission Evaluation and Monitoring System (TEM). Real-time data measurement, analysis, and post-processing across the network are all capabilities of TEMs. The measurement setup consists of a 4G Android phone acting as the mobile station (MS), a USB connector, GPS, a laptop, serial cables, and TEMs mobile system dongle software. In a repeated campaign scenario, reference signal received power (RSRP) was measured for all

sites between 23 m and 2.7 km in suburban, and urban regions of Bekwai using all of these linked devices. In order to reduce the Doppler effect, the vehicle moved at a constant speed. At an operational frequency ranging from 1.8 GHz to 2.1 GHz, numerous drive tests were carried out across networks.

For a transmitter-receiver distance ranging from 23 meters to 2.7 kilometers, the RSRP data were registered on the computer screen. The path loss was then calculated for every measured RSRP throughout the drive route. The first drive route received 2000 measurements, and the second drive route received 2330 measurements, for a total of 4330 measurements over the two drive routes. To achieve high precision, measurements were made at each spot eight times, with the average being determined. A reliable data preparation technique was used in the ensemble model building.

Before training the model, sufficient data preparation is required in order to attain accuracy in path loss predictions. Hundreds of features may be present in the actual data utilized for machine learning. Inaccurate path loss prediction might occur from both omitting or keeping unimportant data. The size of the input space has a significant impact on how machine learning models behave. Therefore, normalizing the data should be done before training the data. To prepare the data for this study, the following procedures were implemented. The data captured was imported into Python from an Excel file, where it was subsequently read and scrutinized for any instances of duplicate or missing values. Following this, the data underwent a normalization process.

Thirty percent of the pre-processed dataset was used for testing, with the remaining 70% going towards training. The use of uniform random sampling was made. In order to optimize the ensemble model, adjustments were made to the hyperparameter values using the training dataset. This allowed the incorporation of optimized network parameters in the RF, GBR, and SVR models, as well as the ensemble model. The equation below illustrates the calculation of path loss based on the measured received power.

$$Path\ loss\ (dB) = EIRP(dBm) - RSRP(dBm) \quad (10)$$

The aggregate power density transmitted from the base station to the surrounding medium is termed the effective isotropic radiated power (EIRP), represented in dBm.

**Table I:** Descriptive Statistics of Dataset

| Feature | Mean | Min | Max | STD | 25% | 50% | 75% |
|---|---|---|---|---|---|---|---|
| **Distance (m)** | 594.960 | 23.740 | 2677.090 | 431.213 | 311.143 | 479.385 | 742.555 |
| **Frequency (MHz)** | 1930.878 | 1800.000 | 2100.000 | 148.793 | 1800.000 | 1800.000 | 2100.000 |
| **Height of TX (m)** | 45.229 | 25.000 | 63.500 | 9.816 | 39.000 | 43.000 | 55.000 |
| **Height of RX (m)** | 14.896 | 3.998 | 30.033 | 5.230 | 11.087 | 14.850 | 18.293 |
| **Path loss (dB)** | 119.471 | 81.000 | 162.000 | 17.373 | 104.000 | 120.000 | 134.000 |

## IV. Results and Discussions

The research study's findings are presented in this section along with illuminating discussions. The findings are divided into separate segments, each of which sheds light on a different aspect of the examination, in order to be consistent with the study's aims. The anticipated values of path loss are contrasted with the measured values and the values obtained by applying empirical models in order to assess the viability of the suggested strategies.

### *A. Comparison of Different Models*

This section evaluates and compares the efficacy of machine learning models in predicting path loss against an empirical model (Cost-Hata). Multiple models are employed to predict the path loss value for each test data point. A total of 4330 samples were collected along the routes, each comprising path loss data and antenna separation distance computed using GPS information. For the training dataset, 70% of the samples were randomly selected, while the remaining 30% constituted the test dataset. To predict path loss values in the test dataset, three models—GBR, SVR, and RF—were utilized. The SVR model's regularization coefficient, insensitive loss, and kernel function parameter was set to 0.1, 10, and linear, respectively. The maximum tree depth and ensemble size for the RF and GBR based models, respectively, were 7 and 5, and there were 200 ensemble members. For comparison, the Cost-Hata model was also taken into account.

The measured data and the results that the various models are shown in Figures 6 to 9. The separation between the sending and receiving antennas is shown on the x-axis.


**Fig. 6.** Prediction accuracy of the Cost-Hata model on the test dataset

A significant disparity between the Cost-Hata model's predictions and the observed measured data indicates that the model struggles to accurately represent the actual path loss behavior. This gap reveals the inherent limitations of traditional empirical models, which are based on generalized assumptions and static parameters, making them less effective in accounting for the unique and dynamic features of specific environments.

To address this challenge, machine learning approaches, particularly ensemble techniques, can be employed. Ensemble models, which integrate the capabilities of multiple algorithms like Random Forest, Gradient Boosting, and Support Vector Regression, excel at learning directly from real-world data.

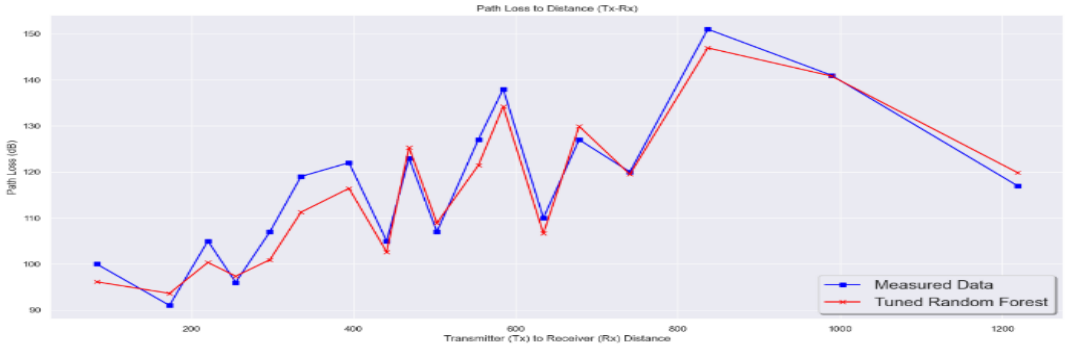


**Fig. 7.** Prediction accuracy of the Random Forest Regressor model on test dataset


**Fig. 8.** Prediction accuracy of the Gradient Boosting Regressor on test dataset


**Fig. 9.** Prediction accuracy of the SVR model on test dataset

Various models were employed to predict path loss values for each location within the test dataset. Subsequently, these estimated values were contrasted with the observed data, and the prediction errors were computed. The performance measures MAE, MAPE, RMSE, and MaxPE used to assess prediction performance are as follows:

$$MAE = \frac{1}{Q}\sum_{q=1}^{Q}\left|PL_q - PL'_q\right| \quad (11)$$

$$MAPE = \frac{100}{Q}\sum_{q=1}^{Q}\left|\frac{PL_q - PL'_q}{PL_q}\right| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(PL_q - PL'_q\right)^2} \quad (13)$$

$$ESD = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}\left(PL_q - PL'_q\right)^2} \quad (14)$$

$$MaxPE = \max\left(PL_q - PL'_q\right) \quad (15)$$

**Table II:** Performance evaluation of various models using 30% test samples from the measured dataset (up to 2.1GHz).

| METRIC | RF | GBR | SVR | Cost-Hata |
|---|---|---|---|---|
| **MAE (dB)** | 3.347 | 2.760 | 5.704 | 10.368 |
| **MAPE (%)** | 2.94 | 2.42 | 4.94 | 8.84 |
| **RMSE (dB)** | 4.220 | 3.574 | 7.359 | 12.267 |
| **MSE (dB)** | 17.812 | 12.772 | 54.152 | 150.475 |
| **MaxPE (dB)** | 15.622 | 16.930 | 25.047 | 31.557 |

The machine learning models' prediction errors are displayed in Table 2 above.

It is evident that the machine learning techniques performed well and outperformed the empirical model (Cost-Hata). In the above table, the Gradient Boosting Regressor algorithm outperformed the SVR, Random Forest Regressor, and Cost-Hata models based on the selected hyperparameters.

### *B. Development of the Ensemble Pathloss Models*

Various unique models are integrated in ensemble techniques to produce a more dependable and accurate prediction model. By using the diversity of numerous models to capture diverse features of the data, these strategies reduce the biases and mistakes prevalent in single models. Multiclassification approaches, multistage learning, and the integration of machine learning algorithms are all names for ensemble methods. The goal of ensemble approaches is to build a model through combination that is more accurate than a single isolated model. The challenge determines which ensemble machine learning techniques to use, including bagging, stacking (blending), and boosting. In this study, path loss is predicted using bagging and blending models. The ensemble method employs a labeled dataset to establish a mapping from the input to the corresponding output, constituting a supervised machine learning procedure. Labelled data were used in each of the basis models that were used to create the ensemble model.

### *C. Bagging Ensemble Pathloss Prediction Model*

A reliable strategy for improving the precision and dependability of prediction models is the bagging ensemble method, often known as bootstrap aggregating. In order to provide a coherent final prediction, this technique combines several basic learners. Bagging computes the mean of predictions from these base models in the context of regression tasks, such as path loss prediction. It's important to remember that each base learner is taught using replacement learning on a randomly chosen portion of the initial training data.

This study focuses on Random Forest, Support Vector Machine, and Gradient Boosting Regressor, three independent regression models that have proven adept at detecting complex patterns inside data. The fact that these models can handle non-linear interactions and incorporate complex dependencies makes them interesting options.

Following training of the foundation models, we use a deterministic averaging strategy to combine them into an ensemble model. The final forecast for each data point is produced by combining the predictions from the RF, GBR, and SVM models. The ensemble will benefit from each base model's unique qualities thanks to this integration.

*D. Blending Ensemble Pathloss Prediction Model*

The predictive capacity of various base models, such as Random Forest, Support Vector Machine, and Gradient Boosting Regressor, is combined using the ensemble technique of blending to produce a reliable and accurate path loss prediction model. This strategy uses a combiner model to carefully combine the predictions of the various models while leveraging their strengths. A synopsis of the blended ensemble method and how it can be used to forecast path loss is provided in this study.

A dataset with attributes useful for path loss prediction is used to train each of the three base models, RF, GBR, and SVR. The fundamental links between input data and path loss are captured in different ways by these models as they develop. The base models' hyperparameters, such as the number of trees in RF and the learning rate in GBR, are adjusted to maximize each model's performance. Each base model is optimized for accuracy through the use of hyperparameters. A fusion model, a regression model incorporating predictions from the RF, GBR, and SVR models, was trained using the validation dataset. The test phase commenced with the results obtained from the trained RF, GBR, SVR, and fusion models. Path loss for the RF, GBR, and SVR models that were returned from the training phase was estimated for the blended ensemble model's prediction phase using the input data. The combiner model then used the same sampled data to forecast path loss using the prediction outcomes from the three models. The combiner models' prediction results were then sent back.

*E. Experimental Results*

Nine candidate variables altogether, made up of system parameters and environmental parameters, were included in the preliminary data for this study. The Cost-Hata model, an empirical model, was used to perform the first step's calculations. In this model, there were only seven parameter variables used: the distance, the frequency, the height of the TX and the RX, the angle between the RX and the main beam TX (vertical and horizontal), and the height of the ambient building. These data were employed to generate the estimated path loss value and determine the delta value, representing the disparity between the calculated path loss and the measured path loss. The process of feature selection was then used to analyses which of the nine candidate variables were chosen as the best variables, as displayed in Table 2.

**Table III:** Candidate variables

| Name | Description | Level |
|---|---|---|
| Distance | transmitter (TX) and receiver (RX) distance | meters |
| Frequency | Frequency used in signal transmission | MHz |
| Height TX | Transmitter antenna height + altitude location | meters |
| Height RX | Receiver antenna height + altitude location | meters |
| Terrain | The characteristics of the geographical landscape between the transmitter (TX) and receiver (RX) | rural, suburban, urban |
| Height of Building | Surrounding building height | meters |
| Distance between Building | Distance between surrounding buildings | meters |
| Vertical angle | The angular disparity between the vertical orientation of the antenna and that of the receiver. | degree |
| Horizontal angle | The angular difference between the horizontal azimuth of the antenna and the horizontal orientation of the receiver. | degree |

*F. Ensemble Methods Evaluation*

In this study, three distinct machine learning models were employed and integrated into an ensemble method to enhance path loss prediction. Correlation plots between the measured data and the predictions from the machine learning models are presented for both the training and test datasets. Validation was carried out using metrics such as mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and maximum prediction error (MaxPE). The dataset samples were partitioned, with 70% assigned to the training set and 30% to the test set.

To enhance the performance of the models depicted in Figure 10, adjustments to hyperparameters were implemented. Through a method known as grid search or random search, which investigates different parameter combinations to find the most efficient configuration for each model, the values of the hyperparameters are methodically adjusted. Table 3 lists the tuning-relevant hyperparameters.

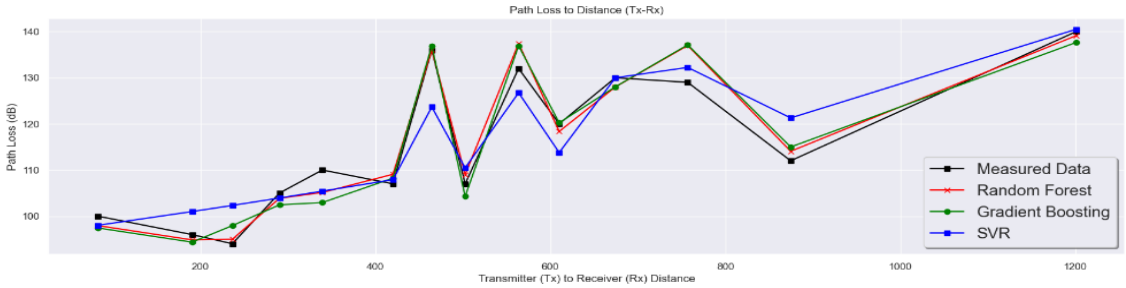


**Fig. 10.** Pathloss prediction of the various prediction models on the test **dataset** without hyperparameter tuning.

**Table IV :** Tuned Hyperparameter Values

| Hyperparameters | RF | GBR | SVM |
|---|---|---|---|
| n_estimators | [100, 200, 300] | [100, 200, 300] | NA |
| max_depth | [3, 5, 7] | [3, 5, 7] | NA |
| learning_rate | NA | 0.1 | NA |
| Kernel | NA | NA | ['linear', 'rbf'] |
| C | NA | NA | [0.1, 1, 10] |
| Epsilon | NA | NA | [0.01, 0.1, 1] |
| Gamma | NA | NA | ['scale', 'auto'] + list (np.logspace(-3, 3, 7)) |
| Optimization Algorithm | GridSearchCV | GridSearchCV | RandomSearchCV |


**Fig. 11.** Pathloss prediction of the various prediction models on the **test** dataset with hyperparameter tuning.

Hyperparameter adjustment is done in order to improve the functionality of the models in Figure 11. Through a method known as grid search or random search, which investigates different parameter combinations to find the most efficient configuration for each model, the values of the hyperparameters are methodically adjusted. Table IV lists the tuning-relevant hyperparameters.


**Fig. 12.** Prediction accuracy of Bagging Ensemble on the test dataset.


**Fig. 13.** Prediction accuracy of Blending Ensemble on the test dataset.

The superiority of the blended ensemble model over other machine learning models is evident from the plots in Figures 11, 12, and 13, as well as the performance metrics detailed in Table IV. Consequently, it is deemed suitable for achieving precise signal propagation. In machine learning, the incorporation of an ensemble is a strategy employed to boost the overall model performance, a goal successfully achieved in this study. Notably, the blending algorithm demonstrated the lowest error rates for both the training and test datasets. The SVR standalone model has the greatest number of mistakes. As a result, the model's performance, robustness, and forecast accuracy were all enhanced by the blended ensemble method. Additionally, the bagging technique outperformed the RF, GBR, and SVR models.

Figures 11 depict each machine learning model plotted alongside the measured path loss. The blending ensemble model stands out with a smaller mean distance between the points around the fitted line compared to the other three models. As the mean distance increases, the MSE value decreases. Specifically, the MSE value for the blending ensemble model is 10.397 dB, as detailed in Table IV. This result indicates that the blended ensemble model effectively and closely predicted path loss in accordance with the measured data. According to Table IV, the blending ensemble model exhibits the highest accuracy and the least overall error in predicting path loss among the considered models.

**Table V:** Performance Metrics for the Developed Models on 30% of the Dataset (Test Set)

| Metrics | RFR | GBR | SVR | BAGGING | BLENDING |
|---|---|---|---|---|---|
| MAE (dB) | 3.347 | 2.760 | 5.704 | 3.544 | 2.792 |
| MAPE (%) | 2.94 | 2.42 | 4.94 | 3.09 | 1.94 |
| RMSE (dB) | 4.220 | 3.574 | 7.359 | 4.464 | 3.160 |
| MSE (dB) | 17.812 | 12.772 | 54.152 | 19.923 | 10.397 |

## V. CONCLUSION

This study presents an innovative approach to path loss prediction models through ensemble techniques, specifically employing Support Vector Regression, Random Forest and Gradient Boosting models. The ensemble models, including bagging and blending, demonstrated enhanced performance in mitigating errors and reducing variation. Notably, the blended ensemble method stood out, showcasing superior accuracy in path loss prediction for wireless communication channels. This suggests that ensemble approaches outperform standalone machine learning models, providing a robust mechanism for accurate signal categorization in diverse scenarios. For future research, recommendations include focused efforts on data collection, emphasizing diversity and uniformity, strategic feature selection methodologies, addressing hyperparameter optimization challenges, and exploring incremental learning algorithms to adapt to evolving training datasets.

For the collection of training data, considerations of sample diversity and uniformity are crucial, requiring careful planning of measurement routes to ensure high generalization properties. Feature selection methodologies should be developed to guide the inclusion of relevant characteristics in path loss predictors, avoiding the "curse of dimensionality." Additionally, addressing hyperparameter optimization challenges and exploring incremental learning algorithms can enhance the adaptability of path loss predictors to evolving datasets over time, improving overall accuracy and efficiency in real-world scenarios.

## REFERENCES


[1] A. Saakian, *Radio wave propagation fundamentals*. Artech House, 2020.

[2] Oladimeji,et al "Propagation path loss prediction modelling in enclosed environments for 5G networks: A review," *Heliyon*, vol. 8, 2022.

[3] Sun, Shu and Rappaport,et al "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 2843--2860, 2016.

[4] Zhang, Yan and Wen,et al "Path loss prediction based on machine learning: Principle, method, and data expansion," *Appl. Sci.*, vol. 9, 2019.

[5] Wright, Aidan GC and Kruege, "The structure of psychopathology: toward an expanded quantitative empirical model," *J. Abnorm. Psychol.*, vol. 122, 2013.

[6] Aram, Morteza Ghaderi and Guo, "Site-Specific Outdoor Propagation Assessment and Ray-Tracing Analysis for Wireless Digital Twins," *arXiv Prepr. arXiv2410.14620*, 2024.

[7] Hussain, Sajjad, "Geometrical Features based mmWave UAV Path Loss Prediction using Machine Learning for 5G and Beyond," *IEEE Open J. Commun. Soc.*, 2024.

[8] Angelov, Plamen P, "Empirical approach to machine learning," *Springer*, 2019.

[9] Mao, Qian and Hu, "Deep learning for intelligent wireless networks: A comprehensive survey," *IEEE Commun. Surv.*, vol. 20, pp. 2595--2621, 2018.

[10] J. R. de F. Cabral, "A Machine Learning Approach for Path Loss Estimation in Emerging Wireless Networks," pp. 1–54, 2019.

[11] Iliev, Ilia and Velchev, Yuliyan and Petkov, "A Machine Learning Approach for Path Loss Prediction Using Combination of Regression and Classification Models," *Sensors*, vol. 24, 2024.

[12] Frattasi, Simone and Della Rosa, *Mobile positioning and tracking: from conventional to cooperative techniques*. John Wiley & Sons, 2017.

[13] Elmezughi, Mohamed K and Salih, "Comparative analysis of major machine-learning-based path loss models for enclosed indoor channels," *Sensors*, vol. 22, 2022.

[14] Zhang, Yan and Wen, "Path loss prediction based on machine learning: Principle, method, and data expansion," *{Applied Sci.*, vol. 9, 2019.

[15] A. Sani, "Modelling and characterisation of antennas and propagation for body-centric wireless communication," Queen Mary University of London, 2010.

[16] Elmezughi, Mohamed K and Salih, "Path loss modeling based on neural networks and ensemble method for future wireless networks," *Heliyon*, vol. 9, 2023.

[17] Wang, Chenlong and Ai, "Channel path loss prediction using satellite images: A deep learning approach," *IEEE Trans. Mach. Learn. Commun. Netw.*, 2024.

[18] Ojo, Stephen and Imoize, "Radial basis function neural network path loss prediction model for LTE networks in multitransmitter signal propagation environments," *Int. J. Commun. Syst.*, vol. 34, 2021.

[19] Wen, Jinxiao and Zhang, "Path loss prediction based on machine learning methods for aircraft cabin environments," *IEEE Access*, 2019.

[20] Saito, Kentaro and JIN, "Two-step Path Loss Prediction Method by Artificial Neural Network for Wireless Service Area Planning," *{IEICE Tech. Rep.*, 2020.

[21] Tahat, Ashraf and Edwan, "Simplistic machine learning-based air-to-ground path loss modeling in an urban environment," IEEE, 2020.

[22] Popoola, Segun I and Adetiba, "Optimal model for path loss predictions using feed-forward neural networks," *Cogent Eng.*, 2018.

[23] Singh, Harsh and Gupta, "Path loss prediction in smart campus environment: Machine learning-based approaches," IEEE, 2020.

[24] Khalid, Samina and Khalil, "A survey of feature selection and feature extraction techniques in machine learning," {IEEE, 2014.

[25] Wu, Di and Zhu, "Application of artificial neural networks for path loss prediction in railway environments," IEEE, 2010.

[26] Zineb, Aymen Ben, "A multi-wall and multi-frequency indoor path loss prediction model using artificial neural networks," *Arab. J. Sci. Eng.*, 2016.

[27] Pearlmutter, Barak A, "Gradient calculations for dynamic recurrent neural networks: A survey," *IEEE Trans. Neural networks*, vol. 6, 2009.

[28] Isabona, Joseph and Srivastava, "Hybrid neural network approach for predicting signal propagation loss in urban microcells," IEEE, 2016.

[29] Mahesh, "Machine learning algorithms-a review," *Int. J. Sci. Res.*, vol. 9, 2020.

[30] H. and B. Yuliana, "Hyperparameter Optimization of Random Forest Algorithm to Enhance Performance Metric Evaluation of 5G Coverage Prediction," *Bul. Pos dan Telekomun.*, vol. 22, 2024.

[31] Chen, Tianqi and Guestrin, "A scalable tree boosting system," 2016.

[32] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann. Stat.*, 2001.

[33] P. and R. Bermolen, "No TitleSupport vector regression for link load prediction," *Comput. Networks*, vol. 53, 2009.

[34] A. G. and H. Wilson, "Deep kernel learning," 2016, pp. 370--378.

[35] B. and X. Chen, "Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering},"  *IEEE Trans. Signal Process.*, vol. 65, 2017.