IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 October 2024, accepted 28 October 2024, date of publication 6 November 2024, date of current version 18 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3492706

RESEARCH ARTICLE

Received Signal Strength Indicator Prediction for Mesh Networks in a Real Urban Environment Using Machine Learning

MARLON JESKE[1], BRUNILDE SANSÒ[2], (Senior Member, IEEE), DANIEL ALOISE[2], AND MARIÁ C. V. NASCIMENTO[1]

[1]Aeronautics Institute of Technology, São José dos Campos, SP 12228-900, Brazil
[2]Polytechnique Montréal, Montreal, QC H3T 1J4, Canada

Corresponding author: Marlon Jeske (marlonjeske03@gmail.com)

This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under Grant 309385/2021-0, Grant 403735/2021-1, and Grant 142311/2019-7; in part by São Paulo Research Foundation (FAPESP) under Grant 2022/05803-3 and Grant 2013/07375-0; and in part by the Brazilian Federal Agency for Support and Evaluation (CAPES) under Finance Code 001.

ABSTRACT Mesh networks are self-managing wireless systems with dynamic topology. These networks differ from broadcast and mobile networks because their mesh nodes can directly exchange information without the intervention of any other infrastructure. However, the radio propagation environment in urban regions, characterized by dense building clusters and human-made structures, influences signal attenuation and path loss. Therefore, deploying these networks brings distinct challenges from the more intensively studied indoor or rural scenarios. In line with this, predicting radio signal propagation attenuation is crucial for planning and deploying reliable networks. The literature on received signal strength indicator (RSSI) prediction for mesh networks in urban areas is scarce. This paper proposes machine learning-based RSSI prediction models for highly urbanized areas. We highlight the most influential features, including the distance between the transmitter and receiver, obstruction details in the first Fresnel zone, and terrain variability measures. Considering data from two mesh networks in the Metropolitan Region of São Paulo, Brazil, owned by a power utility company, we trained a Random Forest and a Support Vector Regression model for the RSSI prediction task. Comparative analysis indicates an improvement of up to 66% in the RSSI prediction error using the Random Forest approach in comparison with classical and empirical models.

INDEX TERMS Feature importance, machine learning, mesh networks, network planning, RSSI prediction.

I. INTRODUCTION

Mesh networks are wireless networks with a dynamic topology that changes based on network and ambient conditions. Compared to conventional wireless networks, mesh networks require low installation and maintenance costs. These networks are increasingly present in our daily lives with applications in home, corporate, and metropolitan environments [1].

The radio propagation environment and the technological parameters of the devices are essential to define the network connectivity. In line with this, the terrain and its obstacles largely influence the network's functioning under any other external conditions. As such, evaluating the transmission path allows us to identify different propagation mechanisms that cause signal attenuation, such as signal reflections or diffractions. Particularly in urban areas, the obstructions in the radio wave propagation path are occasioned mainly by manufactured obstacles such as a region of dense and tall buildings, a mixed area of houses and structures, and residential or industrial areas.

Estimating the point-to-point signal strength considering atemporal variables is fundamental to the planning of mesh networks. There are classical and empirical approaches to measure the attenuation of the radio signal occasioned by these several obstructions in the propagation path. Classical models are derived from electromagnetic theory, whereas empirical models are based on field measurements of the

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

received signal strength indicator (RSSI). For example, the widely employed classical model Friis Equation [2] considers parameters like the transmission power and the gains of transmitting and receiving antennas to predict RSSI. However, these models do not include the geographic and environmental data that compose the radio propagation environment.

Empirical models demonstrate high computational efficiency in prediction due to the simplicity of their mathematical equations representing the statistical relationships of measurements obtained in field tests [3], [4]. Moreover, these models consider a few network parameters, such as the distance from the transmitter to receiver antennas, height of antennas, and frequency. The drawback of these models is that they were developed in specific regions that may have particularities in their geography and environment, such as Japan for Okumura-Hata [5], [6] and the United States for Egli [7]. Therefore, such models might not be efficient and accurate when applied in scenarios different from those for which they were developed in [8].

According to [9], machine learning (ML) strategies, primarily supervised regression methods, have shown promising results in path loss prediction, proving to be an alternative to traditional models. For indoor localization of objects, Guidara et al. [10] presented an RSSI predictor based on deep learning to estimate the distance between the transmitter and receiver. In their model, ambient conditions, such as humidity, temperature and noise were instrumental features to improve the model's accuracy. The authors highlight the importance of this investigation for indoor applications, for which Global Positioning System (GPS) satellite signals fail.

Numerous ML approaches have been applied to predict path loss in outdoor scenarios, primarily in urban and suburban areas [4]. The path loss prediction through ML approaches is mainly concerned with data-driven techniques to learn the relationships between the characteristics in the radio propagation environment (input) and the RSSI (output). The learning process occurs through training data composed of real-world RSSI measurements or simulation tools. The inputs of these models, referred to as features, are based on the parameters used in traditional models and on details observed in each particular scenario where the network is located, such as topographic data. Moreover, in the literature, the prediction models based on machine learning have demonstrated superior accuracy over classical and empirical models [8], [11], [12], [13], [14], including International Telecommunication Union - Radiocommunication Sector (ITU-R). In particular, the studies presented in [3] and [15] showed that the ML approaches outperformed the ITU-R P.452 recommendation [16]. Also, the ML model proposed in [17] achieved better prediction results than ITU-R P.1546 [18].

Despite the several studies, the literature on ML approaches to predict RSSI in urban and suburban regions has been mostly developed for cellular and broadcast networks [4]. Notably, there have been applications in radio [19] and television [17] systems, and in cellular networks from 2G to 5G [9], [11], [20], [21]. Wireless mesh networks, on the other hand, differ significantly from broadcast and mobile networks because their mesh nodes can directly exchange information without the intervention of any other infrastructure. They are often located in existing poles that limit their coverage conditions and operate at different frequency ranges.

More specifically, the literature on machine learning approaches dedicated to path loss prediction in mesh networks is very scarce. To our knowledge, there is only one ML model, but it focuses on mountainous areas [12]. However, the radio propagation environment in urban regions, characterized by dense building clusters and human-made structures, presents distinct challenges and complexities compared to mountainous areas, where rugged terrains, elevation changes, and natural obstructions predominantly influence signal attenuation and path loss. Indeed, the authors in [12] highlight that crucial features of their scenario, such as tree canopy coverage, are not necessarily relevant for other applications and that the developed model only applies to the environment in which it was created. Thus, the criteria to define the appropriate features that describe the region where the network is located must be carefully chosen to emphasize the characteristics regarding the propagation environment. Additionally, as suggested in [9], disregarding pertinent features or retaining unrelated ones can result in an inaccurate predictor.

The objective of this paper is to propose a comprehensive ML approach to predict the RSSI for urban mesh networks to support its planning and design. In fact, our RSSI prediction values enable the construction of a middle-layer propagation grid, such as proposed in [22], for future what-if studies from service providers. The introduced methodology comprises the selection of features, the determination of appropriate ML algorithms, and the strategies to train and validate the model. Furthermore, we provide a performance comparison between the suggested ML approach and both classical and urban empirical models. Lastly, we highlight a feature importance analysis, indicating which features most significantly contribute to the machine learning process. Performing the feature importance enables understanding the impact and relationship of the features representing the radio propagation environment with the RSSI.

To validate our proposal, a case study was created by using real data extracted from two urban mesh networks located in the Metropolitan Region of São Paulo, Brazil. We gathered, among others, a dataset of 1,117 mesh links (different transmitter-receiver pairs), terrain profile data, the percentage of obstruction in the first Fresnel zone, and the distance between the devices. As far as we know, this is a novel proposal of an ML approach for RSSI prediction in urban wireless mesh networks.

The primary contributions of this paper are summarized next.

- We introduce a novel approach based on machine learning to predict the RSSI for urban mesh networks;
- We present a detailed analysis of feature importance on urban mesh networks;
- The introduced approach outperforms empirical models regarding prediction errors.

The rest of this paper is organized as follows. Section II presents a literature review on machine learning for RSSI prediction in urban scenarios. Section III shows the classical and empirical path loss prediction models used in this investigation. Section IV presents the machine learning approach proposed to predict the RSSI. Section V reports the results, feature importance analysis, and discussion. Finally, Section VI presents the conclusions and future work directions.

II. LITERATURE REVIEW

This section presents a literature review on ML for the RSSI prediction in different environments and networks.

A. RSSI AND PATH LOSS PREDICTION IN URBAN AREAS

This section gives an overview of machine learning approaches for predicting received signal strength indicator and path loss in urban scenarios. We also suggest a comparative analysis of recent literature on ML approaches contrasted with other algorithms, including ITU-R recommendations and empirical models.

1) MACHINE LEARNING FOR RSSI PREDICTION

In urban and suburban areas, obstructions in radio wave propagation are mainly caused by manufactured obstacles like dense high-rises, mixed residential or industrial areas, and houses and buildings. In rural or forest areas, obstacles are primarily due to terrain complexity, vegetation coverage, tree density, and seasonality [12], [23]. Several ML approaches for RSSI prediction have been proposed, including the use of Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Random Forest. In studies such as [8], [9], and [20], ANN outperformed empirical models like Egli, ECC-33, COST-231, and Okumura-Hata, showing superior prediction accuracy. In [9], the authors demonstrated that Random Forest, SVR, and ANN outperformed the classical log-distance path loss model in predicting RSSI for mobile networks. Studies like [11], [13], [15], [17], [19], [21], [24], [25], [26], and [27] have explored various ML approaches, including Random Forest, Elastic-Net Regression, Adaptive Boosting (AdaBoost), and ANN, showcasing improved performance compared to traditional models and recommendations.

2) COMPARATIVE ANALYSIS AND CONTRAST WITH OTHER ALGORITHMS

Studies such as [11], [13], [15], [17], [19], [21], [24], [25], [26], and [27] conducted comparative analyses against other algorithms, including ITU-R recommendations and empirical models. ML approaches consistently outperformed traditional models like COST-231, ITU-R recommendations, log-linear regression, two-ray models, and more, as demonstrated by reduced root mean squared error (RMSE) values. In [15] and [17], the authors specifically compared ML approaches to ITU-R recommendations, with ML models, particularly ANN, consistently outperforming the literature algorithms.

3) LACK OF STUDIES IN URBAN MESH NETWORKS

Even though the machine learning approaches proposed in the studies previously described are for networks located in urban areas, to our knowledge, there are no similar proposals for urban wireless mesh networks. This is despite the fact that with the growing number of wireless services with high performance demands, new type of architectures, such as mesh architectures are going to be deployed. Therefore, as indicated by [4], new propagation models are needed. Thus, ML is a useful tool to address these demands.

That is why the primary contribution of this paper is to fill such a gap. The next subsection details the only work in the literature that deals with mesh networks, even though it is not in an urban setting.

B. RSSI PREDICTION IN MESH NETWORKS

Traditional mobile and broadcast networks have a well-defined structure with a base station (transmitter) strategically placed to cover a specific area where user equipment (receivers) is located. This strategic placement ensures optimal coverage. In contrast, mesh networks, while also planned for coverage, face limitations due to the specific locations of nodes, such as light poles or buses. Additionally, the antennas used in mesh networks differ from those in cellular or broadcast networks. These differences significantly impact the evaluation of RSSI.

Despite these challenges, there is limited research on using machine learning for RSSI prediction in mesh networks. Only the work by [12] addresses this gap, focusing on mesh networks in mountainous regions with distinct radio propagation characteristics compared to urban areas. In their study, the authors in [12] utilized a dataset comprising 2,218 links from mesh networks at the America River Hydrologic Observatory, USA. These networks monitored environmental indicators like soil temperature, snow depth, and air temperature. The ML approach incorporated four algorithms: Random Forest, AdaBoost, ANN, and K-nearest Neighbors (KNN). The model input featured seven characteristics tailored to mountainous areas, including transmitter-receiver distance, average tree canopy coverage, terrain and vegetation standard deviation, angle between line of sight (LOS) and horizontal plane, and canopy coverage at transmitter and receiver locations. Beyond ML algorithms, the authors incorporated empirical models for forested environments and models accounting for varying vegetation (seasonality). These included Weissberger's modified exponential decay model, ITU-R recommendation, COST-235, as well as models based on Friis Equation and the Plane Earth

TABLE 1. A summary of the related literature.

Network	Reference	RoI	Data	Main Features	Number of Features	Algorithms	Feature Importance
Radio and TV	[8]	Urban	Real	Distance, Rx height, elevation	3	ANN, Adaptive Neural Fuzzy Inference System (ANFIS) and Kriging	No
	[13]	Urban	Real	Distance, elevation, latitude, and longitude	4	ANN	No
	[17]	Mixed city-river	Real	Distance, distance over the river, elevation, and Fresnel zone radius	5	ANN and ANFIS	No
	[19]	Rural with small scale industries	Real	Distance	1	AdaBoost, ANN, and SVR	No
Mobile (3G to 5G)	[9]	Urban	Real	Distance	1	ANN, SVR, and Random Forest	No
	[11]	Urban, suburban, and open areas	Real	Distance, antenna's angles, enviroment type, and frequency	7	Random Forest	Yes
	[15]	Urban	Simulated	Distance, antenna's angles, LOS status, and clutter type	13	ANN, KNN, Decision Tree, and Linear Regression	No
	[20]	Urban, suburban, and open areas	Real	Distance, Tx and Rx heights, frequency, and Rx elevation	5	ANN	No
	[21]	Street canyons	Real	Distance, LOS distance, clutter type, and street width	6	Lasso Regression, Elastic Net, Random Forest and SVR	Yes
	[24]	Urban	Real and simulated	Distance, Tx height, building density, and average building width	8	ANN	No
	[25]	Urban	Real	Distance, antenna's angles, frequency, and Tx and Rx coordinates	10	ANN	No
	[26]	Suburban	Real	Distance, Tx and Rx heights, and Tx and Rx height difference	4	ANN	No
	[27]	Urban, suburban, and rural	Real	Distance, elevation, altitude, and clutter height	5	ANN and SVR	No
Mesh	[12]	Mountainous region	Real	Distance, LOS angle, tree canopy coverage, and canopy coverage at Tx and Rx	7	AdaBoost, ANN, KNN, and Random Forest	Yes
	This work	Urban	Real	Distance, Tx and Rx heights, elevation, and percentage of obstruction in the Fresnel zone	8	Random Forest and SVR	Yes

model. Results indicated that the Random Forest algorithm outperformed other ML algorithms (KNN, ANN, AdaBoost) and empirical models, reducing the average prediction error by 37%. In summary, the investigation in [12] demonstrated the effectiveness of ML, specifically Random Forest, for RSSI prediction in challenging terrain, shedding light on the importance of specific features in the learning process.

Table 1 shows a summary of the literature review discussed in this section.

III. CLASSICAL AND EMPIRICAL MODELS

The first reported prediction models were derived from electromagnetic theory, which began in the 40s with the significant contribution of the pioneering work of Harald Trap Friis [2]. With the evolution and worldwide expansion of VHF networks (broadcasting and TV) in the 60s and 70s, other models for predicting the received signal strength were proposed and developed considering measurements performed in cities such as New York and Tokyo. The classical and empirical models for path loss prediction, despite being dated, are still employed in recent studies as the primary approach to solving grid planning problems for networks [28] and in comparative analysis of the results with ML algorithms [29].

In this paper, we chose empirical models considering the characteristics of the urban region where the mesh networks are located and the carrier frequency of 920 MHz. Thus, the selected empirical models were those proposed by Egli, Edwards-Durkin [30], and the Okumura-Hata. We also consider in our study the classical models derived from electromagnetic theory. More specifically, the Friis Equation [2], the Free Space Path Loss (FSPL), and the variant of FSPL that considers reflection (FSPL-R) [31].

The following sections briefly present the classical and empirical models largely employed to evaluate the received signal strength.

A. FREE SPACE PATH LOSS

The FSPL is the loss of signal strength caused by the natural propagation of radio waves, often referred to as beam divergence [32]. The radio frequency signal power spreads over large areas as the signal propagates from an antenna, and, as a result, the signal strength is attenuated. The FSPL can be calculated, in dBm, as in

$$\text{FSPL} = 20\log_{10}\left(\frac{\lambda}{4\pi d}\right), \tag{1}$$

where d is the distance between the transmitter and receiver, and λ is the wavelength. As stated by [31], it is possible to consider reflections in the propagation path, and it can be calculated using

$$\text{FSPL-R} = \min\left\{10\log_{10}\left[\frac{(h_{Tx}h_{Rx})^2}{d^4}\right]; \text{FSPL}\right\}, \tag{2}$$

where h_{Tx} and h_{Rx} are the height of the transmitter and receiver in meters, respectively. Also, the FSPL is obtained by (1).

B. FRIIS

The Friis Equation, developed by Harald Trap Friis, is a fundamental equation based on electromagnetic theory. This equation is used to calculate the received power level of the radio signal propagated from transmitter to receiver, considering the transmitted power, the gain of the transmitting and receiving antennas, the wavelength of the signal, and the distance between the transmitter and receiver. Also, this equation was based on the environment that does not consider any obstacles or interference along the path between the antennas, generally referred to as free space condition.

The Friis Equation can be defined as in

$$\text{Friis} = P_{Tx}G_{Tx}G_{Rx}\left(\frac{\lambda}{4\pi d}\right)^2, \tag{3}$$

where P_{Tx} and G_{Tx} are the transmission power and transmitter antenna gain, respectively; G_{Rx} is the receiver antenna gain; and the term in-between parentheses refers to FSPL. The Friis Equation can be calculated in dBm and is defined by applying the log function as in

$$\text{Friis} = P_{Tx} + G_{Tx} + G_{Rx} + \left[20\log_{10}\left(\frac{\lambda}{4\pi d}\right)\right]. \tag{4}$$

C. EGLI

In [7], Egli performed several measurements on irregular terrain, mainly in New York City, using systems with frequencies between 40 MHz and 1000 MHz. Later, in [33], the authors proposed a mathematical expression for calculating the received signal strength based on Egli results. The Egli model can be defined as in

$$\text{Egli} = 20\log_{10}(f) + 40\log_{10}(d) - 20\log_{10}(h_{Tx}) + k, \tag{5}$$

where f is the frequency in MHz; h_{Tx} and h_{Rx} are the heights in meters of the transmitting and receiving antenna, respectively; and d is the distance between the transmitter and receiver in kilometers. The term k refers to the height of the receiver and is defined as:

$$k = \begin{cases} 76.3 - 10\log_{10}(h_{Rx}), & \text{if } h_{Rx} \leq 10 \\ 85.9 - 20\log_{10}(h_{Rx}), & \text{if } h_{Rx} > 10 \end{cases}. \tag{6}$$

D. EDWARDS-DURKIN

The Edwards-Durkin model [30] was obtained from measurements carried out in the United Kingdom by Durkin [34]. Later, with these measurements, Edwards and Durkin used the FSPL equation to propose a correction factor due to propagation loss, as in

$$\begin{aligned}\text{Edwards-Durkin} = {} & 118.7 - 20\log_{10}(h_{Tx}) \\ & - 20\log_{10}(h_{Rx}) + 40\log_{10}(d),\end{aligned} \tag{7}$$

where h_{Tx} and h_{Rx} are the heights in meters of the transmitting and receiving antenna, and d is the distance between the transmitter and receiver in kilometers.

E. OKUMURA-HATA

The Okumura-Hata model was developed based on measurements taken by Okumura in Tokyo city [5], using systems with frequencies ranging from 150 MHz to 2000 MHz [35]. Later, Hata [6] refined the model based on the results obtained by Okumura. Hata introduced three mathematical expressions to represent different environments where the networks are located. Additionally, the author provided a correction factor for small to medium-sized cities and another for large cities. The Okumura-Hata model has become widely known and is one of the most referenced models in the literature [36].

In particular, we used the correction factor for large cities and the equation for urban regions. The correction factor (α) is calculated as in

$$\alpha = 3.2[\log_{10}(11.75\, h_{Rx})]^2 - 4.97, \tag{8}$$

where h_{Rx} is the height in meters of the receiver antenna. The Okumura-Hata Equation is calculated as follows [31]:

$$\begin{aligned}\text{Okumura-Hata} = {} & 69.55 + 26.16\log_{10}(f) \\ & - 13.82\log_{10}(h_{Tx}) - \alpha \\ & + [44.90 - 6.55\log_{10}(h_{Tx})]\log_{10}(d),\end{aligned} \tag{9}$$

where f is the carrier frequency in MHz, h_{Tx} is the height in meters of the transmitting antenna, and d is the distance in kilometers from the transmitter to the receiver.

IV. MACHINE LEARNING APPROACH

This section introduces the proposed method for predicting RSSI in urban wireless mesh networks. We provide an overview of the employed machine learning algorithms, details regarding the two mesh networks and their respective urban settings, strategies for data collection, and the comprehensive data processing steps leading to the final

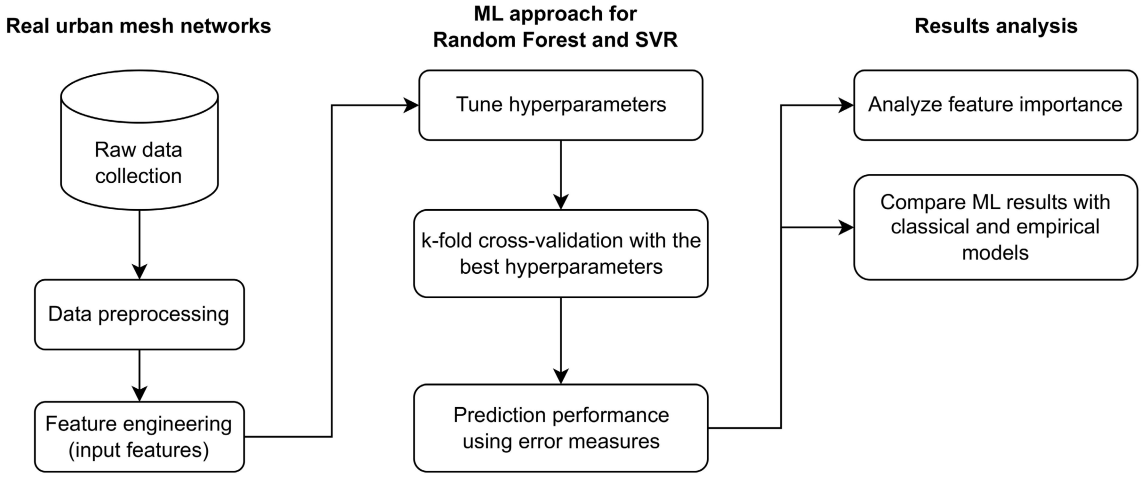


FIGURE 1. Workflow of the machine learning approaches.

RSSI dataset. Additionally, we discuss the features utilized as input for machine learning algorithms, the error metrics employed to assess algorithm performance, the process of hyperparameter tuning, and the methodology for training and validating the algorithms. Fig. 1 illustrates the workflow of the ML approach proposed in this paper.

The following sections describe the steps of the workflow in Fig. 1. Section IV-B covers the Raw Data Collection and Data Processing stages. Feature Engineering, as depicted in the figure, is detailed in Section IV-C. Section IV-D discusses the ML approach, including hyperparameter tuning, training and testing strategies, and performance evaluation using error measures. Section V presents a comparative analysis between the results obtained by the proposed ML approach and the classical and empirical models. Finally, Section V-A reports the feature importance analysis in the prediction process.

A. MACHINE LEARNING ALGORITHMS

According to [12], empirical and classical models face limitations when dealing with the diverse scenarios encountered in measuring received signal strength. In light of this, our investigation focuses on machine learning-based propagation models that have demonstrated promise in recent related literature [4], [25], [37], [38].

Roughly, the RSSI prediction through machine learning algorithms involves training a regression model. For this model, the target variable (output) is the RSSI value, and the features (input) may include the distance from the transmitter to the receiver, the heights of the transmitter and receiver antennas, and other values that represent the radio environment propagation or antenna parameters.

Building on the extensive review provided in [9], we find that the Support Vector Regression and Random Forest algorithms have exhibited strong performance in predicting received signal strength.

1) RANDOM FOREST

The Random Forest [39] is an ensemble technique that effectively addresses regression problems by combining multiple decision trees, leading to accurate predictions. In the context of regression, where the goal is to predict continuous numerical values, each decision tree in the Random Forest considers input features to offer predictions, averaged to obtain the final output.

Decision trees are non-parametric supervised learning algorithms with a hierarchical structure that starts from the root node (with no incoming branchings) to the bottom levels. Internal nodes are non-terminal nodes, also known as decision nodes, from which branches depart to two (or more) children nodes. Internal nodes contain a split condition formulated from a subset of features that divides the training set into two (or more) subsets. The split aims to find the best way to separate the training set according to the rule formulated with the selected feature. Leaf or terminal nodes represent the final predictions of the tree. The random forest enhances the robustness of the decision tree algorithm by combining multiple uncorrelated decision trees.

To build a decision tree in the Random Forest algorithm, a subset must be obtained from the original training set. This subset is obtained through the bootstrap method, that is, by resampling with replacement from the original set. In the first step, a feature must be chosen from this subset to perform the first split, to define the decision rule of the root node. For example, suppose that the feature antenna height is under evaluation to guide such a splitting. The training set is sorted in increasing order of the antenna height values and the pairwise mean of consecutive values of this sorted feature considered to define the splitting value. The splitting value is the value that would define the branches from this node, where the left branch would correspond to the samples whose antenna height is lower than or equal to this threshold, whereas the right branch would contain the data greater than

it. The chosen threshold of a given feature is the pairwise mean value with the minimum mean squared error (MSE). The MSE is calculated considering the prediction of a given input (from the training set) the average expected output of all terminal nodes from the branch to which the rule it applies. Besides, the feature picked for the splitting is the one with the lowest MSE among all features in the subset. After selecting the feature, the leaf nodes are evaluated for splitting (following the breadth-first search order). Therefore, the decision question of the root and decision nodes is based on the feature from a new sampled set of features that better splits the training data in the given iteration. The process stops when a stopping criterion is reached. In general, the stopping criteria are hyperparameters of the Random Forest, such as the maximum depth of the tree, the minimum number of samples to perform a split, and others. In the proposed RF the stopping criterion is the minimum number of samples in the node for splitting, which is 5. The tuning of this and other hyperparameters are better discussed in Section IV-D2.

After reaching a stopping criteria, the final prediction value, $\hat{y}$, of the Random Forest is obtained by averaging the prediction value y_i obtained in each tree i of the n_T trees, as in:

$$\hat{y} = \frac{1}{n_T}\sum_{i=1}^{n_T} y_i. \tag{10}$$

The preference of the Random Forest algorithm for predicting RSSI was based upon different motivations. The primary reason is that the Random Forest algorithm was employed in the existing literature concerning the prediction of RSSI in wireless mesh networks, proposed by [12]. The authors showed that the best prediction results were achieved through this approach when compared with alternative well-established machine learning algorithms such as ANN, KNN, and AdaBoost. Moreover, additional support for the advantage of the Random Forest algorithm in RSSI prediction is drawn from recent investigations presented by [9] and [11], wherein it was evidenced that the Random Forest algorithm achieved significant accuracy.

2) SUPPORT VECTOR REGRESSION

The SVR [40] is a technique that extends the principles of Support Vector Machines (SVM) [41] from classification to regression problems. The SVR can be formulated as an optimization problem.

Let a training set be represented by $\{(x_i, y_i)\}$ such that i is the index of the samples ranging from 1 to n. Let $x_i \in \mathbb{R}^m$ be the input feature vector of the i-th sample and $y_i \in \mathbb{R}$ the corresponding target value. We aim to find a function $f(x_i)$ that has at most ϵ deviation from the actual values y_i, ensuring that $f(x_i)$ is as smooth as possible. The ϵ deviation is one of the hyperparameters of the SVR. The function $f(x_i)$ can be defined as:

$$f(x_i) = \langle w, x_i\rangle + b_i, \tag{11}$$

where $\langle w, x_i\rangle$ is the dot product between the weight vector w and the feature vector x_i, and b_i represents the bias. To ensure that $f(x_i)$ is as smooth as possible for every i, the norm of the weight vector must be minimized. This can be defined as an optimization problem as follows:

$$\min \quad \frac{1}{2}\|w\|^2 \tag{12}$$

$$\text{subject to} \quad |y_i - f(x_i)| \le \epsilon \quad \forall i = 1, \ldots, n. \tag{13}$$

To avoid being too punitive and to handle cases where $f(x_i)$ might not exist, we can introduce slack variables ξ_i and ξ_i^* and add a constant C to determine the trade-off between the smoothness and the deviations greater than ϵ. The constant C is also a hyperparameter of the SVR. The new formulation of the minimization problem can be described as follows:

$$\min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{14}$$

$$\text{subject to} \quad y_i - f(x_i) \le \epsilon + \xi_i \quad \forall i = 1, \ldots, n \tag{15}$$

$$f(x_i) - y_i \le \epsilon + \xi_i^* \quad \forall i = 1, \ldots, n \tag{16}$$

$$\xi_i, \xi_i^* \ge 0, \quad \forall i = 1, \ldots, n. \tag{17}$$

One way to solve the above optimization problem is by transforming the original formulation using the Lagrangian formulation with the Lagrange multipliers α_i, α_i^*, η_i, and η_i^*. Thus, the original problem can be formulated as follows:

$$\begin{aligned} L = \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{n}\alpha_i[\epsilon + \xi_i - y_i + f(x_i)] \\ & - \sum_{i=1}^{n}\alpha_i^*[\epsilon + \xi_i^* + y_i - f(x_i)] \\ & - \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*). \end{aligned} \tag{18}$$

To find the optimal Lagrange multipliers α_i and α_i^*, we can obtain the dual problem, which can be represented as follows:

$$\begin{aligned} \max \quad & \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) - \epsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) \\ & - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \end{aligned} \tag{19}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C \quad \forall i = 1, 2, \ldots, n \tag{20}$$

$$0 \le \alpha_i^* \le C \quad \forall i = 1, 2, \ldots, n \tag{21}$$

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0. \tag{22}$$

After obtaining the optimal solution for the dual problem, that is, finding the optimal values for α_i and α_i^*, the final prediction function of the SVR is defined as follows:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x_i, x\rangle + b. \tag{23}$$

Finally, the SVR can use kernel functions to map the input features properly. The kernel function, denoted as $K(x_i, x_j)$, replaces the dot product $\langle x_i, x_j \rangle$. There are various kernel functions; one way to choose one is by testing different functions on the same training set. One well-known kernel is the Radial Basis Function (RBF), which replaces the traditional dot product as follows:

$$K(x_i, x_j) = e^{(-\gamma \|x_i - x_j\|^2)}, \tag{24}$$

where γ is a parameter of the RBF. After defining the kernel function, it can be replaced in the final prediction function of the SVR as follows:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b. \tag{25}$$

Similar to the reasons for selecting the Random Forest algorithm, Support Vector Regression has been successfully employed in the literature for path loss prediction. In [27], the authors demonstrated that SVR exhibited superior performance compared to outcomes achieved by ANN.

B. REAL-WORLD DATASET

In this section, we explain the characteristics of the collected RSSI data, describe the raw data, data preparation, and data cleaning. The data preprocessing and manipulation were conducted using the *dplyr* package [42].

1) MESH NETWORKS IN URBAN ENVIRONMENT

To develop the machine learning approach to predict the RSSI, we collected data from two mesh networks, BVI and SLU, located in a dense urban area in the Metropolitan Region of São Paulo, Brazil. The region is characterized by a mix of residential houses, commercial buildings, and industrial factories.

These mesh networks are used to automate and monitor a set of reclosers installed in utility poles from an overhead electric power distribution system. The reclosers are automatic devices that verify the existence of overcurrent and promptly restore power to the line. According to [43], about 70% of overhead electric distribution system faults are temporary. Therefore, using reclosers in such types of systems is crucial to ensure that temporary faults or disturbances are swiftly addressed, minimizing disruptions in the power supply and enhancing the overall reliability of the electrical grid.

By integrating reclosers with mesh networks in electrical distribution systems, real-time monitoring becomes achievable, making it possible to identify and rapidly respond to faults. This integrated system enables centralized management that supports strategic decision-making and data analysis, while direct communication between the reclosers ensures autonomy and automation. As a result, there is a significant decrease in the need for manual interventions, reducing human efforts and saving time and resources for electrical distribution companies.

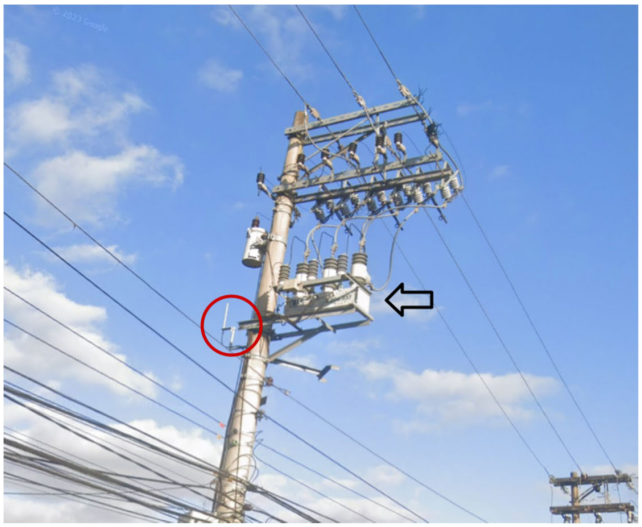


FIGURE 2. A utility pole in the overhead electric distribution system with a recloser (black arrow) and a mesh node with an omnidirectional antenna (red circle).

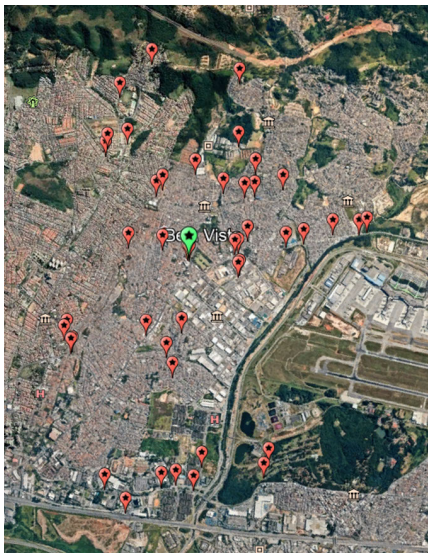
In particular, the structure of BVI and SLU mesh networks for the reclosers automation is composed of two types of devices, the mesh nodes and concentrator. The mesh node is installed in each recloser and is capable of sensing, transmitting, and receiving data. The concentrator is unique for each mesh network and acts similarly as a gateway, responsible for receiving all data routed in the network and then transmitting it to an external server. In Fig. 2, we show a mesh node installed in a recloser in its respective utility pole from a network used in this paper.

The BVI and SLU mesh networks are composed of 45 and 31 mesh nodes (reclosers), respectively. All mesh nodes and concentrators from BVI and SLU have the same technical specifications. Each device has an omnidirectional antenna with the transmission power equal to 24 dBm, antenna gain of 6 dBi, operating at 922 MHz, and the radio sensitivity level is equal to −120 dBm. Additionally, the mesh nodes are installed on utility poles at a height of 5 meters above the ground, close to their respective reclosers, as shown in Fig. 2. The concentrator's antenna is positioned at a height of 25 meters above the ground on a dedicated tower.

In Fig. 3, the aerial views of the device distribution of the BVI (Fig. 3a) and SLU (Fig. 3b) networks are presented.

In BVI and SLU mesh networks, after each mesh node performs its sensing, the gathered data should be immediately forwarded to the concentrator. Each mesh node can act both as a transmitter and a receiver. The data transmission can occur via a single hop, where the mesh node sends data directly to the concentrator, or via multi-hop, where data is sent through intermediate mesh nodes before reaching the concentrator. The path to successfully send data from a mesh node to the concentrator depends on the signal quality between the devices, referred to as link quality.

A link in the network is formed between a transmitter, which is exclusively a mesh node, and a receiver, which can



(a) BVI network

(b) SLU network

FIGURE 3. The aerial views of the BVI and SLU mesh networks located in the Metropolitan Region of São Paulo. The red markers represent the mesh nodes, and the green markers represent the concentrator.

be either another mesh node or the concentrator. The link exists if the receiving device is in the neighborhood of a specific mesh node. The neighbor nodes are those devices that receive the minimal signal quality required, defined by the radio sensitivity level, to establish communication.

2) DATA GATHERING

We obtained data from the networks described in the previous section through the Supervisory Control and Data Acquisition (SCADA) system to develop the machine learning approach. In particular, every 5 minutes, each network node identifies available neighbor nodes to transmit its data. Each node then sends a data package containing information about all its neighbors to the concentrator. Thus, through the SCADA system, we obtained a log file from the concentrator containing the list of neighbors from each mesh node.

Due to the complex urban propagation environment where the mesh networks are located, the list of neighbors sent by each node can change throughout the day. Typically, neighbors with weak RSSI tend to appear and disappear from the neighbor list. As a result, a specific neighbor node might

```
189 2020-09-16T00:07:08:176 NBR_STATS: 3038353031002a0080003b01b5003c013038
mac 30:38:35:30:31:00:2a:00; etx: 128; rank: 315; rssi: 181; last-tx: 316
mac 30:38:35:30:30:00:1c:00; etx: 367; rank: 475; rssi: 88; last-tx: 93
mac 30:38:35:30:31:00:36:00; etx: 179; rank: 275; rssi: 96; last-tx: 2079
mac 30:38:35:30:30:00:3b:00; etx: 139; rank: 307; rssi: 130; last-tx: 41
```

FIGURE 4. Raw log files from the SCADA system.

or might not consistently appear in a node's neighbor list across all the transmitted log files. Therefore, considering these variations in the measured RSSI, we conducted an 18-day consecutive data collection campaign to measure the RSSI for each link reliably.

Fig. 4 illustrates an example of a log file received from the concentrator via the SCADA system.

The figure shows an example of information regarding a particular mesh node and its neighbors. The number 189 serves as an identifier for the packet that the node sent to the concentrator. The date and time this packet was dispatched are described as 2020-09-16T00:07:08:176, that is, on September 16, 2020, at 00:07:08:176 (hours, minutes, seconds, milliseconds), the concentrator received this packet. The "NBR STATS" means neighbor status and represents the encoded information derived from the neighbor list in the packet.

Each neighbor node in the network is identified by its Media Access Control (MAC) address. In this example, the MAC address of the first neighbor is 30:38:35:30:31:00:2a:00. Additionally, every neighbor has metric values that represent its status and the quality of the link. These metrics are:

- Expected Transmission Count (ETX): measures how many transmission attempts are expected to deliver a packet over that link successfully;
- Rank: represents the neighbor's distance relative to the concentrator;
- Received signal strength indicator: represents the RSSI from the neighbor and the receiver;
- Last transmission (last-tx): provides information about the time, in seconds, when the last packet was sent to the neighbor.

Among these metrics, only the RSSI value is pertinent for the introduced machine learning approach. On the other hand, the ETX, rank, and last-tx values are employed to determine routing rules according to the defined routing protocol.

In addition to these log files, the energy company provides pertinent details about each mesh network. This file contains a label of the utility pole where a recloser is situated, MAC address, port number, and the respective geographical coordinates (latitude and longitude). Table 2 shows an example of this file.

TABLE 2. Complementary data for the reclosers.

Label	MAC address	Port	Latitude	Longitude
564312	30:38:35:30:31:00:2a:00	10044	-23.409	-46.500
563865	30:38:35:30:30:00:1c:00	10008	-23.418	-46.465
503171	30:38:35:30:31:00:3b:00	10073	-23.413	-46.491

In the first line of Table 2, a utility pole is identified by the number 564312, which has an associated mesh node with a MAC address of 30:38:35:30:31:00:2a:00. The port number for this node is 10044, and its coordinates are −23.409, −46.500.

3) DATA PROCESSING

In the first step of data processing, we extracted the MAC address to identify the neighbor nodes and their respective RSSI values from each node's log files sent to the concentrator. Since the information in the log file is encoded, the following procedure is required to obtain the RSSI value in dBm units:

$$\text{RSSI (dBm)} = \left(\frac{\text{RSSI}}{2}\right) - 130. \tag{26}$$

For instance, if the RSSI value is 181 in the log file, by using (26), we obtained an RSSI equal to −40 dBm.

As mentioned earlier, we recorded information from all packets sent by each mesh node to the concentrator every 5 minutes for 18 consecutive days. As the RSSI measured for a particular link can vary throughout this period, we calculated the median of the RSSI values, in dBm, to determine the definitive RSSI of each existing link. The median was chosen because it is less susceptible to extreme values and provides a more centralized representation of the RSSI distribution without being influenced by potential outliers.

After calculating the RSSI medians, we combine the files by the MAC address of the respective transmitter and receiver for each link. We obtained 662 links from the BVI network and 455 links from the SLU network. Then, we combine both data in a unique dataset comprising 1,117 RSSI measures. Finally, in the definitive RSSI dataset, we have the labels and geographic coordinates of the transmitter and receiver and the median of the RSSI for each link. Table 3 presents examples of links and their respective information.

TABLE 3. Definitive RSSI dataset.

Label	Value
Tx label	550135
Rx label	509413
Tx latitude	-23.430
Tx longitude	-46.500
Rx latitude	-23.422
Rx longitude	-46.515
Distance (m)	183.66
RSSI (dBm)	-65

C. FEATURE ENGINEERING

In this section, we elaborate on the features utilized in the proposed machine learning approach. The selection of appropriate input features is pivotal in machine learning, as the performance of the model hinges on the quality and relevance of the input data. For RSSI prediction, it is essential that the features distinctly capture the characteristics of the radio propagation environment.

Furthermore, in [4], the authors underscored the potential of machine learning methods in simplifying input requirements for developing path loss models in complex environments. They argued that an increased number of features (factors influencing signal attenuation) does not necessarily guarantee improved accuracy in prediction models. On the contrary, a large number of features can diminish performance and add complexity to the machine learning approach, particularly in terms of computational time and efforts required for feature extraction. Although various factors impact signal attenuation, their incorporation can make machine learning approaches more intricate, both in extracting relevant information to be used as features and in achieving optimal prediction performance.

Addressing these challenges, in [21], the authors emphasized the difficulties encountered when comparing machine learning approaches to deterministic propagation models like ray-tracing models. The latter necessitates detailed information about the environment, encompassing geometric and material properties such as topography details or material composition and their respective influences on signal propagation, including absorption loss. This detailed information significantly heightens the complexity of modeling. Notably, several studies in the literature have demonstrated that even utilizing a single feature in machine learning approaches for predicting path loss, such as employing distance as the sole feature, can yield results superior to traditional propagation prediction models [9], [19].

Given these considerations, this paper approaches feature selection with an emphasis on achieving a balance between comprehensiveness and practicality. The features were chosen based on classical and empirical models and the recent literature on machine learning for RSSI prediction in urban contexts discussed in Section II. Additionally, we explored the features proposed by [12] for predicting RSSI in mesh networks situated in mountainous regions. Despite the differences in the regions, we drew inspiration from features that could be relevant in both scenarios.

For features extraction, we use the labels of the transmitter and receiver and their respective longitude and latitude for each of the 1,117 links from the dataset detailed in Section IV-B2 and terrain data. The terrain elevation data was obtained using the *elevatr* package [44], which provides a series of repositories for geographical data. In particular, with the function *get_elev_raster*, we obtained terrain elevation data from the Amazon Web Services Terrain Tiles [45]. Therefore, combining these two data sources, we extract the following proposed features for each link.

1) EFFECTIVE HEIGHT OF THE TRANSMITTER AND RECEIVER ANTENNAS

For both the transmitter and receiver, we considered the antenna height installed at the utility pole plus the terrain elevation in its respective coordinates.

2) DISTANCE FROM THE TRANSMITTER TO THE RECEIVER

To calculate the distance for each link, we use the transmitter and receiver coordinates through the function *geodist* from the package *geodist* [46].

3) TERRAIN ELEVATION STATISTICS

A set of statistical measures concerning the terrain elevation in the path between the transmitter and receiver. Each of the following measures represents an input in our approach: the maximum and minimum terrain elevations and the mean, median, and standard deviation of the terrain elevation along the path.

4) PERCENTAGE OF OBSTRUCTION IN THE FIRST FRESNEL ZONE

Given a radio link, the space between the transmitter and receiver can be divided into a set of zones (ellipsoids) called Fresnel zones, in honor of the physicist Augustin-Jean Fresnel (1788-1827). The radius of the first Fresnel zone is calculated according to:

$$r = \sqrt{\frac{d_1 d_2 \lambda}{d_1 + d_2}}, \tag{27}$$

where λ is the wavelength, d_1 is the distance from the transmitter to the highest obstacle in the path profile, and d_2 is the distance from the highest obstacle in the path profile to the receiver.

According to [31], when the first Fresnel zone is obstructed in more than 40%, attenuation caused by obstacles will probably occur. Also, according to [25], the obstacles are the most critical components in a radio propagation environment. Therefore, including geometric patterns of radio wave propagation suggests potential relevance for RSSI prediction, specifically within urban regions as presented by [17].

The proposed machine learning approach considers as a feature the percentage of obstruction in the first Fresnel zone varying from 0% to 100%. The percentage for each link is calculated based on the highest obstruction in the path profile and compared to the diameter ($2r$) of the first zone. For example, if the highest obstacle in the path profile hits exactly in the line of sight, then the obstruction in the first Fresnel zone is equal to 50%.

The procedure used to calculate the percentage of obstruction has the following steps:

- Extract terrain elevation data between the transmitter and receiver using their heights and latitude and longitude coordinates;
- Draw a straight line between the transmitter and receiver antenna, which represents the line of sight – which will be one of the diameters of the ellipse of the Fresnel zone;
- Consider the first Fresnel zone using (27) to define the other ellipse radius.
- Calculate the obstruction percentage of the first Fresnel zone, considering the maximum altitude in the terrain profile between the transmitter and receiver. Then, calculate the radius using (27), considering d_1 the distance from the transmitter to the maximum altitude and d_2 the distance from the maximum altitude to the receiver. Finally, the obstruction percentage is obtained between the maximum altitude and the bottom of the Fresnel zone two times the radius.

Fig. 5 displays the elevation profile in meters (black curved line) to sea level rise (y-axis) and the distance from the transmitter (T) to the receiver (R) in meters (x-axis). Furthermore, the line of sight (straight dashed blue line) and the first Fresnel zone (ellipse region) are shown in the same figure.

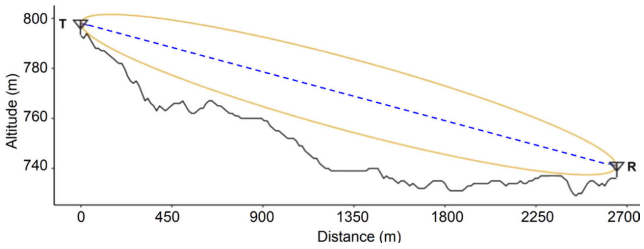


FIGURE 5. Elevation profile (black), line of sight (blue), and the first Fresnel zone (orange).

In the link example illustrated in Fig. 5, the percentage of obstruction in the first Fresnel zone is equal to 0% since there is no obstruction inside the ellipsoid.

D. LEARNING PROCESS AND PERFORMANCE METRICS

This section describes the process of defining the hyper-parameters of the Random Forest and Support Vector Regression algorithms, the strategies for training and testing the models, and the error metrics employed to validate the accuracy of the regressors. The implementation of the ML approach described in this section is publicly available on GitHub [47].

1) ERROR METRICS

Considering the importance of accurate and reliable RSSI predictions in urban mesh networks, we chose error metrics that evaluate the performance of our models in terms of both error magnitude (in dBm units) measured by the RMSE and the error percentage by the Mean Absolute Percentage Error (MAPE).

The RMSE is calculated as in

$$\text{RMSE} = \sqrt{\frac{1}{n_l}\sum_{i=1}^{n_l}\left(\hat{RSSI}_i - RSSI_i\right)^2}. \tag{28}$$

The MAPE is calculated as in

$$\text{MAPE} = \frac{1}{n_l}\sum_{i=1}^{n_l}\left|\frac{RSSI_i - \hat{RSSI}_i}{RSSI_i}\right| 100. \tag{29}$$

In (28) and (29), n_l is the number of links, $\hat{RSSI}_i$ is the i-th estimated value of the received signal strength indicator,

and $RSSI_i$ is the i-th real value of the received signal strength indicator.

2) HYPERPARAMETER TUNING

Hyperparameters, prevalent in most machine learning algorithms, play a crucial role in determining their performance.

To implement the Random Forest algorithm in R language, we used the *randomForest* function from the package *randomForest* [48]. In Random Forest, we set the hyperparameter `mtry` to integer numbers from 1 to 9. The parameter `mtry` refers to the number of random variables chosen as candidates for each split in the tree.

The optimal performance was determined using a grid search combined with 10-fold cross-validation. The grid search involved specifying a range of hyperparameters to be tested and then systematically evaluating the performance of the model for each combination of hyperparameters. The data was repeatedly split into 10 folds, where 9 folds were used for training and 1 fold for validation in each iteration. This process was repeated 10 times, ensuring that each fold was used exactly once for validation. The performance metric used for evaluation was Root Mean Squared Error (RMSE). The best result was achieved at an `mtry` value of 4. For the other parameters, we assigned the default values of the *randomForest* function, which include the number of trees (`ntree`) set to 500 and the minimum number of samples in a node required to split (`nodesize`) set to 5. Additionally, `nodesize` serves as the stopping criterion of the algorithm for splitting nodes.

The Support Vector Regression algorithm was implemented in R language using the *svm* function from the *e1071* package [49]. In SVR, we tuned the following hyperparameters:

- Epsilon (ϵ): specifies a threshold below which prediction errors are not penalized. To tune ϵ we define a set of values $[0.1, 0.2, \ldots, 1]$
- Cost (C): determines the penalty for prediction errors. A larger C will prioritize minimizing errors but might increase the possibility of overfitting. To tune C we define a set of values $[2^0, 2^1, \ldots, 2^9]$
- Kernel function: both the RBF and sigmoid kernels were considered for the kernel function.

Finally, considering the RMSE obtained in the tuning process, the definitive values chosen for the hyperparameters were 0.5 for ϵ, 128 for C, and RBF as the kernel function.

3) TRAIN AND TEST

To increase the reliability of the results obtained by the predictive algorithms, we used the k-fold cross-validation method. In this method, the dataset is divided into k approximately equal-sized subsets, where $k-1$ subsets are used as training data, and the remaining subset is the test set. This process is repeated k times, each of them using a different subset as the test set. The final performance of the algorithm is the average performance in each test subset. In our dataset, we define $k = 10$. Therefore, the performance of the SVR or Random Forest algorithm is the average of the errors obtained in these ten tests.

V. RESULTS AND DISCUSSION

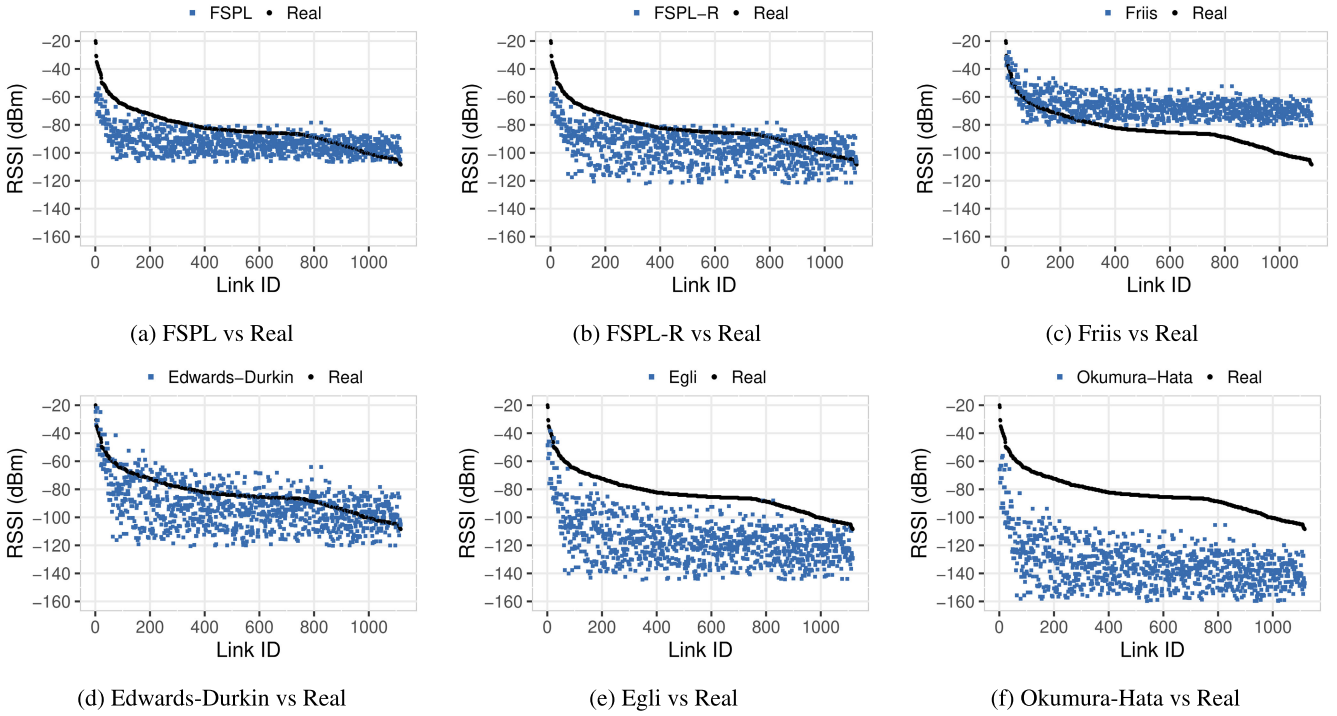
We present a comparison and the prediction results of RSSI obtained by the classical, empirical, and machine learning models. In Fig. 6 and Fig. 7, we depict the distribution of both the real RSSI and the predicted RSSI from each model. For better visualization, we sorted the 1,117 links from the real RSSI dataset in descending order from strong to weak RSSI.

From Fig. 6a to Fig. 6c, which correspond to the classical models, the Free Space Path Loss and its variation considering reflection in the path underestimated mainly the RSSI with strong signal, and both concentrated its prediction RSSI in the range from -110 dBm and -80 dBm. Furthermore, in our urban scenario, the FSPL-R slightly increases the error prediction compared to the FSPL. This result was counterintuitive, given that the links are not predominantly LOS in such scenario. Regarding the Friis Equation, almost all predictions were overestimated, concentrating the predicted RSSI between -80 dBm to -60 dBm. In conclusion, from the classical models, the best prediction results are obtained from the original formula of the Free Space Path Loss.

From Fig. 6d to Fig. 6f, which correspond to the urban empirical models, there is a tendency in all models to underestimate the real RSSI, mainly in the prediction results obtained from the Okumura-Hata model. Also, different from the prediction results in the classical models, the range of the predicted RSSI from Okumura-Hata and Egli exceeded the radio sensitivity level of the mesh nodes and concentrator. The predicted RSSI values from Okumura-Hata are concentrated in the range of -160 dBm to -120 dBm, while the Egli model is from -140 dBm to -100 dBm. On the other hand, the Edwars-Durkin was the unique empirical model that concentrated its prediction RSSI values according to the radio sensitivity level and had the best fitting distribution related to real RSSI.

As discussed earlier, the empirical models are still frequently used. However, these models are influenced by the unique characteristics of the regions where the data were gathered. Even though the selected urban empirical models consider the carrier frequency of our networks, the urban characteristics found in the Metropolitan Region of São Paulo may differ from the original urban scenario in which these models were formulated. Furthermore, another reason that might have influenced the high prediction error is that these models are derived based on data obtained from mobile networks, and as discussed in Section II, there are some considerable differences from the wireless mesh networks.

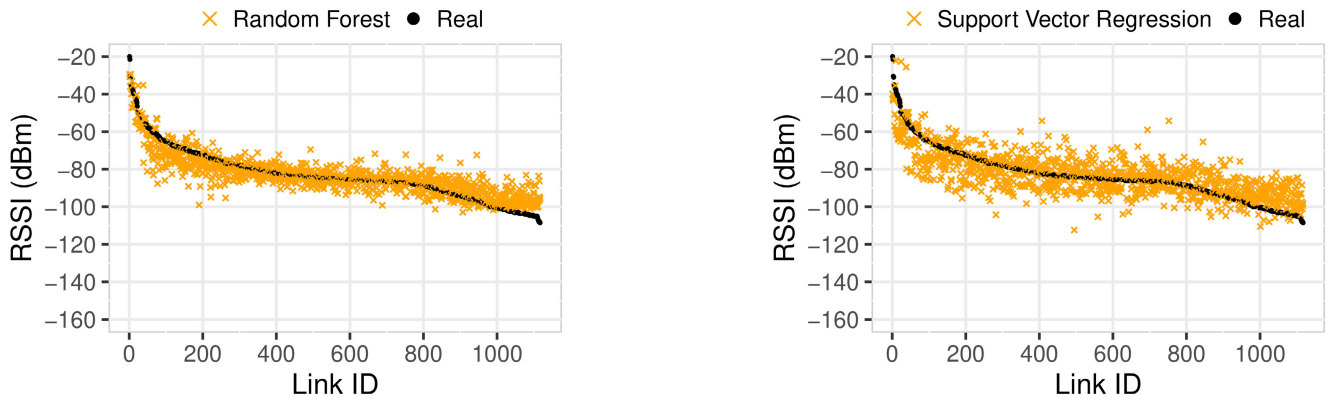
From our proposed approach using the Random Forest and Support Vector Regression algorithms, illustrated in Fig. 7a and Fig. 7b, both algorithms had a similar performance. However, the Random Forest exhibited a slightly more accurate prediction than SVR. Additionally, both algorithms



(a) FSPL vs Real (b) FSPL-R vs Real (c) Friis vs Real

(d) Edwards-Durkin vs Real (e) Egli vs Real (f) Okumura-Hata vs Real

FIGURE 6. **Comparison between the real RSSI and the prediction by classical and empirical models.**



(a) Random Forest vs Real (b) Support Vector Regression vs Real

FIGURE 7. **Comparison between the real RSSI and the prediction by machine learning models.**

showed the same propensity to underestimate the RSSI values between the range of −80 dBm and −50 dBm. In contrast, in the −110 dBm to −90 dBm range, both algorithms tended to overestimate weak RSSI values close to the radio sensitivity level.

To quantitatively evaluate the performance, in Table 4, we presented the RMSE and MAPE of the prediction models.

According to the values in Table 4, the machine learning models outperformed the classical and empirical models. In particular, the Random Forest algorithm had the best performance with a prediction error of approximately 6 dBm, followed by the Support Vector Regression with an error of 8 dBm.

TABLE 4. **Error measurements from all prediction models.**

Models	RMSE (dBm)	MAPE
FSPL	14.6	0.16
FSPL-R	18.2	0.20
Friis	19.9	0.20
Egli	35.5	0.41
Edwards-Durkin	16.5	0.17
Okumura-Hata	52.4	0.63
Random Forest	**5.6**	**0.05**
Support Vector Regression	8.0	0.08

Regarding the classical models, the best performance was obtained from the Free Space Path Loss model with a prediction error of around 15 dBm, while the FSPL-R achieved approximately 18 dBm and the Friis Equation equal to 20 dBm.

Concerning the urban empirical models, the Edwards-Durkin model outperformed the Egli and Okumura-Hata models. The prediction error using the Edwards-Durkin was around 16 dBm, while the Okumura-Hata had the worst performance with an error of approximately 52 dBm.

In conclusion, the two machine learning algorithms used in our approach to predict RSSI for urban mesh networks outperformed all traditional models. The prediction performance using Random Forest increased the accuracy by 62% compared to the best results from classical models (FSPL) and 66% compared to the best results from empirical urban models (Edwards-Durkin). Moreover, the prediction error of 6 dBm achieved using our proposed approach with the Random Forest algorithm is considered acceptable according to the literature for predicting the path loss in urban environments [38].

A. FEATURE IMPORTANCE ANALYSIS

The analysis of feature importance provides insights into how the features, representing distinct aspects of the propagation environment, interact with each other and influence the machine learning process.

We analyzed the feature importance from the results obtained in Random Forest, which performed best. According to [50], the most relevant measure to evaluate the feature contribution in Random Forest is the percentage increase in mean squared error (*%IncMSE*). The *%IncMSE* is calculated for each input feature by measuring how much the mean squared error increases when that feature is randomly permuted in the test data. Thus, if the error increases significantly after permuting a specific feature, this indicates that this feature holds substantial importance for the prediction model. On the contrary, if the increase in error is minimal or unchanged, the feature has less influence or is potentially irrelevant to the model. Therefore, based on the *%IncMSE* values, we can evaluate which features have the most significant impact on the prediction and that best describes the relationship between signal attenuation along the path from the transmitter to the receiver. Table 5 shows the *%IncMSE* for each feature used in the Random Forest.

TABLE 5. %*IncMSE* values of the features in random forest.

Feature	*%IncMSE*
Percentage of obstruction in the first Fresnel zone	90.9
Distance between the transmitter and receiver	78.6
Total height of the receiver	56.9
Total height of the transmitter	37.6
Standard deviation of the altitudes on the link path	34.7
Median of the altitudes on the link path	34.0
Mean of the altitudes on the link path	33.1
Maximum altitude	30.3
Minimum altitude	29.2

According to Table 5, the feature with the highest importance in the RSSI prediction was the obstruction percentage in the first Fresnel zone, followed by the distance from the transmitter to the receiver. These results showed the importance of considering the geometric patterns of radio wave propagation by quantifying the obstruction in the Fresnel zone as input for the machine learning approach. Although this feature is rarely considered in the literature, in our ML approach to predicting the RSSI for urban mesh networks, this feature has shown a substantial contribution.

Furthermore, as the second most influential feature, the distance from the transmitter to the receiver still plays a fundamental role in the RSSI prediction, even in urban regions where signal attenuation factors extend beyond the relation of decreasing the RSSI as the distance increases. Similar to the distance feature, the effective height of the receiver and transmitter antennas, present in classical and empirical models, also contribute significantly to the RSSI prediction.

The last feature group is the statistical features derived from the profile terrain elevation. The feature that better indicates the terrain complexity was the standard deviation followed by the median of the elevation along the path profile.

Our work pioneers the application of machine learning for RSSI prediction in urban mesh networks. Unlike the previous work presented by [12], our analysis incorporates a broader set of features tailored to urban environments. For a direct comparison, these features are presented in Table 6, ordered by its importance. This comprehensive approach highlights the unique factors influencing urban mesh networks, demonstrating the adaptability and specificity of our machine learning model in addressing these challenges.

B. EXTENSION AND APPLICABILITY TO OTHER URBAN MESH NETWORKS

Considering that this work introduces a machine learning approach for RSSI prediction in urban mesh networks for the first time, several points need discussion regarding its application to all mesh networks embedded in similar characteristics. As highlighted throughout the article, the selected features were based on traditional models and literature related to other networks, such as mobile, television, and radio networks, with consideration for ease of obtaining them in real-world scenarios. Consequently, the approach presented here is potentially applicable to other mesh networks in urban scenarios.

The chosen features account for the region's particularities, considering obstruction percentage in the line of sight, as well as aspects of the network devices, such as antenna height and distance between them. The results from the analysis of the importance of each feature in the RSSI prediction process demonstrate that certain features are as important in urban scenarios as those in mountain conditions presented by [12].

Moreover, the results indicate that the proposed approach is a superior alternative to traditional and empirical propagation models, validated and applied to predict RSSI in two

TABLE 6. Comparison of features used in mesh networks in order of importance.

Features in Urban Mesh Networks (this work)	Features in Mountainous Mesh Networks [12]
Percentage of obstruction in the first Fresnel zone	Distance between the transmitter and receiver
Distance between the transmitter and receiver	Terrain standard deviation
Total height of the receiver	Vegetation standard deviation
Total height of the transmitter	Average tree canopy coverage
Standard deviation of the altitudes on the link path	Angle between line of sight (LOS) and horizontal plane
Median of the altitudes on the link path	Canopy coverage at transmitter location
Mean of the altitudes on the link path	Canopy coverage at receiver location
Maximum altitude	
Minimum altitude	

networks operating in real-world applications. This approach for predicting RSSI in urban scenarios may be extended to other mesh networks based on the expected error presented.

It is essential to note that predicting RSSI for other mesh networks operating at significantly different frequencies may however lead to variations not identified in this study.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced a pioneering machine learning approach to predict RSSI for mesh networks situated in urban areas. Our methodology was developed using real-world RSSI measurements extracted from two mesh networks in the Metropolitan Region of São Paulo. A review of existing literature revealed that, while several studies have developed machine learning approaches to predict path loss and RSSI, their primary focus has been on broadcast and mobile networks. We underscored the significant differences between these networks and mesh networks, emphasizing aspects such as topology, architecture, and antenna type. Based on this literature and the unique characteristics of urban mesh networks, we defined a set of features that describe the radio propagation environment, serving as inputs for the introduced machine learning algorithms.

To predict RSSI for the investigated urban mesh networks, we adopted the Random Forest and Support Vector Regression, which have shown significant performance results in RSSI prediction. Additionally, we selected appropriate classical and empirical models for comparative prediction results, such as the Free Space Path Loss and Okumura-Hata. To evaluate the results from all these prediction models, we used the RMSE and MAPE metrics. Based on the RMSE analysis, quantifying the error in dBm units, the Random Forest approach outperformed all models, achieving a prediction error of 6 dBm.

Moreover, from the Random Forest results, we presented a feature importance analysis describing the contributions of the set of features in RSSI prediction for urban mesh networks. From this analysis, we showed that the feature not usually considered in the literature, representing the percentage of obstruction in the first Fresnel zone, was the most important for the introduced model, followed by the distance from the transmitter to the receiver. Additionally, the effective height of the transmitter and receiver antennas and the terrain variation along the propagation path, measured by the terrain elevation's standard deviation, significantly contributed to the prediction results.

Considering that this is the first time a machine learning approach to predict RSSI was explored for mesh networks in urban regions, for future work, we intend to include new features, such as building information, that are present in the radio propagation environment found in urban areas. We will then compare the performance of these features with those presented in the approach proposed in this paper.

REFERENCES

[1] B. De Beelde, M. Vantorre, G. Castellanos, M. Pickavet, and W. Joseph, "MmWave physical layer network modeling and planning for fixed wireless access applications," *Sensors*, vol. 23, no. 4, p. 2280, Feb. 2023.

[2] H. T. Friis, "A note on a simple transmission formula," *Proc. IRE*, vol. 34, no. 5, pp. 254–256, May 1946.

[3] U. Masood, H. Farooq, A. Imran, and A. Abu-Dayya, "Interpretable AI-based large-scale 3D pathloss prediction model for enabling emerging self-driving networks," *IEEE Trans. Mobile Comput.*, vol. 22, no. 7, pp. 3967–3984, Aug. 2023.

[4] A. Seretis and C. D. Sarris, "An overview of machine learning techniques for radiowave propagation modeling," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 3970–3985, Jun. 2022.

[5] Y Okumura, E Ohmori, T Kawano, and K Fukuda, "Empirical formula for propagation loss in land mobile radio services," *Rev. Electr. Commun. Lab.*, vol. 3, no. 3, pp. 317–325, Aug. 1968.

[6] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. Veh. Technol.*, vol. VT-29, no. 3, pp. 317–325, Sep. 1980.

[7] J. Egli, "Radio propagation above 40 MC over irregular terrain," *Proc. IRE*, vol. 45, no. 10, pp. 1383–1391, 1957.

[8] N. Faruk, S. I. Popoola, N. T. Surajudeen-Bakinde, A. A. Oloyede, A. Abdulkarim, L. A. Olawoyin, M. Ali, C. T. Calafate, and A. A. Atayero, "Path loss predictions in the VHF and UHF bands within urban environments: Experimental investigation of empirical, heuristics and geospatial models," *IEEE Access*, vol. 7, pp. 77293–77307, 2019.

[9] Y. Zhang, J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: Principle, method, and data expansion," *Appl. Sci.*, vol. 9, no. 9, p. 1908, May 2019.

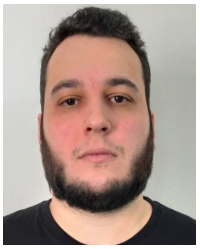
[10] A. Guidara, G. Fersi, M. B. Jemaa, and F. Derbel, "A new deep learning-based distance and position estimation model for range-based indoor localization systems," *Ad Hoc Netw.*, vol. 114, Apr. 2021, Art. no. 102445.

[11] M. F. A. Fauzi, R. Nordin, N. F. Abdullah, H. A. H. Alobaidy, and M. Behjati, "Machine learning-based online coverage estimator (MLOE): Advancing mobile network planning and optimization," *IEEE Access*, vol. 11, pp. 3096–3109, 2023.

[12] C. A. Oroza, Z. Zhang, T. Watteyne, and S. D. Glaser, "A machine-learning-based connectivity model for complex terrain large-scale low-power wireless deployments," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 576–584, Dec. 2017.

[13] S. I. Popoola, A. Jefia, A. A. Atayero, O. Kingsley, N. Faruk, O. F. Oseni, and R. O. Abolade, "Determination of neural network parameters for path loss prediction in very high frequency wireless channel," *IEEE Access*, vol. 7, pp. 150462–150483, 2019.

[14] M. Sousa, A. Alves, P. Vieira, M. P. Queluz, and A. Rodrigues, "Analysis and optimization of 5G coverage predictions using a beamforming antenna model and real drive test measurements," *IEEE Access*, vol. 9, pp. 101787–101808, 2021.
[15] U. Masood, H. Farooq, and A. Imran, "A machine learning based 3D propagation model for intelligent future cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
[16] *Prediction Procedure for the Evaluation of Interference Between Stations on the Surface of the Earth At Frequencies Above About 0.1 GHz*, document ITU-R P.452, 2013.
[17] A. D. S. Braga, H. A. O. D. Cruz, L. E. C. Eras, J. P. L. Araujo, M. C. A. Neto, D. K. N. Silva, and G. P. S. Cavalcante, "Radio propagation models based on machine learning using geometric parameters for a mixed city-river path," *IEEE Access*, vol. 8, pp. 146395–146407, 2020.
[18] *Method for Point-to-Area Predictions for Terrestrial Services in the Frequency Range 30 MHz To 3000 MHz*, document ITU-R P.1546, 2013.
[19] O. J. Famoriji and T. Shongwe, "Path loss prediction in tropical regions using machine learning techniques: A case study," *Electronics*, vol. 11, no. 17, p. 2711, Aug. 2022.
[20] N. Faruk, Q. R. Adebowale, I.-F.-Y. Olayinka, K. S. Adewole, A. Abdulkarim, A. A. Oloyede, H. Chiroma, O. A. Sowande, L. A. Olawoyin, S. Garba, A. D. Usman, Y. A. Adediran, and L. S. Taura, "ANN-based model for multiband path loss prediction in built-up environments," *Scientific Afr.*, vol. 17, Sep. 2022, Art. no. e01350.
[21] A. Gupta, J. Du, D. Chizhik, R. A. Valenzuela, and M. Sellathurai, "Machine learning-based urban canyon path loss prediction using 28 GHz Manhattan measurements," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 4096–4111, Jun. 2022.
[22] F. Malandra, H. Mellah, A. D. Firouzabadi, C. Wetté, and B. Sansò, "A layered and grid-based methodology to characterize and simulate IoT traffic on advanced cellular networks," *IEEE Internet Things Mag.*, vol. 6, no. 1, pp. 134–140, Mar. 2023.
[23] N. Moraitis, L. Tsipi, D. Vouyioukas, A. Gkioni, and S. Louvros, "Performance evaluation of machine learning methods for path loss prediction in rural environment at 3.7 GHz," *Wireless Netw.*, vol. 27, no. 6, pp. 4169–4188, Aug. 2021.
[24] R.-T. Juang, "Explainable deep-learning-based path loss prediction from path profiles in urban environments," *Appl. Sci.*, vol. 11, no. 15, p. 6690, Jul. 2021.
[25] L. Wu, D. He, B. Ai, J. Wang, H. Qi, K. Guan, and Z. Zhong, "Artificial neural network based path loss prediction for wireless communication network," *IEEE Access*, vol. 8, pp. 199523–199538, 2020.
[26] H.-S. Jo, C. Park, E. Lee, H. K. Choi, and J. Park, "Path loss prediction based on machine learning techniques: Principal component analysis, artificial neural network, and Gaussian process," *Sensors*, vol. 20, no. 7, p. 1927, Mar. 2020.
[27] S. Ojo, A. Sari, and T. P. Ojo, "Path loss modeling: A machine learning based approach using support vector regression and radial basis function models," *Open J. Appl. Sci.*, vol. 12, no. 6, pp. 990–1010, 2022.
[28] K. E. R. Morico, K. J. G. Porras, J. M. Judan, M. F. D. De Guzman, and C. A. G. Hilario, "Assessment of television white space in the greater metro Manila area through geospatial and empirical approaches," in *Proc. ISAP*, 2021, pp. 149–150.
[29] S. I. Popoola, N. Faruk, N. T. Surajudeen-Bakinde, A. A. Atayero, and S. Misra, "Artificial neural network model for path loss predictions in the VHF band," in *Proc. Conf. ICDLAIR*, M. Tripathi and S. Upadhyaya, Eds. Cham, Switzerland: Springer, 2021, pp. 161–169.
[30] R. E. Edwards and J. Durkin, "Computer prediction of field strength in the planning of radio systems," in *Proc. Inst. Electr. Engineers*, 1969, pp. 1493–1500.
[31] J. S. Seybold, *Introduction To RF Propagation*. Hoboken, NJ, USA: Wiley, 2005.
[32] D. D. Coleman and D. A. Westcott, *CWNA Certified Wireless Network Administrator*, 3rd ed. Indianapolis, IN, USA; Wiley, 2012.
[33] G. Y. Delisle, J.-P. Lefevre, M. Lecours, and J.-Y. Chouinard, "Propagation loss prediction: A comparative study with application to the mobile radio channel," *IEEE Trans. Veh. Technol.*, vol. VT-34, no. 2, pp. 86–96, May 1985.
[34] R. Edwards and J. Durkin, "Computer prediction of service areas for V.H.F. mobile radio networks," *Proc. Inst. Electr. Engineers*, vol. 116, no. 9, p. 1493, 1969.
[35] S. R. Theodore, *Wireless Communications: Principles and Practice*. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
[36] T. Jawhly and R. C. Tiwari, "The special case of egli and Hata model optimization using least-square approximation method," *Social Netw. Appl. Sci.*, vol. 2, no. 7, pp. 1–10, Jul. 2020.
[37] C. Huang, R. He, B. Ai, A. F. Molisch, B. K. Lau, K. Haneda, B. Liu, C.-X. Wang, M. Yang, C. Oestges, and Z. Zhong, "Artificial intelligence enabled radio propagation for communications—Part II: Scenario identification and channel modeling," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 3955–3969, Jun. 2022.
[38] N. Moraitis, L. Tsipi, and D. Vouyioukas, "Machine learning-based methods for path loss prediction in urban environment for LTE networks," in *Proc. 16th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2020, pp. 1–6.
[39] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[40] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 1–12.
[41] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[42] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, "Dplyr: A grammar of data manipulation," R package version 1.1.0, Tech. Rep., 2023.
[43] S. Azarhazin, H. Farzin, and E. Mashhour, "An MILP model for reliability-based placement of recloser, sectionalizer, and disconnect switch considering device relocation," *Sustain. Energy, Grids Netw.*, vol. 35, Sep. 2023, Art. no. 101127.
[44] J. Hollister, T. Shah, A. L. Robitaille, M. W. Beck, and M. Johnson, "Elevatr: Access elevation data from various Apis," R package version 0.3.1, Tech. Rep., 2020.
[45] Accessed: Aug. 8, 2023. [Online]. Available: https://registry.opendata.aws/terrain-tiles
[46] M. Padgham, "Geodist: Fast, dependency-free geodesic distance calculations," R package version 0.0.7, Tech. Rep., 2021.
[47] M. Jeske. (2024). *Received Signal Strength Indicator Prediction for Mesh Networks in a Real Urban Environment Using Machine Learning*. [Online]. Available: https://github.com/marlonjeske/rssipredictionurbanmesh
[48] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[49] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071)," R package version 1.7-4, Tech. Rep., 2020.
[50] C. Dewi and R.-C. Chen, "Random forest and support vector machine on features selection for regression analysis," *Int. J. Innov. Comput. Inf. Control*, vol. 15, no. 6, pp. 2027–2037, 2019.



MARLON JESKE received the B.S. degree in mathematics from the Regional University of Blumenau (FURB), in 2015, and the M.S. degree in operations research from the Aeronautics Institute of Technology (ITA) and the Federal University of São Paulo (UNIFESP), in 2019, where he is currently pursuing the Ph.D. degree in operations research. His research interests include employing mono and multiobjective optimization techniques, metaheuristics, and machine learning methodologies to address challenges in planning and deploying wireless networks.

BRUNILDE SANSÒ (Senior Member, IEEE) is currently a Full Professor in telecommunication networks with the Department of Electrical Engineering, Polytechnique Montréal. She is also the Director of the LORLAB, a research group dedicated to developing effective methods for the design and performance of wireless and wireline telecommunication networks. She has published extensively in the telecommunications and operations research literature and has acted as a consultant for telecommunication operators, equipment manufacturers, and the mainstream media.

Miss. Sansò has received several awards and honors.

DANIEL ALOISE received the Ph.D. degree in applied mathematics from Polytechnique Montréal, Montreal, QC, Canada, in 2009. He is currently a Full Professor with the Computer and Software Engineering Department, Polytechnique Montréal. He has published articles in leading machine learning and operations research journals during his career. His research interests include data mining, optimization, and mathematical programming, and how these disciplines interact to tackle problems in the big data era. He is a member of the Group for Research in Decision Analysis (GERAD) and a fellow of Canada Excellence Research Chair in Data Science for Real-Time Decision-Making.

MARIÁ C. V. NASCIMENTO received the Ph.D. degree in computer science and applied mathematics from the University of São Paulo (USP), Brazil, in 2010.

She is currently an Associate Professor with the Computer Science Division, Aeronautics Institute of Technology (ITA). She has published extensively in the operations research literature. Her research interests include operations research and machine learning in a wide range of applications, such as telecommunications, industry, and health care. She is an associate editor of leading journals.