Enhancing Reference Signal Received Power Prediction Accuracy in Wireless Outdoor Settings: A Comprehensive Feature Importance Study

Marlon Jeske, Daniel Aloise, Brunilde Sansò, *Senior Member, IEEE*, and Mariá C. V. Nascimento

***Abstract*—Predicting the reference signal received power (RSRP) in wireless communication is crucial for improving network performance, allocating resources, and ensuring good signal coverage, especially in advanced technologies like 5G and beyond. To create an accurate prediction model, we need to look at different aspects of the radio environment and understand the importance of each factor. In our study, we suggest using machine learning (ML) to predict RSRP. We analyze the importance of features by studying their impact on the received signal power. We developed an ML approach using 64 features taken from recent literature and new ones proposed in this study from real-world received signal power measurements in outdoor areas, including cities and suburbs. Using this data, we trained a random forest (RF) model for received signal power predictions. After training, we analyzed the importance of each feature to create a simpler ML model that maintains good prediction accuracy. Our results show that we can use only the 25 most important features to build a less complex model with a small error difference of 0.14 dB compared with the original model with 64 features.**

***Index Terms*—Feature analysis, machine learning (ML), propagation models, received signal power prediction, wireless network planning and deployment.**

I. Introduction

As WIRELESS communication systems progress from 5G to 6G and beyond, the accurate prediction of signal losses between transmitters (Txs) and receivers (Rxs) assumes growing significance. This is particularly noteworthy with the utilization of terahertz frequencies, where an escalation in signal losses and a subsequent reduction in coverage are observed [1], [2], [3], [4]. Consequently, the ability to predict the received signal power becomes indispensable for ensuring reliable communication services. This involves the efficient estimation of signal coverage and the identification of optimal Tx placement strategies.

Various prediction models are available in the literature. Deterministic prediction models, based on electromagnetic theory and methods like ray tracing, predict path loss by considering the physical characteristics of the propagation environment and solving Maxwell's equations [5]. Although accurate, they require detailed site-specific information and are computationally intensive. In contrast, empirical models, such as Hata [6], Edwards and Durkin [7], [8], and Egli [9], rely on extensive propagation measurements across different environments and frequency bands, offering computational efficiency. However, these models do not fully capture the complexities of actual deployment environments, leading to potential inaccuracies in prediction reliability.

In recent times, there has been a surge in the utilization of machine learning (ML) models for received signal power and path loss prediction [10], [11]. The underlying concept involves leveraging data-driven techniques to discern patterns and relationships between the radio propagation environment (input) and received signal strength (output). This is achieved through extensive sets of training data acquired from real-world campaigns or simulation tools, generating synthetic data.

The literature widely supports this paradigm shift, demonstrating that ML algorithms consistently outperform empirical models [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. Notable quantifiable results include a 65% increase in prediction accuracy using the proposed ML approach compared with the best-performing empirical models [21]. In addition, a substantial 28.8-dB reduction in prediction error was identified in comparison to the Okumura-Hata model [14].

However, the use of ML for received signal power and path loss prediction comes with challenges [30]. Acquiring extensive data through campaigns is costly and logistically challenging, limiting the data to existing technologies [31]. Furthermore, there is significant data variability dependent on network and environmental conditions. Compounding these challenges is the lack of consensus in the literature regarding criteria for creating, extracting, or determining features for ML-based signal loss prediction models. The choice of

Received 9 October 2024; revised 22 April 2025; accepted 30 May 2025. Date of publication 10 June 2025; date of current version 14 October 2025. This work was supported in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant 309385/2021-0, Grant 403735/2021-1, and Grant 142311/2019-7; in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2022/05803-3 and Grant 2013/07375-0; and in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) under Finance Code 001. *(Corresponding author: Marlon Jeske.)*

Marlon Jeske and Mariá C. V. Nascimento are with the Computer Science Division, Aeronautics Institute of Technology, São José dos Campos, São Paulo 12228-900, Brazil (e-mail: marlonjeske03@gmail.com; mariah@ita.br).

Daniel Aloise is with the Department of Computer Engineering and Software Engineering, Polytechnique Montréal, Montreal, QC H3T 1J4, Canada (e-mail: daniel.aloise@polymtl.ca).

Brunilde Sansò is with the Department of Electrical Engineering, Polytechnique Montréal, Montreal, QC H3T 1J4, Canada (e-mail: brunilde.sanso@polymtl.ca).

Digital Object Identifier 10.1109/TAP.2025.3576492

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

features significantly influences prediction performance, as models establish relationships between independent (input) and dependent (output) variables. Feature engineering for reference signal received power (RSRP) involves parameters from propagation models [13], [29], geographic data, Tx technological parameters, and information from data collection campaigns. The set of features may contain characteristics relevant to the propagation environment closely linked to path loss [11], [31].

As demonstrated by Zhang et al. [10], neglecting relevant features or including unrelated ones can compromise predictor quality. Therefore, assessing the importance of features becomes fundamental for developing more precise propagation models using ML frameworks. Feature importance analysis elucidates how features representing different propagation environment characteristics interrelate and the impact of their combinations on the learning process. Despite its significance, only a limited number of recent studies have investigated feature importance analysis in urban and suburban scenarios [18], [21], [30], [32], [33].

In this article, we propose a comprehensive study and application of ML for predicting received signal power. The main contributions are as follows.

1) *Random Forest (RF) Model to Predict RSRP:* For the prediction, we performed a thorough investigation of the features considered in the literature. A taxonomy of the features based on their characteristics is presented.
2) *In-Depth Feature Importance Analysis in Real-World Data Collected From Hundreds of Base Stations Located in Diverse Urban Scenarios:* We conduct a feature importance analysis using the random permutation method for RF, highlighting the most and least significant features for RSRP prediction. The scarce literature that approaches the feature importance in urban and suburban scenarios is limited to a specific location or relies on synthetic data. In this article, we use real RSRP data obtained from measurement campaigns conducted in various urban scenarios with frequencies ranging from 2.585 to 2.624 GHz.
3) *Method to Significantly Reduce the Complexity of the ML Model and the Number of Features Based on Their Importance Ranking*: Leveraging the significance of each feature, we propose reduced feature sets by combining the most important features. Finally, we evaluate the performance of these reduced-feature ML regressions and provide practical recommendations for constructing prediction models characterized by diminished complexity in both feature selection and dataset variety.

The rest of this article is organized as follows. In Section II, we present the literature review and explore the criteria to choose and define the features for received power prediction models. Section III is devoted to presenting all the features we use to develop our ML approach. The raw data, the methodology to extract the path profile information and features, and statistics analysis with highlights about the general scenarios are given in Section IV. Section V contains the ML approach and preliminary results from computational experiments. In Section VI, we show and discuss the importance of each feature in the learning process and suggest new models to predict the RSRP with reduced complexity. Finally, concluding remarks and future research directions are provided in Section VII.

II. Literature Review

This literature review examines recent work using ML to predict received signal strength indicator (RSSI), RSRP, or path loss in urban and suburban environments across a wide frequency range from 62.25 MHz to 70 GHz. This comprehensive analysis evaluates various propagation scenarios and the criteria for feature selection in ML models, aiming to develop a robust prediction model with a concise set of relevant features for estimating signal propagation losses. Although RSSI, RSRP, and path loss are distinct metrics, they all reflect the signal power received at a specific location and are influenced by the same environmental and propagation factors between Tx and Rx. Consequently, the features representing the propagation environment, including Tx and Rx information, are relevant across all these measures.

A. General Overview

According to [30], ML approaches face three significant challenges in path loss prediction: obtaining data, extrapolating to new scenarios, and reducing the complexity of the models and features. The data collection is usually expensive, time-consuming, and limited by the equipment configurations, which makes it difficult to execute in various scenarios [31], [34]. A common way to obtain data involves campaigns that use radio Rxs installed on vehicles to measure the signal strength at different points in the region of interest [16], [18], [25], [35], [36], [37]. Thus, data from distinct scenarios and environments, as well as in different operating frequencies, are scarce. As a consequence, extrapolation to new scenarios is limited as the models may be inadequate to handle various characteristics of these new regions or operating frequencies. For example, a prediction model developed in a suburban residential area, predominantly composed of houses, may be insufficient to capture the intrinsic diversities of an urban area with several skyscrapers. Finally, the predictions obtained from ML models are often difficult to interpret in terms of how the features are related to path loss prediction [30].

B. Factors Considered in the Literature

The signal propagation losses can be influenced by a wide variety of factors, ranging from the physical characteristics of the radio propagation environment to the parameters of the transmitting and receiving antennas. Therefore, properly choosing and using these factors as features in the ML approach is crucial to improving prediction accuracy and achieving better performance than traditional models. In this article, we will address various characteristics highlighted in the literature that have been used to define the set of features for received signal power and path loss prediction. We will present and discuss these characteristics in Sections II-B1–II-B8 that include the following:

1) scenarios related to network location;
2) frequency;
3) distance;
4) geographical coordinates and line of sight (LOS) existence;
5) path profile;
6) location environment;
7) technological parameters;
8) model-based elements.

1) Location Scenarios: Location characteristics are broadly categorized as indoor or outdoor scenarios.

In outdoor scenarios—urban, suburban, and rural—obstructions vary. Rural areas face terrain complexity, vegetation, tree density, and seasonal changes [38], [39]. Urban and suburban areas encounter obstacles from tall buildings, mixed zones, and industrial or residential areas.

Indoor scenarios differ significantly, with shorter distances and higher environmental variability [40]. These distinctions arise from building structures, room layouts, human presence, and materials [40], [41], [42]. Prediction models for outdoor scenarios may not apply indoors [42], [43].

Outdoor scenario classification (urban, suburban, and rural) is subjective, based on user perceptions [13]. Other classifications exist, such as dense urban, urban, dense suburban, suburban, and rural [14], lacking explanation. The absence of a consensus makes determining successful features challenging, emphasizing the need to identify area characteristics for effective feature representation.

2) Frequency: The carrier frequency significantly influences signal attenuation in various scenarios. Higher frequencies experience greater signal loss over distance, and millimeter waves (30–300 GHz) are particularly affected by atmospheric absorption [44]. ML models use diverse operating frequencies, with some focusing on a single frequency and others on frequency intervals. In the very high frequency (VHF) band, common features include Tx–Rx distance, antenna height, and elevation terrain data. Applications often involve TV, radio, and smart metering networks [15], [45], [46], [47]. The ultrahigh frequency (UHF) band attracts the most studies for signal propagation losses prediction models, using features common for UHF models with additional features such as coordinates, clutter data, and technological parameters [10], [12], [13], [14], [18], [20], [21], [22], [25], [26], [35], [36], [37], [48], [49], [50], [51], [52], [53], [54], [55]. In the super high frequency (SHF) band, models commonly use features such as distance, building statistics, and path profile obstacles [30], [33], [56]. Work across operating frequencies (VHF, UHF, SHF, and EHF) and application scenarios from TV to 5G show that the most common features are those used in UHF models [14], [27], [45], [51], [52], [56]. Despite earlier considerations that frequency influences prediction models, recent work often neglects it as a feature, focusing on distance as a primary predictor.

3) Distance: Distance from Tx to Rx is an essential factor in predicting signal propagation losses regardless of the approach adopted. First, a signal will be attenuated as it propagates, even without attenuation caused by obstructions, absorption, or diffraction. The free space path loss is the intensity loss of the signal caused by the natural widening of radio waves. In a transmission scenario without obstruction in the propagation path, the signal loss will be a function of the distance and frequency [57]. This idea can be seen in the first propagation models, such as the well-known Friis equation [58].

According to [12], all empirical and deterministic models use distance as one of the parameters. Moreover, the logarithm of distance is largely employed in heuristic channel models [30]. Several other models for predicting RSSI based on distance can also be found in [59]. Among the work reviewed in this article, only [52] does not use distance as a feature in their ML approach.

4) Coordinates: Geographical coordinates of Tx and Rx antennas are also used as features [22], [26], [27], [29], [32], [52], [54], [55], [60]. However, their direct use is limited to local prediction.

5) Path Profile and LOS: A good number of features can be considered by combining information on the terrain and buildings along the environment between Tx and Rx, typically called the path profile. Such features can describe, for example, the street width [30], [48], the irregularity of the terrain elevation [46], the indication if there is an obstruction in the LOS [18], [20], [21], [22], [50], the distance from Tx or Rx to the first obstacle or building [32], [47], [56], [60], the height of the highest building that blocks the LOS [47], the height of the nearest building to the Rx [32], [56], the number of buildings that obstruct the LOS, statistical information about clutter type and obstructions along the path profile, and terrain elevation, such as the standard deviation and mean height of buildings [13], [15], [20], [21], [24], [32], [46], [53], [56], [60], how much of the LOS is being obstructed by the terrain or buildings, often referred to in the literature as the portion through buildings (indoor distance) or terrain [20], [21], [47], and can be more specific, such as the length of the LOS through each type of clutter (e.g., vegetation, building, road, water, and others), as presented by Ayadi et al. [13], Masood et al. [20], [21], and Braga et al. [35].

Cavalcanti and Cavalcante [48] used as feature a recommendation from ITU-R [61], called terrain clearance angle, which is the angle formed by the Rx antenna in relation to the maximum point of obstruction caused by the terrain.

Losses due to diffraction caused by obstacles in the path profile are also considered and calculated in different ways, such as the Deygout method [13], [62], the cascade knife edge (CKE) [37], [63], or simply considering the distance from the Tx to the first and last diffraction points, as presented by Masood et al. [20], [21]. From the elements of the Fresnel zone, only [35] used the radii of the first Fresnel zone as a feature.

In addition to the path profile, a set of other features can still be extracted by analyzing the LOS formed by the antennas. In the literature, features related to it include the height of the antennas [12], [13], [15], [16], [17], [22], [29], [30], [36], [46], [48], [51], [52], [56], the length of the LOS, also referred to as 3-D distance [30], the ratio or the difference of the antenna heights, angular deviation of antennas [12], [18], [20], [21], [29], [36], [47], [48], [50], [51], [54], and the orientation of

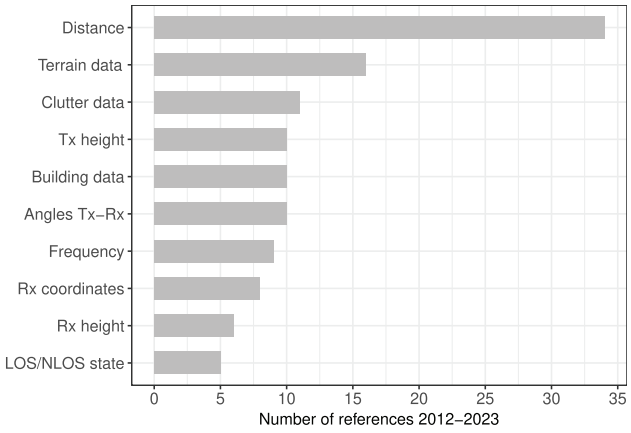


Fig. 1. Ten most frequent features and categories of features used in the literature for path loss prediction between 2012 and 2023.

the Tx antenna relative to the horizontal and vertical planes, defined by the azimuth and downtilt angles, respectively.

6) Environment: According to [54], when information about the geometry and materials of propagation environments is missing, the use of environment features (clutter types) can help describe the characteristics of those environments. Regarding these features, Wu et al. [54] and Zhu et al. [55] used a variety of information from the region of interest, often attributing information such as the type of clutter as presented by Bolli [15], which defines the clutter categories according to the percentage of water, trees, buildings, and other elements contained in the radio environment. Also, Masood et al. [20], [21] used the specific clutter type location of the Tx and Rx as features. The clutter height was used by Ojo et al. [24], Popoola et al. [26], and Ojo et al. [53]. In addition, some authors consider particular characteristics of the region of interest. Gupta et al. [30] considered specific features that characterize street canyons, such as indicating the presence or absence of buildings on both sides of the street and the width of the street. The distinction between the distance of the path and the distance only above the river is taken into account in [35]. In [18] and [33], a categorical feature describing if the region, where data were measured is urban, suburban, or an open area is considered. Also, in [33], another categorical feature is used to define if the propagation wave occurs below or above the rooftop.

7) Technology: Technological parameters have also been considered as features. Examples are the antenna gain [49], [54], the transmit power [54], and the feeder loss [54].

8) Model-Based Elements: Some features are based on the parameters of classical models. For example, [13] defined the set of features according to the parameters used in the standard propagation model (SPM) model [29], [64] used the path loss value obtained by the Okumura-Hata model [6] as a feature, and Ferreira et al. [37] developed their ML approach based on ITU-R P.526 recommendations [63], [65], using values obtained by diffraction models as features.

Fig. 1 presents the ten most used features and categories of features in the literature for received power and path loss prediction via ML approaches. According to it, the distance from Tx to Rx (distance) is the most used feature in the ML approach. The second is a set of features derived from data related to terrain elevation (terrain data), followed by a set of features derived from the environment classification of the region of interest (clutter data), the height of the Tx (Tx height), the set of features derived from information about the radio propagation environment, such as 3-D data (building data), and a set of features that represents antenna tilt and elevation angles formed between the Tx and Rx (angles Tx–Rx). Then, information about the Rx and the frequency bands are often used, including the operating frequency (frequency), the Rx coordinates (Rx coordinates), and the Rx height (Rx height). Finally, the tenth most used is a categorical feature (LOS/non-LOS (NLOS) state) that indicates if the link between the Tx and Rx antennas is in an LOS or in an NLOS.

C. Feature Importance

Several studies in the last decade have proposed propagation prediction models for diverse outdoor scenarios, frequency ranges, and applications. Recent research has focused on assessing the importance of features in ML-based models specifically tailored for urban and suburban networks.

In [32], simulated urban data with 23 features, including XGBoost and RF models, emphasized the significance of features like building information and the distance from the Tx to the tallest building.

For urban and suburban regions, Nguyen and Cheema [33] used real-world data with four features, revealing the distance from the Tx to the Rx as the most influential.

In [30], data from 13 urban street canyons at 28 GHz, employing seven features, highlighted the importance of LOS length and clutter density per street.

Simulated data at 2110 MHz in [21], with 16 features, identified LOS through buildings and Tx–Rx distance as crucial using LightGBM and SHAP analysis.

Real-world 4G data at 1800 and 2600 MHz in [18], with seven features, pinpointed azimuth offset angle and region classification as most influential through RF predictions.

Despite these contributions, limitations like reliance on simulated or limited real-world data, use of specific frequencies, and incomplete feature evaluations persist. Our work aims to bridge these gaps by conducting a comprehensive feature importance analysis using ML models trained on a rich dataset from RSRP measurements performed in diverse urban and suburban scenarios [66]. More information on this dataset will be provided in Section IV-A.

III. Features Description

In this section, we detail a comprehensive set of features for our ML approach to predict RSRP. Considering the factors discussed in Section II influencing feature selection in ML models for path loss prediction, our criteria aim to encompass diverse aspects affecting signal attenuation along the Tx-to-Rx path.

The section is divided into three parts: the first delves into the features from existing literature (generally discussed in Section II), whereas the second introduces new features.

Each feature is presented in an italic font for easy reference throughout this article. Finally, the third part provides a feature categorization based on the three potential data sources. We note that, throughout this article, we use the term Tx to refer to the Tx (base station) and Rx to refer to the Rx.

A. Features From the Literature

1) *Distance From Tx to Rx (Distance):* This feature consists of the Euclidean distance D from Tx to Rx coordinates. It is worth noting that in the dataset used in this article, the largest links between Txs and Rxs are approximately 1000 m. Therefore, when the distance is calculated, it does not regard the Earth's curvature. Let $D = \left\|\mathbf{p}_{\mathrm{Rx}} - \mathbf{p}_{\mathrm{Tx}}\right\|_2$ be the distance between Tx and Rx, where ($\mathbf{p}_{\mathrm{Tx}}$) and ($\mathbf{p}_{\mathrm{Rx}}$) are their position vectors.
2) *Terrain Complexity Measurement:* According to [38], the variability of the terrain elevation, considering the level sea, in the path between Tx and Rx can be defined as the standard deviation of the terrain elevation (*TerrainSD*).
3) Building complexity measurements: similar to the terrain complexity measurements, statistical measures concerning the existing buildings in the path between the Tx and Rx compose these features. This set of features considers the height of all buildings (clutter height) along the path profile. To represent this variability, [56] used the mean (*BuildingMean*) and the standard deviation of the building heights (*BuildingSD*). Also, Uccellari et al. [47] considered the maximum height in the path (*BuildingMax*). To calculate these building-related features, we considered the building height plus the respective elevation at the location of the building. For example, if a building is 80 m and is located 320 m above sea level (elevation), then the height of this building is 400 m. We evaluate all the statistics features above based on the set of these existent heights in the path profile.
4) *Terrain elevation of Tx and Rx antennas (TxElevation, RxElevation):* These two features are the elevation of the terrain at the coordinates of Tx and Rx.
5) *Height of Tx and Rx Antennas (TxHeight, RxHeight):* We consider that the Tx height is defined as the sum of the heights of the antenna tower and the building where the Tx is installed. The Rx height is defined as the elevation at sea level where it was measured.
6) *Effective Height of the Tx and Rx Antennas (TxEffectiveHeight, RxEffectiveHeight):* The effective height of Tx and Rx is defined as the sum of the antenna elevation, tower height, and the building height where it was installed. The building height in the Tx feature is considered because sometimes the base station is located at the top of the building. On the other hand, this does not happen with the Rx because all RSRP measures were done at the ground level.
7) *Difference Between the Effective Tx and Rx Heights (DiffTxRxEffectiveHeight):* This feature represents the difference in the Tx and Rx effective heights in the vertical plane.
8) *Transmission Power (TxPower):* The amount of power used to send information from Tx to Rx.
9) *Downtilt and Azimuth (Downtilt, Azimuth):* The downtilt is the downward angle at which an antenna is tilted from the horizontal plane, whereas the azimuth refers to the angle in degrees between the direction in which an antenna is pointed and a reference direction, which is generally geographic north. If the antenna is perfectly horizontal, the downtilt is 0. If the antenna is pointed directly toward the geographic north, the azimuth is 0. This feature is relevant especially when the data are from directional antennas.
10) *Frequency (Frequency):* This feature consists of the frequencies used in the communication system.
11) *LOS length (LOSDistance):* This feature means the distance from Tx-to-Rx antennas relative to the LOS, also called 3-D distance [30].
12) *Angular Difference Between Tx and Rx (AngularDifferenceTxRx):* The angular difference in degrees between the Tx and Rx, is also interpreted as the elevation angle between the Tx and Rx. It is calculated by comparing the direct horizontal distance between Tx and Rx with the LOS distance that takes into account any obstacles between them. The result gives an angle that represents how much the signal path deviates vertically from a straight horizontal line.
13) *Portion Through Building (PTB):* It is defined concerning the direct ray formed by Tx and Rx and passes inside the buildings. Also defined as indoor distance [20], [21]. Let B be the total number of segments of the direct Tx–Rx ray that pass through buildings, each of length d_b for (b = 1, . . . , B). The *PTB* gives the percentage of the overall distance D that lies inside the buildings

$$PTB = 100 \times \frac{\sum_{b=1}^{B} d_b}{D}. \tag{1}$$

14) *Portion Through Terrain (PTT):* Similar to *PTB*, *PTT* is defined concerning the direct ray formed by the Tx and Rx that passes through the terrain, also known as portion through ground [47]. Let T be the total number of segments of the Tx–Rx ray that traverse terrain (ground), each of length d_t for ($t = 1, \ldots, T$). The *PTT* is define as follows:

$$PTT = 100 \times \frac{\sum_{t=1}^{T} d_t}{D}. \tag{2}$$

15) *Portion Through Clutter (PTC_i):* It is defined in relation to the direct ray formed by Tx and Rx that passes through each type of clutter. It is also known as crossed distance [13]. For each clutter type i (out of $\mathrm{N}_{\mathrm{clutter}}$ types), let C_i be the number of segments that intersect clutter i, each length d_c^i for ($c = 1, \ldots, C_i$) The PTC_i represents the percentage of D that crosses clutter type i

$$PTC_i = 100 \times \frac{\sum_{c=1}^{C_i} d_c^i}{D} \quad (i = 1, \ldots, N_{\mathrm{clutter}}). \tag{3}$$

The types of clutter may vary depending on the quality and quantity of geographic information obtained from

the region of interest. The most common categories of clutter found in the literature are the characteristics of the building's density, vegetation types, roads, and water [15]. In this work, we consider 20 types of clutter, meaning 20 features.

16) *Portion Through Free Space (PTFS):* It is defined as the portion of the direct ray, formed by Tx and Rx, in the free space (without obstructions by buildings and terrain), also known as outdoor distance [20], [21]. The *PTFS* is defined in the following equation:

$$PTFS = 100 - (PTB + PTT). \tag{4}$$

17) *Distance From Tx to the Nearest Obstruction Caused by Building (DistanceTxObsBuilding):* The distance from Tx to the nearest obstruction in the path between Tx and Rx considering the building.

18) *Distance From Rx to the Nearest Obstruction Caused by Building (DistanceRxObsBuilding):* The distance from Rx to the nearest obstruction in the path between Tx and Rx considering the building. For the features *DistanceTxObsBuilding* and *DistanceRxObsBuilding*, if there is no obstruction in the path between Tx and Rx, the feature value is considered the distance from Tx to Rx.

19) *LOS or NLOS State (LOSIndicator):* It is a binary variable used to indicate whether there are obstructions in the LOS. In this sense, if the LOS is obstructed, the value assigned to the feature is 1.

20) *Radius of the First Fresnel Zone (FresnelRadii):* The radius of the first Fresnel zone is calculated according to the following equation [35]:

$$r = \sqrt{\frac{d_1 d_2 \lambda}{d_1 + d_2}} \tag{5}$$

where λ is the wavelength, d_1 is the distance from Tx to the middle of the path profile, and d_2 is the distance from the middle of the path profile to Rx. Also, the values of d_1 and d_2 may be defined as the distance from Tx and Rx to a particular obstacle anywhere in the LOS. In this work, we define $d_1 = d_2 = (D/2)$.

21) *Percentage of Obstruction in the First Fresnel Zone (FresnelObs):* This feature is defined by comparing the tallest obstruction in the path profile and the diameter of the first Fresnel zone ($2r$) (5). For instance, when the tallest obstruction blocks the signal in the LOS, the obstruction percentage is equal to 50%. According to [67], if the Fresnel zone is obstructed up to 40%, the signal transmission will not have attenuation due to the incidence of the obstacles. The only work we found in the literature that considered information from the Fresnel zones was proposed by Braga et al. [35], which used only the radius of the first Fresnel zone as a feature.

22) *Terrain Clearance Angle (TCA):* It is the angle formed by the Rx antenna with respect to the highest point of the terrain (TCA). If the obstruction is at the same level or lower than the Rx, the TCA will be equal to 0°. The TCA is from ITU-R P.1546 recommendation [61] and used by Cavalcanti and Cavalcante [48] in their ML approach.

23) *Diffraction Loss Based on CKE (DiffractionLoss):* It considers the loss caused by diffraction that may occur along the path between Tx and Rx. To calculate the diffraction loss, we use the CKE model from the recommendation of ITU-R P526-12 [63]. According to [37], the CKE model has better results compared with another diffraction model from ITU-R P526-12 called Delta-Bullington [65]. To calculate the diffraction loss that occurs in the path between Tx and Rx, it is necessary to identify the main obstacle in the path and then divide the path into two segments. Tx and the point of the main obstacle will define the first segment, whereas the point of the main obstacle and Rx form the second segment. All geometrical information in these segments is calculated in a single parameter v, also known as the Fresnel–Kirchhoff diffraction parameter, obtained as follows:

$$v = h\sqrt{\frac{2}{\lambda}\left(\frac{1}{d_1'} + \frac{1}{d_2'}\right)} \tag{6}$$

where h is the height of the main obstacle in the LOS and could be positive if the top is above the LOS or negative; otherwise, d_1' is the distance from Tx to the point of the main obstacle, d_2' is the distance from the point of the main obstacle and Rx, and λ is the wavelength. The diffraction loss, using the parameter v, can be calculated as follows:

$$J(v) = -20log \times \left(\frac{\sqrt{[1 - C(v) - S(v)]^2 + [C(v) - S(v)]^2}}{2}\right) \tag{7}$$

$$S(v) = \int_0^v \sin\left(\frac{\pi s^2}{2}\right) ds \tag{8}$$

$$C(v) = \int_0^v \cos\left(\frac{\pi s^2}{2}\right) ds. \tag{9}$$

In (7), $C(v)$ and $S(v)$ are the real and imaginary parts, respectively, of the complex Fresnel integral. To calculate the CKE, in this article, we follow the same process as suggested by Ferreira et al. [37]. After obtaining the value of the diffraction loss, named $J_1(v)$, caused by the main obstacle, the same procedure to obtain $J_1(v)$ is repeated to calculate the diffraction losses $J_2(v)$ and $J_3(v)$ that is obtained from the new segments created from the two first segments of the main obstacle. Then, the CKE loss feature is the sum of $J_1(v)$, $J_2(v)$, and $J_3(v)$.

B. Proposed Features

We describe next a set of features proposed in this article that focus on improving RSRP prediction.

1) *Terrain Complexity Measurements:* We propose a set of statistical measures to expand the representations of the terrain elevation complexity, which is only utilized in the literature by Oroza et al. [38] through the standard deviation of elevation. Each of the following

measures is considered a feature in our ML approach. To assess the terrain irregularity along the path between Tx and Rx, we define the mean (*TerrainMean*), median (*TerrainMedian*), first quartile (*Terrain1q*), third quartile (*Terrain3q*), the skewness (*TerrainSkewness*), and the kurtosis (*TerrainKurtosis*) of the terrain elevation. In addition, similar to the analysis of buildings in the path from features proposed in the literature, we have defined the maximum (*TerrainMax*) and minimum (*TerrainMin*) values of the terrain elevation as features. As shown in [47], even in urban scenarios where the terrain is not typically considered as an attenuation factor, it is still relevant to consider obstructions caused by terrain.

2) Building complexity measurements: similar to the terrain complexity measurements, we propose a set of statistical measurements to enhance the representation of the building's distribution along the path profile. In addition to those presented by Juang [56], which used the mean and the standard deviation of the building's height, we also define as features the median (*BuildingMedian*), first quartile (*Building1q*), third quartile (*Building3q*), the skewness (*BuildingSkewness*), and the kurtosis (*BuildingKurtosis*) of the buildings. In addition to the maximum height of buildings, used as a feature by Uccellari et al. [47], we also consider the minimum value of the building height as a feature (*BuildingMin*).

3) *Difference Between Tx and Rx Coordinates (DeltaCoordX, DeltaCoordY):* These features are the absolute value of the subtraction of X- and Y-coordinate values of Tx and Rx. These features replace the latitude and longitude coordinates of Tx and Rx, previously explained in Section II-B.4.

4) *Building Clearance Angle (BCA):* This feature is proposed in this work as a variation of the ITU-R P.1546 recommendation [61], TCA. The BCA considers the obstruction caused by the highest point of the buildings instead of the maximum point of the terrain.

5) *Effective Height Differences:* These features concentrate on the differences in the effective height of the Tx and Rx antennas related to the highest building and the highest terrain elevation. Each of the following differences is considered a feature in our ML approach.
   a) Difference between the effective Tx height and the maximum terrain point in the path profile (*DiffTxEffectiveHeightTerrainMax*).
   b) Difference between the effective Tx height and the maximum building point in the path profile (*DiffTxEffectiveHeightBuildingMax*).
   c) Difference between the effective Rx height and the maximum terrain point in the path profile (*DiffRxEffectiveHeightTerrainMax*).
   d) Difference between the effective Rx height and the maximum building point in the path profile (*DiffRxEffectiveHeightBuildingMax*).

6) *Distance From Tx to the Nearest Obstruction Caused by Terrain (DistanceTxObsTerrain):* It corresponds to the distance from Tx to the nearest obstruction in the path between Tx and Rx considering the terrain. This feature is created based on the feature used in [47] and [60], which considered only obstruction caused by buildings.

7) *Distance From Rx to the Nearest Obstruction Caused by Terrain (DistanceRxObsTerrain):* It corresponds to the distance from Rx to the nearest obstruction in the path between Tx and Rx considering the terrain. This feature is created based on the feature used in [47] and [56], which considered only obstruction caused by building. For the features *DistanceTxObsTerrain* and *DistanceRxObsTerrain*, if there is no obstruction in the path between Tx and Rx, the feature value is considered as the distance from Tx to Rx.

8) *Number of Obstructions in the LOS (NumObsLOS):* It considers how many obstacles occasioned by buildings or terrain are blocking the LOS. Here, we sum up the number of blocking incidences along the path profile. This feature was created based on the feature used in [21] and [32], which considered only the sum of the obstructions caused by the building.

C. Three Major Sources of Features

The features defined for our study are categorized into three major groups: Tx- and Rx-related, buildings, and terrain information, as illustrated by a Venn diagram in Fig. 2. Tx and Rx information include technological parameters, coordinates, and LOS. Terrain data involves features related to terrain complexity, such as ground elevation statistics. Building data provides features related to building complexity, such as building height statistics, and characteristics of the outdoor region.

We combined data from these three sources to create additional features. For instance, merging Tx and Rx data with terrain data results in features like the length of the LOS obstructed by terrain and terrain elevation at Tx and Rx locations. Combining Tx and Rx data with building data creates features like the length of the LOS obstructed by buildings and the BCA. The Venn diagram not only illustrates these relationships but also aids in generating new features for future studies and provides a method to adapt when information from one of the three sources is unavailable, such as lacking a digital surface map (DSM) or LiDAR data in certain urban regions.

IV. Data Processing

A. Dataset Description

The data used in the experiments belong to the well-known Huawei dataset that is available in [66], which was first used in [68] and that it was presented in the context of a mathematical modeling competition. Other authors have used this dataset [69], [70]; however, the literature does not provide information on the cities or the service providers the data was extracted from. The dataset is remarkable because of the large number of base stations (4000 Tx) and RSRP measurements (12011833), the fact that covers roads, urban and suburban areas of different types as well as for providing many other geographical and context details.

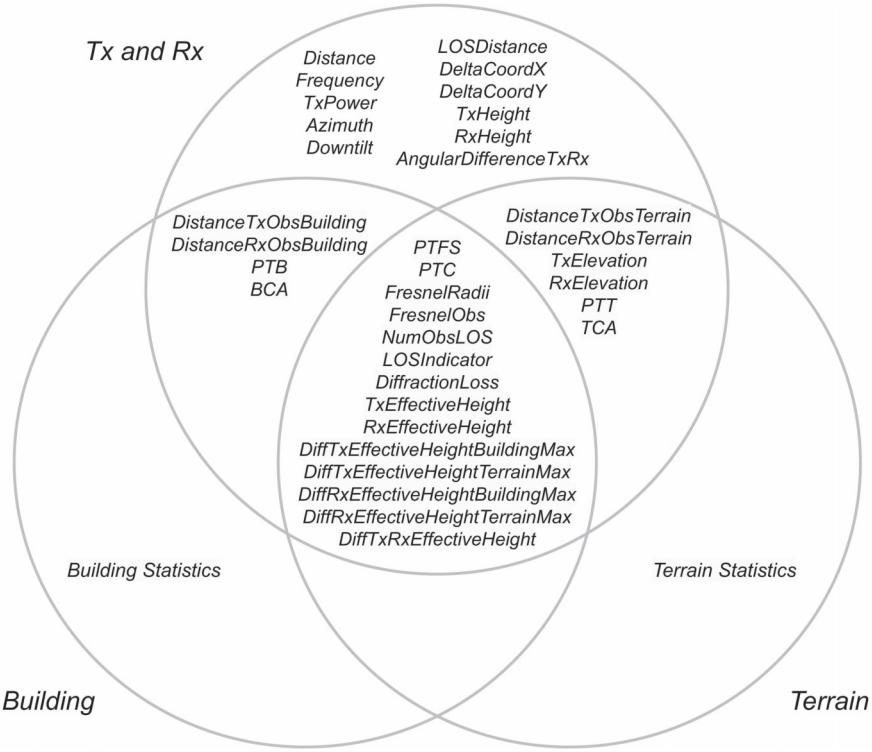


Fig. 2. Venn diagram of the features.

TABLE I
RAW DATA MEASUREMENTS

Description	Value	Unit
Tx index	150445	-
Tx coordinate X	4894894	-
Tx coordinate Y	4894894	-
Azimuth	300	degree
Frequency	2585	MHz
Transmission Power	18	dBm
Electrical Downtilt	3	degree
Mechanical Downtilt	8	degree
Tx tower height	35	m
Tx altitude	480	m
Tx building height	28	m
Tx clutter index	3	-
Rx coordinate X	8484565	-
Rx coordinate Y	8484566	-
Rx altitude	520	m
Rx clutter index	6	-
RSRP	-84.5	dBm

In fact, information on the propagation environment are obtained for each grid, including terrain elevation (relative to sea level), building height, clutter index, geographic coordinates, and technological parameters of the respective Tx. Table I presents the information contained in the original dataset and their respective units.

In Table I, we show an instance of raw data measurements for a Tx and Rx couple to provide the reader with the type of information contained in the database. Based on that information, we propose a methodology to extract the links (path profile) according to the location of the measured points, enabling the analysis of the communication between Tx and Rx in a continuous way. With the creation of the path profile, we can analyze the variation of the terrain elevation, and the variation of the buildings, and mainly verify obstructions that may exist in the LOS, as well as assist in creating features for the ML model.

B. Path Profile and Feature Extraction

To extract the path profile, we create the LOS between the Tx and Rx, achieved by tracing a line segment whose endpoints are the coordinates of the Tx and Rx. Then, we identify and extract all grids touching the LOS to obtain the information along the path profile. All this process was implemented in the R language using the package *raster* [71]. This procedure is illustrated in Fig. 3.

After performing the extraction of the path profile information illustrated in Fig. 3, we have a path profile instead of individual measurements as in the original dataset. In Fig. 4, the path profile is illustrated after extracting the information from the grids belonging to the path between Tx and Rx.

The path profile extraction procedure was limited to 150 h of processing, resulting in a new dataset with 186 635 valid path profiles (RSRP) from 326 Tx (base stations). Then, from these path profiles, we are able to extract all features described in Section III for our feature importance analysis.

C. Feature Statistical Analysis

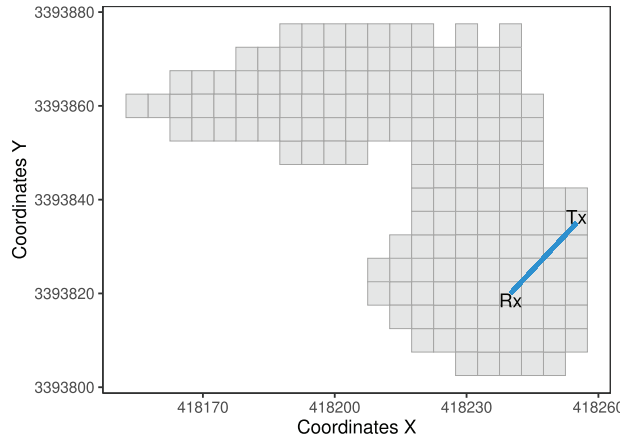
Table II reports for all 72 extracted features, their range, and mean values. For better comprehension, we organize the table following the same feature sequence presented in Section III, starting from the left column to the right column.

TABLE II
STATISTICS OF THE FEATURES AND TARGET

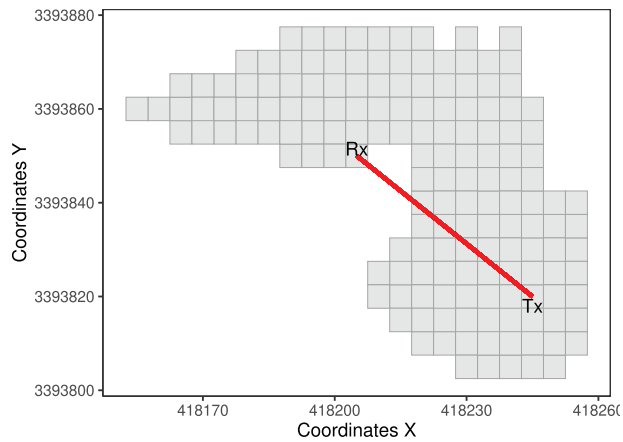
Feature	Min	Max	Mean	Unit	Feature	Min	Max	Mean	Unit
Distance	7.1	1086.4	129.7	m	*Downtilt*	0.0	27.0	9.8	degree
TerrainMean	470.8	546.6	504.9	m	*Azimuth*	0.0	350.0	186.2	degree
TerrainMedian	471.0	547.0	504.9	m	*Frequency*	2585.0	2624.6	2586.8	MHz
TerrainSD	0.0	3.2	0.6	m	*LOSDistance*	7.1	1086.5	135.2	m
Terrain1q	470.0	546.0	504.5	m	*AngularDifferenceTxRx*	3.2	90.0	71.3	degree
Terrain3q	471.0	547.0	505.3	m	*NumObsLOS*	0.0	15.0	0.5	-
TerrainSkewness	-8.4	7.5	0.0	m	*PTB*	0.0	89.5	4.2	%
TerrainKurtosis	-2.4	69.1	0.1	m	*PTT*	0.0	97.2	1.3	%
TerrainMax	472.0	548.0	505.7	m	PTC_1	0.0	0.0	0.0	%
TerrainMin	470.0	545.0	504.1	m	PTC_2	0.0	100.0	0.8	%
BuildingMean	470.8	628.0	507.3	m	PTC_3	0.0	0.0	0.0	%
BuildingMedian	471.0	628.0	505.9	m	PTC_4	0.0	0.0	0.0	%
BuildingSD	0.0	131.3	4.3	m	PTC_5	0.0	100.0	50.9	%
Building1q	470.0	628.0	504.9	m	PTC_6	0.0	100.0	19.7	%
Building3q	471.0	628.0	508.4	m	PTC_7	0.0	100.0	4.1	%
BuildingSkewness	-8.4	9.0	1.0	m	PTC_8	0.0	100.0	1.2	%
BuildingKurtosis	-2.4	80.2	2.3	m	PTC_9	0.0	0.0	0.0	%
BuildingMax	472.0	835.0	517.1	m	PTC_{10}	0.0	100.0	1.0	%
BuildingMin	470.0	628.0	504.2	m	PTC_{11}	0.0	100.0	0.7	%
TxHeight	0.0	127.0	27.3	m	PTC_{12}	0.0	100.0	7.3	%
RxHeight	0.0	0.0	0.0	m	PTC_{13}	0.0	100.0	10.4	%
TxElevation	472.0	545.0	504.9	m	PTC_{14}	0.0	100.0	2.9	%
RxElevation	470.0	548.0	504.9	m	PTC_{15}	0.0	100.0	1.1	%
TxEffectiveHeight	489.0	659.0	532.2	m	PTC_{16}	0.0	38.5	0.1	%
RxEffectiveHeight	470.0	548.0	504.9	m	PTC_{17}	0.0	30.8	0.1	%
DiffTxRxEffectiveHeight	-4.0	127.0	27.3	m	PTC_{18}	0.0	0.0	0.0	%
DiffTxEffectiveHeightTerrainMax	-127.0	4.0	-26.5	m	PTC_{19}	0.0	0.0	0.0	%
DiffTxEffectiveHeightBuildingMax	-42.0	297.0	-15.1	m	PTC_{20}	0.0	0.0	0.0	%
DiffRxEffectiveHeightTerrainMax	0.0	8.0	0.8	m	*PTFS*	2.8	100.0	94.5	%
DiffRxEffectiveHeightBuildingMax	0.0	340.0	12.2	m	*LOSIndicator*	0.0	1.0	0.3	-
DistanceTxObsBuilding	3.5	1086.4	115.5	m	*FresnelRadii*	0.4	5.6	1.8	m
DistanceTxObsTerrain	3.5	1086.4	124.4	m	*FresnelObs*	0.0	100.0	36.7	%
DistanceRxObsBuilding	3.5	1086.4	101.6	m	*TCA*	0.0	28.8	0.4	degree
DistanceRxObsTerrain	3.5	1086.4	118.3	m	*BCA*	0.0	85.2	9.8	degree
DeltaCoordX	0.0	1085.0	88.6	-	*DiffractionLoss*	-73.7	385.9	30.4	dB
DeltaCoordY	0.0	520.0	73.8	-	*RSRP*	-139.8	-53.7	-85.7	dBm
TxPower	3.2	18.2	11.9	dBm	-	-	-	-	-

From the feature's statistics presented in Table II, we provide the following highlights.

1) *Features Removed:* As we identified features with minimum and maximum values that are the same, we removed them from the dataset. It is because features with the same value for all samples do not have any impact on the learning process. The removed features are: *RxHeight*, PTC_1, PTC_3, PTC_4, PTC_9, PTC_{18}, PTC_{19}, and PTC_{20}.
2) *Length of the Path Profile:* Regarding the path profile length (the distance from the Tx to Rx), it is essential to emphasize that the longest path in our dataset is approximately one kilometer, whereas the mean length is around 130 m. This characteristic relates to the nature of the mobile network and the outdoor scenario where the data were collected.
3) *Diversity of the Environment:* Although the scope of this work is the outdoor scenario in urban and suburban regions, we can observe a variety of characteristics that occur along the path profile between Tx and Rx. This variety can be verified by the mean values of the features PTC_i, where each index i represents an environmental characteristic. For example, on average, all path profiles are 50% composed of urban development areas ($i = 5$), 20% of roads ($i = 6$), 10% of urban high-density buildings ($i = 13$), and 7% of urban medium–high buildings ($i = 12$). In addition, some natural characteristics are found, with 4% of vegetation cover area ($i = 7$), as well as 1% of shrubs ($i = 8$) and 0.8% of inland lakes ($i = 2$).
4) *Portion Through Obstacles:* As expected in urban and suburban areas, obstructions in the LOS are predominantly caused by various buildings. Besides analyzing the type of clutter, we can verify the features *PTT* and *PTB*, where on average, the path profiles suffer more from obstacles deriving from buildings than terrain variation.
5) *Technological Parameters:* Regarding the technological parameters, we noticed a slight variation in the operating frequency of the networks, with the minimum frequency equal to 2.585 GHz and the maximum frequency equal to 2.624 GHz. Although there is a slight variation, we decided to keep the feature in the dataset as suggested in the literature. There is also a considerable variation in the transmission power, with the minimum power equal to 3.2 dBm and the maximum power equal to 18.2 dBm. This factor should also be considered since



(a)



(b)

Fig. 3. Gray area shows 5 × 5 m grids where RSRP measurements were taken, while the white area indicates where no measurements occurred. Together, they outline a Tx's coverage area. In (a), a blue line representing a path between Tx and Rx intersects with gray grids, from which data are extracted. In (b), a red line path lacks intersecting information-filled grids, making it invalid and excluded from the final dataset. (a) Feasible path. (b) Infeasible path.

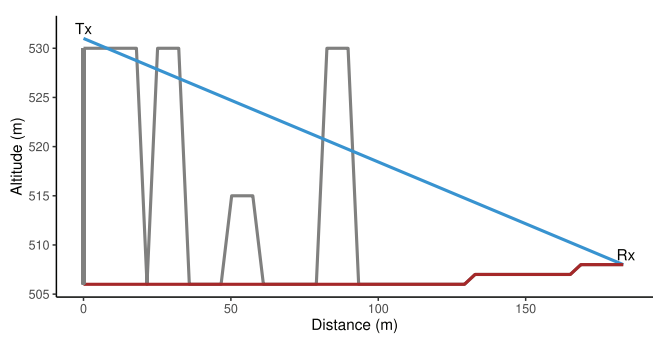


Fig. 4. Distance (m) axis shows the metered space between Tx and Rx, and the Altitude (m) axis indicates meters above sea level. The Tx is located in the top left, and the Rx is in the bottom right. The blue line illustrates the LOS, the brown line indicates the terrain elevation relative to sea level, and the gray line represents buildings and obstacles.

the amplitude and range of the main beam are determined based on the power and other parameters, such as the antenna gain. Finally, the average height of the Tx antennas is 27 m. It is worth noting that this value refers only to the height of the Tx antenna and does not consider the elevation or the building height in cases where the Tx is located on top of a building.

After all the manipulations in the original dataset, we updated it to make it ready for use in the features importance study for RSRP prediction. The updated dataset has 64 features, with one target and 186 635 samples.

V. Methodology

In this section, we present our methodology to evaluate feature importance for RSRP prediction. It includes the data preparation for training the used RF model as well as its testing pipeline. We also show the error metrics used to evaluate the performance of the ML model.

A. RF and Feature Importance

RF [72] is an ML model based on decision trees that can be used for classification and regression problems. The main difference, when compared with the classical decision tree, is that RF uses and combines not one but multiple decision trees to make predictions. In a regression problem, the target variable is continuous, and the objective is to predict its value given a set of input features. In this work, the target RSRP is a continuous variable measured in dBm, and the set of features is composed of radio environment characteristics and technological parameters.

There are several compelling reasons behind our choice of an RF model as the preferred ML approach for the present work. First, RF has established itself as a highly effective method for predicting path loss, as evidenced by its extensive application in recent studies [10], [18], [19], [21], [28], [30], [32], [50], [52]. Besides that, RF has presented superior regression performance regarding RSRP prediction when compared with other ML models, e.g., support vector regression (SVR) [10], [28], artificial neural network (ANN) [10], [19], adaptive boosting (AdaBoost) [52], and K-nearest neighbors (KNNs) [52]. For an in-depth exploration of the advantages of using RF in various RSRP prediction scenarios, a comprehensive discussion is available in the study conducted by Seretis et al. [11].

The RF algorithm combines multiple decision trees, indexed by $t = 1, \ldots, T$. For a training dataset with N samples and M features, each tree is trained on a different subset of the training dataset selected at random (with replacement). The selected subset is called a bag, and the left-out samples are called out-of-bag (OOB) samples. The trees are then trained on different bags, and later the prediction from all the trees are aggregated (typically by average).

With RF, it is also possible to compute the predictive importance of each of the M features by averaging the increase in the error made by a tree when the values of each feature are randomly permuted in the OOB samples.

A typical feature importance measure used with RFs is the incremental mse (IncMSE), which is computed for each feature $m = 1, \ldots, M$ as follows:

$$\text{IncMSE}(m) = \frac{1}{T} \sum_{t=1}^{T} \left(\text{MSE}(\text{OOB}^{*}_{t,m}) - \text{MSE}(\text{OOB}_{t}) \right) \quad (10)$$

where OOB_t refers to the out-of-bag sample used to train the tree t and $OOB_{t,m}$ refers to the same OOB but where the values of feature m are randomly permuted in the sample. The mean squared error (mse) of an OOB sample is calculated as the average of the squared differences between the actual target values and the predicted values for all the samples in the OOB.

The higher the value of % IncMSE, the more important the feature m is for the prediction. Conversely, if the increase in error is relatively low or does not show any difference, then the feature is considered irrelevant.

The RF used in this article is implemented in the *R* package *randomForest* [73] with $T = 300$ and a split of (63%, 37%) between the bags and their associated OOB samples. In Sections V-B–V-D, we will describe the training strategies, the tuning of RF hyperparameters, the performance metrics, and the obtained prediction results.

B. Training and Testing Sets

To perform the training and testing of the RF algorithm, we divided the dataset presented in Section IV-A into two sets. The purpose of this division is to ensure that the sample data from the test set is not used in the algorithm training, allowing for the evaluation of the predictive model's performance. The division of the set was performed with the holdout technique in the 80/20 proportion, as used in [10], [30], and [32]. Thus, the training set was created with 80% of the data, with 149 311 measured points in total, while the remaining 20% formed the test set composed of 37 324 measured points.

C. Tuning

RF has two relevant hyperparameters. The first is the size of the features subset randomly selected in each node split, named *mtry*. The second parameter is the number of trees in the forest, called *ntree*. We adjusted these hyperparameters by grid search, ranging *mtry* from 1 to 64 with a step of 4, and *ntree* from 100 to 500 with a step of 100. Our tests allowed us to set *mtry* = 16 and *ntree* = 300.

D. Error Metrics and Prediction Results

Our RF model was evaluated regarding the regression task using root mse (RMSE), mean absolute percentage error (MAPE), and Pearson correlation coefficient (r). RMSE was chosen because it provides the final predictor error in the unit of the target attribute, which in our case is dBm. However, RMSE penalizes the difference between predicted and actual when this difference is high. Hence, in addition to RMSE, we also used the MAPE metric for complementary evaluation. The Pearson correlation coefficient measures the strength and direction of the linear relationship between predicted and actual target values, ranging from –1 to 1. A value of 1 indicates a perfect positive linear relationship, –1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(Y_j - \hat{Y}_j\right)^2} \tag{11}$$

$$\text{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{Y_j - \hat{Y}_j}{Y_j}\right| \times 100 \tag{12}$$

$$r = \frac{\sum_{j=1}^{N}\left(Y_j - \overline{Y}\right)\left(\hat{Y}_j - \overline{\hat{Y}}\right)}{\sqrt{\sum_{j=1}^{N}\left(Y_j - \overline{Y}\right)^2 \sum_{j=1}^{N}\left(\hat{Y}_j - \overline{\hat{Y}}\right)^2}}. \tag{13}$$

In (13), $\overline{Y}$ and $\overline{\hat{Y}}$ are the average of the actual target variable and the predicted value mean, respectively.

The obtained results of the RF with the test set were an RMSE of 3.58 dB, an MAPE of 2.9%, and an r of 0.92. According to [16] and [22], the RMSE equal to 3.58 dB is within the accepted prediction error range for urban scenarios, which is between 7 and 8 dB. Also, a correlation coefficient of 0.92 indicates a strong positive relationship between the predicted and actual values.

VI. Feature Importance Analysis

The main objective of our study is to evaluate the importance of each feature in the learning process. In the Section IV-A, we will present an analysis of the feature importance as well as an analysis of the error metrics when the predictions are not made with all features listed in Section II, but using a different set of features according to their nature.

As explained in Section V-A, it is possible to analyze the importance of each feature in predicting the RSRP according to the values obtained from the mse increase when the feature values are randomly permuted in the tested sample. The feature importance obtained by our RF model is presented in Table III. The values of %IncMSE are shown in descending order from the most important to the least important feature. We also show the cumulative %IncMSE, labeled as %IncMSE C.

We observe from Table III that the following hold.

1) The features that contribute the most to the RSRP prediction is mainly related to the Tx, which includes technical parameters such as downtilt, azimuth, Tx height, and transmission power.
2) The most important features are obtained from the combination of Tx and environment data, such as Tx effective height, the differences between the effective height of the Tx and the maximum point of the terrain and the building, Tx elevation, and the mean elevation of the terrain. To complete the group of the ten most important features in predicting the RSRP, we have the feature related to the effective height difference between Tx and Rx.
3) The two most important features (downtilt and azimuth), which are parameters related to the antenna orientation, exhibit a %IncMSE of approximately 40% each. These results could be explained by having several scenarios from the 326 Txs with directional antennas in the dataset. Also, the experiments performed by Azevedo et al. [74] showed that multiple reflected and diffracted signals reach the Rx at different azimuth angles from the Tx direction. While in the results presented by [75], different downtilt angles have distinct effects on the received power. The same argument could

TABLE III
FEATURE IMPORTANCE RANK

Feature	%IncMSE	%IncMSE C	Rank	Feature	%IncMSE	%IncMSE C	Rank
Downtilt	41.16	6.23	1	*Building1q*	8.32	82.17	33
Azimuth	40.31	12.33	2	*Building3q*	8.00	83.38	34
TxEffectiveHeight	28.09	16.58	3	*BCA*	7.86	84.57	35
TxHeight	25.25	20.40	4	PTC_5	7.82	85.75	36
TerrainMean	23.42	23.94	5	*BuildingMedian*	7.55	86.89	37
TxPower	23.01	27.43	6	*DistanceRxObsTerrain*	6.70	87.90	38
DiffTxEffectiveHeightTerrainMax	22.13	30.78	7	*PTFS*	6.66	88.91	39
TxElevation	19.62	33.75	8	*RxElevation*	6.50	89.90	40
DiffTxRxEffectiveHeight	19.58	36.71	9	*RxEffectiveHeight*	6.15	90.83	41
DiffTxEffectiveHeightBuildingMax	18.12	39.45	10	*DistanceRxObsBuilding*	5.52	91.66	42
BuildingMax	17.70	42.13	11	PTC_{13}	5.51	92.49	43
AngularDifferenceTxRx	15.77	44.51	12	PTC_{12}	5.48	93.32	44
LOSDistance	15.60	46.87	13	*BuildingSkewness*	4.81	94.05	45
FresnelRadii	15.44	49.21	14	*DiffractionLoss*	4.65	94.76	46
FresnelObs	14.83	51.46	15	*TerrainSkewness*	4.18	95.39	47
BuildingSD	14.77	53.69	16	*NumObsLOS*	3.86	95.97	48
DiffRxEffectiveHeightBuildingMax	14.28	55.85	17	*TerrainKurtosis*	3.69	96.53	49
Distance	14.09	57.98	18	*BuildingKurtosis*	3.60	97.07	50
DistanceTxObsTerrain	13.73	60.06	19	PTC_7	3.43	97.59	51
BuildingMean	13.42	62.09	20	*DiffRxEffectiveHeightTerrainMax*	3.17	98.07	52
TerrainMax	12.80	64.03	21	*LOSIndicator*	3.09	98.54	53
DeltaCoordY	12.14	65.86	22	PTC_{14}	2.38	98.90	54
TerrainSD	11.87	67.66	23	*TCA*	1.85	99.18	55
TerrainMin	10.52	69.25	24	PTC_{11}	1.29	99.37	56
DeltaCoordX	10.29	70.81	25	PTC_2	1.08	99.54	57
Terrain3q	10.13	72.34	26	PTC_{10}	0.93	99.68	58
PTC_6	9.98	73.85	27	*Frequency*	0.80	99.80	59
Terrain1q	9.97	75.36	28	PTC_{17}	0.46	99.87	60
PTB	9.44	76.79	29	*PTT*	0.39	99.93	61
TerrainMedian	9.20	78.18	30	PTC_{15}	0.36	99.98	62
DistanceTxObsBuilding	9.13	79.56	31	PTC_8	0.09	100.00	63
BuildingMin	8.90	80.91	32	PTC_{16}	0.02	100.00	64

be considered for the antenna height and transmission power features, which have %IncMSE equal to 25% and 23%, respectively.

4) Besides considering a variety of antenna parameters, when we add the terrain elevation and building information to the dataset, the features that use these related data significantly impact the prediction model. In particular, the feature *TxEffectiveHeight*, constructed with the elevation, antenna height, and building height, is more important than the feature *TxHeight*.
5) Data enrichment enables to identify the main obstacles in the path between the Tx and Rx and improves the analysis of the different elements in the radio propagation environment. Some of these elements could be observed in the most relevant features after the top ten most important ones, including the radius of the first Fresnel zone and its percentage of obstruction, most of the statistics from both terrain and buildings, Tx and Rx antenna angular difference as well as the length of the LOS (*LOSDistance*).
6) The distance from the Tx to Rx and from the Tx to the first obstruction occasioned by terrain is significant in the prediction.
7) The least important features for the prediction model are related to the portion through the clutter (PTC_i), followed by *PTT*, *PTFS*, the features representing the skewness and kurtosis of the terrain elevation and buildings, the ITU recommendations (*DiffractionLoss* and TCA), BCA, and the number of obstacles that block the LOS. Also, except for the difference between the effective height of the Rx and the maximum building, the other features related to the Rx do not show significant improvement in the quality of the predictor.
8) Notably, the feature used to indicate if the link is LOS or NLOS does not significantly improve the model performance. Therefore, there is no necessity to classify all links in the dataset. Although the frequency is among the features that contribute less to prediction, if the frequency range of our dataset covered a broader spectrum, for example, ranging from megahertz to gigahertz, the importance of this feature could be different, as shown in [33].

A. Reduction of Model Complexity

In this section, we will explore the results obtained from the feature importance, aiming to reduce the complexity of the prediction model regarding extracting features while maintaining the quality of the predictor. As mentioned earlier, disregarding relevant features or keeping unrelated ones can lead to a low-quality predictor [10]. Therefore, besides reducing the number of features, we will focus on selecting features that do not increase the error significantly, evaluating the RMSE increase.

TABLE IV
REGRESSION PERFORMANCE OF EVALUATED REGRESSION MODELS

Models	RMSE (dB)	MAPE (%)	r
Set-6	6.60	6.1	0.71
Set-10	5.32	4.7	0.82
Set-15	4.23	3.5	0.89
Set-20	4.06	3.4	0.90
Set-25	3.72	3.0	0.91
All	3.58	2.9	0.92

To develop and analyze the model complexity reduction, we create five different sets of features taking into consideration: 1) important value of each feature identified by our RF model and presented in Table III and 2) the feature categorization described in Section III. We describe each new set of features as *set-number of features*. For example, *Set-6* is a set with six features. These sets are listed as follows.

1) *Set-6* uses only the first six most important features, representing the technological parameters, the height of the Tx, and the elevation terrain average as a measure of terrain complexity.
2) *Set-10* uses the same features as *Set-6* plus four features that represent information about the radio environment, including building and terrain maximum values and the Tx effective height.
3) *Set-15* is composed of the *Set-10* features plus features related to the first Fresnel zone, such as the radio and the percentage of obstruction, which implies the necessity of detailed information in the path profile between the Tx and Rx as well as a distance measure.
4) *Set-20* comprises the most important 15 features with other building statistics and another distance measure.
5) *Set-25* is composed of 25 features, the 20 used in *Set-20* plus features related to Tx and Rx localization and other terrain statistics.

Five different RF models are trained according to the methodology presented in Section V. They are compared regarding the regression error measures RMSE and MAPE and the correlation coefficient against the ML approach, denoted *All*, that corresponds to the RF trained over the whole set of features. These results are reported in Table IV. Fig. 5 shows the convergence of the predicted error according to the cumulative %IncMSE values from Table III, using different sets of most important features presented in Table IV.

According to Fig. 5 and the second column in Table IV, we now compare the performance, using the RMSE, between the prediction obtained using a set composed of all features and each of the reduced set of features.

1) The primary result that we highlighted is that all prediction results using the reduced number of features are worse than the results obtained using all features. There is an increase in the RMSE from *Set-25* to *Set-6* compared with the set with all features. More specifically, the increase in the prediction error is from 0.14 to 3.02 dB, respectively.
2) When we reduce the set of features from 64 to only six features that represent the technological parameters, the Tx height, and one measure of terrain complexity (*Set-6*), the increase in the RMSE is about 3 dB. Considering that the acceptable prediction error for urban scenarios is an RMSE between 6 and 7 dB, using only these six features, we can achieve satisfactory performance since the RMSE obtained is 6.6 dB.
3) When features related to some information about the terrain elevation and building distribution are combined with the Tx (*Set-10*), we can observe an improvement in the prediction performance, with 1.3 dB compared with the (*Set-6*). This result shows that adding more details about the region, particularly those that provide additional information regarding the transmitting antenna, leads to a slight decrease in the RMSE.
4) When we added features that describe in detail the path profile between the Tx and Rx and the first Fresnel zone (*Set-15* and *Set-20*), the error difference when compared with the model with all features is less than 1 dB. This result reinforces the importance of using information about the radio environment, especially details obtained by a 3-D obstruction in the path, in the prediction propagation models.
5) The other information, such as additional statistics about the terrain and building complexity and the differences in the coordinates of Tx and Rx (*Set-25*), also yields a slight improvement in the model. Also, the difference in the RMSE compared with the set of all features is only 0.14 dB, which highlights that using the remaining 39 features does not impact the prediction of the RSRP, and there is no need to consider them in the ML approach. Besides that, from the results obtained in the feature importance analysis, shown in Section V, of the 39 features that can be disregarded, 38 of them have an %IncMSE below 10%.

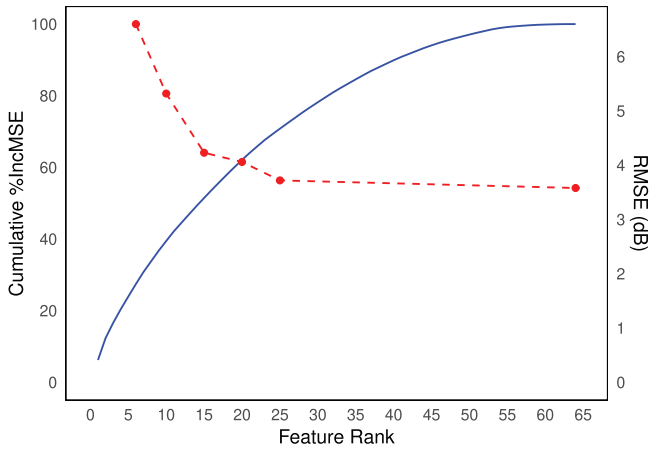


Fig. 5. Blue solid line represents the cumulative %IncMSE of the features, indicating their relative importance in the model. The red points connected by a dashed line represent the RMSE values obtained using different numbers of features.

From our analysis of prediction errors using reduced sets of features, we found it is feasible to develop a simpler prediction model that requires fewer features and less data. For instance, our results indicate that detailed discrimination

of obstructions along the path, as defined by PTC_i features, does not significantly reduce error and may be unnecessary. This is evidenced by the model using only 25 of the original 64 features, achieving a prediction accuracy with just a 0.14 dB difference. Notably, ten of these 25 features were introduced for the first time in this study.

In conclusion, for reliable and accurate RSRP prediction in urban or suburban areas, the following 25 features should be included in the ML approach.

1) The downtilt and azimuth angles, transmission power, Tx height, Tx effective height, Tx elevation, Tx and Rx angular difference, length of the LOS, distance from Tx to Rx, Tx and Rx coordinates differences, the radius and the obstruction percentage of the first Fresnel Zone, distance from Tx to the first obstruction occasioned by the terrain.
2) The differences between Tx and Rx effective heights, Tx effective height and maximum elevation, Tx effective height and maximum building height, and Rx effective height and maximum building height.
3) The statistics to represent the terrain complexity, measured by the mean and standard deviation of the elevation and the maximum and minimum elevation. The statistics to represent the building complexity are represented by the mean and standard deviation of the building heights and the maximum building height.

VII. Conclusion

In this study, we analyzed the impact of various features representing the radio propagation environment on predicting the RSRP, focusing on work from the last decade that suggested features for predicting RSSI, RSRP, or path loss in urban and suburban outdoor scenarios. Using real-world data from 186 635 measurements across 326 base stations, we trained an RF model with 64 input features, achieving an RMSE of 3.58 dB. These features included parameters like frequency, transmission power, azimuth, downtilt, terrain elevation, building complexities, obstructions, and clutter information.

Our feature importance analysis led to the development of five new feature sets, aiming to reduce model complexity while maintaining similar prediction performance. Reducing the features to 15 or 20 resulted in less than a 1 dB increase in RMSE. The ML approach using the 25 most important features (*Set-25*) produced an error difference of only 0.14 dB compared with the 64-feature model, demonstrating that significant model simplification can be achieved without compromising accuracy. These 25 features effectively represent the diverse elements of the radio propagation environment, including antenna information, terrain elevation, and building complexity.

This study focused on the 2.5-GHz frequency range, but the proposed ML approach can be extended to other ranges, including millimeter waves, with possible variations in feature importance for different ranges like the SHF band. The approach can also be adapted for predicting path loss, RSSI, and other signal power measures. Future research could extend this work to specific scenarios and network configurations not covered in our study. We suggest verifying whether the error and feature importance observed here are consistent in other scenarios and, if needed, adding specific features, such as those relevant to underground environments or high-mobility networks like trains.

References

[1] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60–69, Sep. 2019.

[2] Y. Fu, K. N. Doan, and T. Q. S. Quek, "On recommendation-aware content caching for 6G: An artificial intelligence and optimization empowered paradigm," *Digit. Commun. Netw.*, vol. 6, no. 3, pp. 304–311, Aug. 2020.

[3] M. Banafaa et al., "6G mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities," *Alexandria Eng. J.*, vol. 64, pp. 245–274, Feb. 2023.

[4] D. Serghiou, M. Khalily, T. W. C. Brown, and R. Tafazolli, "Terahertz channel propagation phenomena, measurement techniques and modeling for 6G wireless communication applications: A survey, open challenges and future research directions," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 4, pp. 1957–1996, 4th Quart., 2022.

[5] Z. Yun and M. F. Iskander, "Ray tracing for radio propagation modeling: Principles and applications," *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[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, Aug. 1980.

[7] R. Edwards and J. Durkin, "Computer prediction of field strength in the planning of radio systems," *Proc. Inst. Electr. Eng.*, vol. 116, no. 9, pp. 1493–1500, Sep. 1969.

[8] R. Edwards and J. Durkin, "Computer prediction of service areas for v.h.f. mobile radio networks," *Proc. Inst. Electr. Eng.*, vol. 116, no. 9, pp. 1493–1500, 1969.

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

[10] 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.

[11] 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.

[12] J. C. Delos Angeles and E. P. Dadios, "Neural network-based path loss prediction for digital TV macrocells," in *Proc. Int. Conf. Humanoid, Nanotechnol., Inf. Technol., Commun. Control, Environ. Manage. (HNICEM)*, Dec. 2015, pp. 1–9.

[13] M. Ayadi, A. Ben Zineb, and S. Tabbane, "A UHF path loss model using learning machine for heterogeneous networks," *IEEE Trans. Antennas Propag.*, vol. 65, no. 7, pp. 3675–3683, Jul. 2017.

[14] T. A. Benmus, R. Abboud, and M. Kh. Shatter, "Neural network approach to model the propagation path loss for great Tripoli area at 900, 1800, and 2100 MHz bands," in *Proc. 16th Int. Conf. Sci. Techn. Autom. Control Comput. Eng. (STA)*, Dec. 2015, pp. 793–798.

[15] S. Bolli, "Propagation path loss model based on environmental variables," in *Proc. 12th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE)*, Oct. 2020, pp. 368–373.

[16] N. Faruk et al., "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.

[17] N. Faruk et al., "ANN-based model for multiband path loss prediction in built-up environments," *Sci. Afr.*, vol. 17, Sep. 2022, Art. no. e01350.

[18] 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.

[19] R. He, Y. Gong, W. Bai, Y. Li, and X. Wang, "Random forests based path loss prediction in mobile communication systems," in *Proc. IEEE 6th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2020, pp. 1246–1250.

[20] 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.

[21] 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, Jul. 2023.

[22] 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.

[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] S. Ojo, A. Sari, and T. 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.

[25] S. I. Popoola, S. Misra, and A. A. Atayero, "Outdoor path loss predictions based on extreme learning machine," *Wireless Pers. Commun.*, vol. 99, no. 1, pp. 441–460, Mar. 2018.

[26] S. I. Popoola, E. Adetiba, A. A. Atayero, N. Faruk, and C. T. Calafate, "Optimal model for path loss predictions using feed-forward neural networks," *Cogent Eng.*, vol. 5, no. 1, Jan. 2018, Art. no. 1444345.

[27] S. I. Popoola et al., "Determination of neural network parameters for path loss prediction in very high frequency wireless channel," *IEEE Access*, vol. 7, pp. 150462–150483, 2019.

[28] 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.

[29] R. D. A. Timoteo, D. C. Cunha, and G. D. C. Cavalcanti, "A proposal for path loss prediction in urban environments using support vector regression," in *Proc. ICT*, Jul. 2014, pp. 119–124.

[30] 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.

[31] C. Huang et al., "Artificial intelligence enabled radio propagation for communications—Part I: Channel characterization and antenna-channel optimization," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 3939–3954, Jun. 2022.

[32] S. P. Sotiroudis, S. K. Goudos, and K. Siakavara, "Feature importances: A tool to explain radio propagation and reduce model complexity," *Telecom*, vol. 1, no. 2, pp. 114–125, Aug. 2020.

[33] C. Nguyen and A. A. Cheema, "A deep neural network-based multi-frequency path loss prediction model from 0.8 GHz to 70 GHz," *Sensors*, vol. 21, no. 15, p. 5100, Jul. 2021.

[34] C. Huang et al., "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.

[35] A. D. S. Braga et al., "Radio propagation models based on machine learning using geometric parameters for a mixed city-river path," *IEEE Access*, vol. 8, pp. 146395–146407, 2020.

[36] S. Cheerla, D. V. Ratnam, and H. S. Borra, "Neural network-based path loss model for cellular mobile networks at 800 and 1800 MHz bands," *AEU-Int. J. Electron. Commun.*, vol. 94, pp. 179–186, Sep. 2018.

[37] G. P. Ferreira, L. J. Matos, and J. M. M. Silva, "Improvement of outdoor signal strength prediction in UHF band by artificial neural network," *IEEE Trans. Antennas Propag.*, vol. 64, no. 12, pp. 5404–5410, Dec. 2016.

[38] 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. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 576–584, Dec. 2017.

[39] C. Lee and S. Park, "An approach of the diffraction loss prediction using artificial neural network in hilly mountainous terrain," *Microw. Opt. Technol. Lett.*, vol. 59, no. 11, pp. 2917–2922, Nov. 2017.

[40] T. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," *IEEE Antennas Propag. Mag.*, vol. 45, no. 3, pp. 51–82, Jun. 2003.

[41] M. G. Jadidi, M. Patel, and J. V. Miro, "Gaussian processes online observation classification for RSSI-based low-cost indoor positioning systems," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 6269–6275.

[42] S. Bakirtzis, J. Chen, K. Qiu, J. Zhang, and I. Wassell, "EM DeepRay: An expedient, generalizable, and realistic data-driven indoor propagation model," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 4140–4154, Jun. 2022.

[43] W. Chen, Y. Lin, and J. Yang, "Hybrid prediction model for field strength with ray tracing and artificial neural networks," in *Proc. IEEE 14th Int. Conf. Commun. Technol.*, Nov. 2012, pp. 301–305.

[44] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2nd Quart., 2018.

[45] 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.

[46] A. R. Ozdemir, M. Alkan, M. Kabak, M. H. Gulsen, and M. H. Sazli, "The prediction of propagation loss of FM radio station using artificial neural network," *J. Electromagn. Anal. Appl.*, vol. 6, no. 11, pp. 358–365, 2014.

[47] M. Uccellari et al., "On the application of support vector machines to the prediction of propagation losses at 169 MHz for smart metering applications," *IET Microw., Antennas Propag.*, vol. 12, no. 3, pp. 302–312, Feb. 2018. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-map.2017.0364

[48] B. J. Cavalcanti, G. A. Cavalcante, L. M. de Mendonça, G. M. Cantanhede, M. M. M. de Oliveira, and A. G. D'Assunção, "A hybrid path loss prediction model based on artificial neural networks using empirical models for LTE and LTE-A at 800 MHz and 2600 MHz," *J. Microw., Optoelectron. Electromagn. Appl.*, vol. 16, no. 3, pp. 708–722, Sep. 2017.

[49] H. A. O. Cruz, R. N. A. Nascimento, J. P. L. Araujo, E. G. Pelaes, and G. P. S. Cavalcante, "Methodologies for path loss prediction in LTE-1.8 GHz networks using neuro-fuzzy and ANN," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Aug. 2017, pp. 1–5.

[50] M. F. A. Fauzi, R. Nordin, N. F. Abdullah, and H. A. H. Alobaidy, "Mobile network coverage prediction based on supervised machine learning algorithms," *IEEE Access*, vol. 10, pp. 55782–55793, 2022.

[51] 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.

[52] C. E. G. Moreta, M. R. C. Acosta, and I. Koo, "Prediction of digital terrestrial television coverage using machine learning regression," *IEEE Trans. Broadcast.*, vol. 65, no. 4, pp. 702–712, Dec. 2019.

[53] S. Ojo, M. Akkaya, and J. C. Sopuru, "An ensemble machine learning approach for enhanced path loss predictions for 4G LTE wireless networks," *Int. J. Commun. Syst.*, vol. 35, no. 7, p. 5101, May 2022.

[54] L. Wu et al., "Artificial neural network based path loss prediction for wireless communication network," *IEEE Access*, vol. 8, pp. 199523–199538, 2020.

[55] F. Zhu, W. Cai, Z. Wang, and F. Li, "AI-empowered propagation prediction and optimization for reconfigurable wireless networks," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–10, Jan. 2022.

[56] 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.

[57] D. D. Coleman and D. A. Westcott, *CWNA Certified Wireless Network Administrator Official Study Guide*. Hoboken, NJ, USA: Wiley, 2012.

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

[59] C. Phillips, D. Sicker, and D. Grunwald, "A survey of wireless path loss prediction and coverage mapping methods," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 255–270, 1st Quart., 2013.

[60] S. Mohammadjafari, S. Roginsky, E. Kavurmacioglu, M. Cevik, J. Ethier, and A. B. Bener, "Machine learning-based radio coverage prediction in urban environments," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 4, pp. 2117–2130, Dec. 2020.

[61] *Method for Point-to-Area Predictions for Terrestrial Services in the Frequency Range 30 MHz to 3000 MHz*, document Recommendation ITU-R P.1546, 2013. [Online]. Available: https://www.itu.int/rec/R-REC-P.1546/en

[62] J. Deygout, "Correction factor for multiple knife-edge diffraction," *IEEE Trans. Antennas Propag.*, vol. 39, no. 8, pp. 1256–1258, Aug. 1991.

[63] *Propagation By Diffraction*, document ITU-R P.526-11, 2011. [Online]. Available: https://www.itu.int/rec/R-REC-P.526/en

[64] *Atoll 3.1.0 Model Calibration Guide*, document Release AT310-MCG-E1, Forsk, 2011.

[65] *Propagation by Diffraction*, document ITU-R P.526-12, 2012. [Online]. Available: https://www.itu.int/rec/R-REC-P.526/en

[66] Y. Zheng, Z. Liu, R. Huang, J. Wang, W. Xie, and S. Liu, "RSRPSet: The dataset of the 16th CPGMCM," *IEEE DataPort*. [Online]. Available: https://dx.doi.org/10.21227/4ba2-tg21

[67] J. S. Seybold, *Introduction to RF Propagation*, 1st ed., Hoboken, NJ, USA: Wiley, 2005.

[68] Z. Yi, L. Zhiwen, H. Rong, W. Ji, X. Wenwu, and L. Shouyin, "Feature extraction in reference signal received power prediction based on convolution neural networks," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1751–1755, Jun. 2021.

[69] Y. Zheng, J. Wang, X. Li, J. Li, and S. Liu, "Cell-level RSRP estimation with the image-to-image wireless propagation model based on measured data," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 6, pp. 1412–1423, Dec. 2023.

[70] F. Jiang, T. Li, X. Lv, H. Rui, and D. Jin, "Physics-informed neural networks for path loss estimation by solving electromagnetic integral equations," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15380–15393, Oct. 2024.

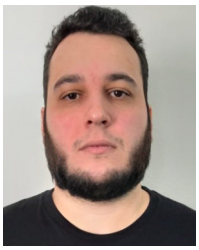
[71] R Package Version 3.6-14. (2023). *Raster: Geographic Data Analysis and Modeling*. [Online]. Available: https://CRAN.R-project.org/package=raster

[72] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324

[73] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: https://CRAN.R-project.org/doc/Rnews/

[74] J. A. Azevedo, F. E. Santos, T. A. Sousa, and J. M. Agrela, "Impact of the antenna directivity on path loss for different propagation environments," *IET Microw., Antennas Propag.*, vol. 9, no. 13, pp. 1392–1398, Oct. 2015.

[75] I. Rodriguez et al., "A geometrical-based vertical gain correction for signal strength prediction of downtilted base station antennas in urban areas," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–5.



Marlon Jeske received the B.S. degree in mathematics from the Regional University of Blumenau (FURB) in 2015, and the joint M.S. degree in operations research from the Technological Institute of Aeronautics (ITA), São José dos Campos, Brazil, and Federal University of Sao Paulo (UNIFESP) in 2019, where he is currently pursuing the Ph.D. degree in operations research.

His research primarily centers on employing mono and multiobjective optimization techniques, metaheuristics, and machine learning methodologies to address challenges in planning and deploying wireless networks.

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, mathematical programming, and how these disciplines interact to tackle problems in the big data era.

Dr. Aloise 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.

Brunilde Sansò (Senior Member, IEEE) is currently a Full Professor in telecommunication networks with the Department of Electrical Engineering, Polytechnique Montreal, Montreal, QC, Canada. She is also the Director of LORLAB, a research group dedicated to developing effective methods for the design and performance of wireless and wireline telecommunication networks.

Dr. Sansò has received several awards and honors, 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.

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

She is currently an Associate Professor with the Computer Science Division, Aeronautics Institute of Technology (ITA), São José dos Campos, Brazil. She is also an associate editor of leading journals and 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.