Received 18 March, 2024; revised .

Machine Learning for Radio Propagation Modeling: A Comprehensive Survey

Manjuladevi Vasudevan and Murat Yuksel
Electrical and Computer Engineering
University of Central Florida, Orlando, Florida
vmanjularamesh@gmail.com, murat.yuksel@ucf.edu

ABSTRACT With recent advancements in the telecommunication industry and the deployment of 5G networks, radio propagation modeling is considered a fundamental task in planning and optimization. Accurate and efficient models of radio propagation enable the estimation of Path Loss (PL) or Received Signal Strength (RSS), which is used in a variety of practical applications including the construction of radio coverage maps and localization. Traditional PL models use fundamental physics laws and regression-based models, which can be guided with measurements. In general, these methods have small computational complexity and have been highly successful in attaining accurate models for settings with trivial environmental complexity (e.g., clear weather or no clutter). However, attaining high accuracy in radio propagation modeling at complex settings (e.g., an urban setting with many buildings and obstacles) has required ray tracing, which computationally complex. Recently, the wireless community has been studying Machine Learning (ML)-based modeling algorithms to find a middle-ground. ML algorithms have become faster to execute and, more importantly, more radio data measurements have become available with the increased deployment of wireless devices. In this survey, we explore the recent advancements in the use of ML for modeling and predicting radio coverage and PL.

INDEX TERMS Path Loss; Machine Learning; Radio Propagation; Wireless Channel Modeling; Neural Networks

List of Abbreviations

3GPP	3rd Generation Partnership Project
AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
AOA	Angle of Arrival
AP	Access Point
BS	Base Station
CNN	Convolutional Neural Network
CWI	COST-Walfisch-Ikegami
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
DTM	Digital Terrain Model
ELM	Extreme Learning Machine
FNN	Feedforward Neural Network
FSPL	Free-Space Path Loss
GAN	Generative Adversarial Network
GPR	Gaussian Process Regression
ITU	International Telecommunication Union
KGNN	Knowledge-Guided Neural Network
KNN	K-Nearest Neighbors
LDPL	Log-Distance Path Loss
LOS	Line-of-Sight
M-PLM	Multiple Path Loss Model
MDT	Minimization of Drive Tests
ML	Machine Learning
MLP	Multi-Layer Perceptron
mmWave	millimeter-wave
NLOS	Non-Line-of-Sight
NN	Neural Network
PL	Path Loss
PLE	Path Loss Exponent
ReLU	Rectified Linear Unit
REM	Radio Environment Map
RF	Random Forest
RFID	Radio Frequency Identification
RMSE	Root-Mean-Square Error

RSRP	Reference Signal Received Power
RSS	Received Signal Strength
RSSI	Reference Signal Strength Indicator
RSSP	Reference Signal Strength Power
Rx	Receiver
SMOGN	Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise
SVM	Support Vector Machine
SVR	Support Vector Regression
TCA	Terrain Clearance Angle
TDOA	Time Difference of Arrival
TOA	Time of Arrival
TRGR	Two-Ray Ground Reflection
Tx	Transmitter
UE	User Equipment
UNET	U-shaped Encoder-Decoder Network
USARP	Ubiquitous Satellite-Aided Radio Propagation
UWB	Ultra-wideband
XGBoost	Extreme Gradient Boosting

I. Introduction

In radio communication, a wireless signal is transmitted from a Transmitter (Tx) to a Receiver (Rx) via unguided free space. When a signal is propagating in free space, during the transmission it can undergo attenuation caused by reflection, absorption, and refraction. This loss in signal strength during propagation is referred to as PL, and understanding this loss in radio signals has attracted a lot of attention over several decades of wireless research as it enables many practical applications such as localization, radio engineering, and cellular system optimization. In general, the PL is measured in decibels (dB) and defined as the difference in the transmitted signal strength to the RSS.

Apart from free space, loss in signal strength can also be caused by terrain type, clutter height, the type of environment (urban, rural and suburban), the nature of propagation medium, the number and height of buildings which the signal needs to penetrate, the antennas' radiation pattern, type of antenna used, and the location of antennas to name a few. PL calculation is useful for link budget calculation, coverage prediction, system performance optimization, and selecting the location for a Base Station (BS). Measuring PL in the location of interest is always an expensive and time-consuming process. To avoid such measurement costs, the calculation of PL is called 'PL prediction' and it enables prediction of how much power is to be received at the Rx end, i.e., RSS prediction, which is typically the main goal of radio propagation models. Several PL prediction techniques are developed, including closed-form models using fundamental physics laws without actual measurements, fitting closed-form models to measured data, and purely empirical models. Radio propagation models are used to calculate PL in different environmental settings. Traditional radio propagation models are formulated based on two ways: (i) Empirical models and (ii) Deterministic models.

Empirical models are based on the measured data and averaged losses, which enables the computation of the received signal level in each propagation medium. Empirical models use the relationship between PL and environmental parameters for modeling and are computationally efficient in many cases. Though results produced by the empirical models are not very accurate as they take only a few environmental parameters for the calculation. Numerous commercially available prediction tools are based on these models such as the Lee model [1], the Okumura-Hata model [2] for urban and suburban environments, the Walfish ilkegami model [3] for the dense urban environment, which is specifically useful in a microcellular system where antennas are deployed lower than building's height.

Deterministic models are also called geometric models which estimate signal power directly from the path profile. Deterministic models use detailed environmental information such as 3D maps, satellite images, terrain information, and antenna type. Ray tracing is one of the most common deterministic methods used for both indoor and urban scenarios [4]. The theory of these models is computed by numerically solving Maxwell's equation and it obeys the physical laws of wave propagation [5]. As deterministic models consider more parameters for calculation, results produced by them are considered more reliable than empirical models. However, they are more computationally expensive than empirical models as they require more time and power to compute. In general, deterministic models are used for short propagation paths as the accuracy of the PL prediction becomes more dependent on the environmental details (e.g., multipath or shadowing effects) for shorter paths.

A. ML-based vs. Traditional PL Models

Increased deployment of wireless devices generates more data about radio propagation behavior. It has become relatively easier to measure radio signals in different scenarios. The availability of such diverse data and the increase in ML computation efficiency enable ML-based radio models to be more accurate [6], which has increased the amount of such ML-based models in the community. In parallel, more wireless device deployment has increased the contention for radio spectrum resources and it has also increased interference. It has become expected that radio signals are generated more intelligently so that interference is minimized [7], [8]. This too has increased the efforts to use ML-based models in radio transmissions. For instance, DARPA's Spectrum Challenge [9] encourages wireless transceivers to figure out how to share highly contentious radio frequencies if no regulation is imposed [10]. Such initiatives have motivated researchers to build radio systems where propagation is inferred using advanced ML-based techniques [11]–[14].

PL prediction using ML models involve four high-level stages, as illustrated in Fig. 1. First and foremost is to collect a diverse set of data from various environments and identify relevant features from raw data to represent the radio

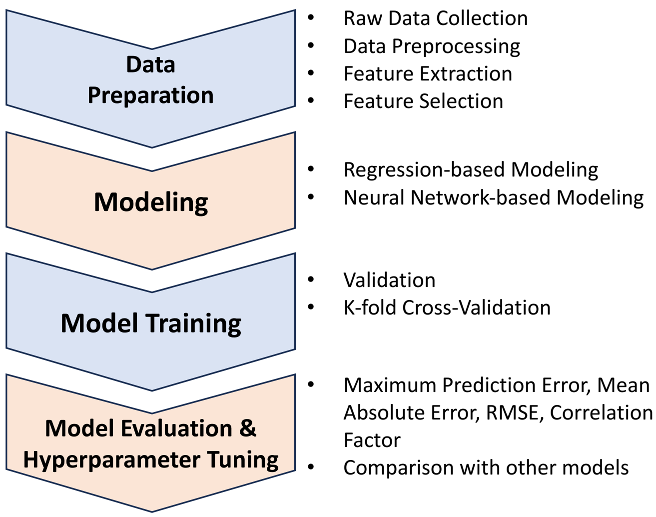

FIGURE 1. ML-based PL Prediction Procedure

TABLE 1. ML-based vs. Traditional PL Models

ML-based Models	Traditional Models
Learn complex relationships between various input features and PL from datasets	Rely on environment and propagation characteristics such as distance, frequency, antenna height, and terrain type
Automated feature selection	Manually selected features
Can adapt to diverse and complex environments or changes in the environment	Often based on specific assumptions and may require manual calibration to capture environmental changes and diversity
Require labelled data for training	Rely on mathematical equations
Computational complexity decreases with less learning capability	Computational complexity increases with more accuracy

environment effectively. Second is to select an ML method that fits well to the setting and frequencies being targeted. Commonly used ML methods for PL prediction are Random Forest (RF), Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Neural Network (NN), Decision Tree (DT), and Convolutional Neural Network (CNN). Then, ML-based models are trained and validated over data collected in different environments. Finally, the models are evaluated and their hyperparameters are tuned using common error calculation methods.

ML-based PL models offer a new opportunity to better capture radio propagation, but they also face several challenges. Table 1 summarizes key differences between traditional and ML-based PL models. Traditional PL models mainly rely on mathematical formulations derived from physical principles as well as limited empirical measurements. On the other hand, ML-based approaches leverage empirical data to model PL [15]. While traditional PL models offer simplicity and computational efficiency, ML-based models' major advantage is their ability to achieve high accuracy in complex environments. In terms of computational complexity, both traditional and ML-based approaches have their own strengths and weaknesses.

B. Surveys of ML in Wireless Modeling

Surveys on radio propagation and coverage date back to several decades ago. A comprehensive survey of radio propagation models is done by Phillips et al. [16] who examine various PL prediction methods, including theoretical, empirical, ray-optical, and measurement-based approaches. The survey captures fundamentals of propagation modeling under various settings. Focusing again on fundamentals, Hrovat et al. [17] reviews advancements in tunnel radio propagation modeling, focusing on numerical methods, waveguide approaches, ray tracing, and two-slope PL models, and evaluates them based on complexity and environmental information required. Recently, Diago-Mosquera et al. [18] have offered a review of indoor radio propagation characteristics in terms of PL mechanisms and fading and shadowing effects, and have introduced a new channel model taxonomy emphasizing the need for accurate modeling in small cells. The survey compares empirical, physical, and hybrid modeling methods – which we also utilize in some of our sub-taxonomy of the ML-based radio propagation modeling literature. Similarly, Al-Saman et al. [19] have recently reviewed mmWave channel measurement studies in indoor environments, focusing on measurement techniques, PL models, and delay spread for frequencies from 28 to 100 GHz. These surveys focus on traditional (e.g., physics- and/or regression-based) models and do not cover the recent ML-based propagation modeling literature.

As the use of ML has increased in the field, there have been several recent surveys on using ML in radio propagation and wireless channel modeling. Chiroma et al. [20] and Mladenovic et al. [21] survey Deep Learning (DL) methods for developing radio propagation models. These surveys focus on DL and do not cover traditional regression-based methods. In their invited articles, Huang et al. survey ML methods for antenna and channel optimization [22] and inference of application-specific scenarios from the outcome of the radio propagation model [23]. These surveys offer a high-level taxonomy of the existing literature on ML use for wireless modeling and do not delve into how ML methods are applied to certain problems. Seretis and Sarris [24] also survey DL methods for radio modeling and present a case study of using an Artificial Neural Network (ANN) to predict RSS through a straight circular-shaped tunnel. The survey also does not cover the legacy methods and only offer a high-level taxonomy of ML in radio modeling. Aldossari and Chen [25] survey legacy as well as reinforcement learning methods on radio wireless channel modeling, but do not cover recent data-driven DL methods that emerged in the last few years.

These prior surveys have made excellent coverage on various aspects of radio propagation and wireless channel modeling using ML. However, a comprehensive survey that covers traditional regression-based methods and the recent DL methods is lacking. With this survey, we aim to fill this gap by making a taxonomy of the most recent ML studies in radio modeling, discuss their pros and cons with respect to the legacy methods as well as among them, and delve into how some of the recent ML methods are applied to various radio propagation modeling problems. Another major difference of our survey is that it offers tutorialistic content both on radio propagation fundamentals and ML methods used for radio propagation modeling. This tutorialistic content is appropriate for researchers new to the area of propagation modeling.

In a similar direction, there have been a number of surveys on indoor localization methods. Liu et al. [26] reviews indoor positioning systems, focusing on triangulation, scene analysis, proximity, and location fingerprinting. Zafari et al. [27] conducts a detailed survey of indoor localization techniques such as Angle of Arrival (AOA), Time of Arrival (TOA), Time Difference of Arrival (TDOA), and examines various technologies including Wi-Fi, Radio Frequency Identification (RFID), Ultra-wideband (UWB), and Bluetooth. The survey investigates these systems based on energy efficiency, availability, cost, range, latency, scalability, and accuracy, and identifies challenges in achieving accurate indoor localization. In response to increased use of ML for indoor positioning, Roy and Chowdhury [28] reviews recent ML advancements for indoor localization, covering methods from traditional algorithms to DL. The survey discusses challenges with dynamic data, improvements from techniques like Extreme Learning Machine (ELM) and NN, and addresses benchmarking issues. Nessa et al. [29] reviews ML techniques for enhancing indoor positioning systems in terms of accuracy and reliability. Their survey covers both supervised and unsupervised ML methods and their integration with various technologies, and contrasts them with traditional algorithms. Though these surveys on indoor localization relate to our survey, we focus on RSS prediction in indoor settings and do not cover indoor positioning literature.

C. Survey Organization and Taxonomy

In this paper, we survey recent advances in ML-based PL prediction in wireless radio systems. We outline key metrics and parameters used for mapping the PL prediction problem to an ML framework. We systematically cover ML methods using regression and DL, and discuss their key pros and cons for the PL prediction problem. With a focus on RSS prediction, we elaborate on various applications of ML-based PL prediction in outdoor and indoor settings. We also briefly discuss future opportunities and challenges in this direction of research. This survey does not introduce new propagation models but offers a new reading of ML-based radio propagation models. We scope the survey to sub-6 GHz bands to make a comprehensive coverage of all key propagation models applicable in these frequencies in a single paper.

We categorize the ML-based radio propagation models in terms of their use of *regression* (Section IV.B) or *neural networks* (Section IV.C), as shown in (Figure 6). With a focus on the type of data being used, we taxonomize the ML-based radio propagation models based on what type of data they use: *ray tracing only* (Table 5), *measured data only* (Table 7), and *ray tracing and measured data together* (Table 6). We also categorize these models in terms of their applicability to frequency bands and urban or suburban settings. We structure the discussion of ML-based RSS prediction approaches in terms of *model-based prediction*, *data-driven prediction*, and *hybrid* models, as shown in Figure 16. Following this structure, we elaborate on the outdoor and indoor RSS prediction methods separately.

The rest of this paper is organized as follows: First, Section II offers a tutorial of fundamentals of radio propagation modeling by covering the commonly used models in the literature. Readers knowledgeable in radio propagation modeling may skip this section. Then, in Section III, we discuss the types of raw radio data used for radio propagation modeling and how these raw data is processed (including feature selection and construction) for use in ML algorithms. In Section IV, we describe how ML algorithms view the problem of radio propagation modeling and give an overview of the ML algorithms, by categorizing them into *regression-based* and *DL-based* algorithms, used in the radio propagation modeling literature. Section V offers a new taxonomy of ML-based methods by defining *model-based*, *data-driven*, and *hybrid* methods for RSS prediction. Following this taxonomy, Sections VI and VII cover the studies about RSS prediction in outdoor and indoor settings, respectively. A difference in Section VII is that it presents a detailed discussion on how RSS predictions in indoor environments are used for building maps of radio coverage. Section VIII offers a discussion of open research problems in radio propagation modeling. Finally, we summarize the paper in Section IX and discuss possible future work.

II. Radio Propagation Fundamentals

Radio propagation models help to predict how waves propagate through different environments and they are an essential tool for network planning and optimization in the telecommunications industry, helping engineers and planners estimate signal strength, coverage, and interference in various propagation environments. We, next, discuss some of the commonly used radio propagation models.

A. Free-Space Path Loss (FSPL)

FSPL model is used to determine the attenuation of a signal as it travels through free space without any obstacles in its Line-of-Sight (LOS). FSPL formula is derived from Friis equation [30], [31]:

$$\frac{P_r}{P_t} = D_t D_r \left(\frac{\lambda}{4\pi}\right)^2 d^{-\rho} \quad (1)$$

where P_t is the signal power at the Tx, P_r is the received signal power at the Rx, D_t and D_r are the degree to which the radiation emitted is concentrated in a single direction of the Tx and Rx respectively, λ is the signal wavelength, d is the distance in meters between the Tx and Rx antennas, and ρ is the Path Loss Exponent (PLE). PL is commonly expressed in dB as:

$$\begin{aligned} L &= 10\log(P_t/P_r) && (2)\\ &= 20\log(4\pi) + 10\rho\log d - 20\log\lambda - 10\log(D_t D_r). && (3)\end{aligned}$$

PL increases exponentially with the distance between the Tx and the Rx and the value of ρ expresses how fast this increase is. ρ can vary between 2 and 4, depending on how lossy the environment is. ρ is 2 for free space; and hence, the FSPL can be expressed in dB as:

$$L_{FSPL} = 20\log(4\pi d) - 20\log\lambda - 10\log(D_t D_r). \quad (4)$$

By assuming speed of light for the radio signal propagation, the above wavelength-based loss model can conveniently be expressed in terms of frequency as:

$$L_{FSPL} = 20\log(df) - 10\log(D_t D_r) - 147.56, \quad (5)$$

where f is the frequency in Hz. The directivity values provide a good handle in terms of incorporating the heterogeneity of the Tx and Rx antennas. For isotropic antennas, directivity is 1. Though specifics depend on how the antenna is designed, a commonly used expression for directivity is $\frac{2}{1-\cos(\theta/2)}$ for a conical antenna beam where θ is the half-power beamwidth.

Higher directivity means that more of the signal power is transmitted or received in comparison to an omni-directional antenna. Usually, this implies that the antenna can Tx or Rx over longer distances. Hence, the directivity of the antennas is expressed as part of their gain in dB. Let $G_t = 10\log D_t$ and $G_r = 10\log D_r$ be the gains of the Tx and Rx antennas, respectively. We can, then, revise (5) as:

$$L_{FSPL} = 20\log(df) - G_t - G_r - 147.56. \quad (6)$$

FSPL is applicable only in far field, i.e., when $d >> \lambda$ and is a simple model for estimating PL in obstacle-free environments. However, a major limitation of this model is that it does not account for obstacles. Therefore, it cannot be applied to many real-world environmental conditions, limiting its practical applicability.

B. Two-Ray Ground Reflection (TRGR) Model

The TRGR model considers both the direct path and ground reflection path between Tx and Rx, as illustrated in Fig. 2. This model is particularly useful with environments with clear LOS and the distance to the ground is small. Further, it applies to the scenarios with the reflective ground surface. The model calculates the PL by using the following formula [32]

$$L_{TwoRay} = \frac{P_r}{P_t} = \frac{G h_B^2 h_R^2}{d^4}, \quad (7)$$

where G is the antenna gain [32], h_B is the height of the Tx, and h_R is the height of the Rx. This PL model can be expressed in dB as:

$$L_{TwoRay} = 40\log d - 10\log G - 20\log(h_B h_R). \quad (8)$$

The main capability of this model is to capture the impact of the signal reflecting from the ground while keeping the model simplistic, similar to the FSPL model. The model is applicable to all frequencies, however, it is more accurate for lower frequencies where ground reflection is more prominent.

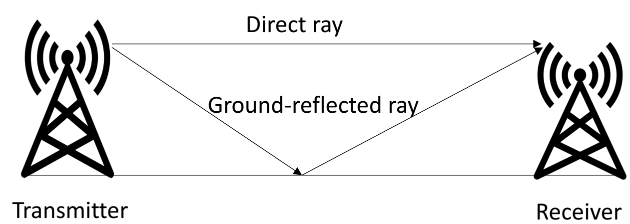

FIGURE 2. TRGR model

C. Okumura-Hata Model

Okumura-Hata Model is one of the widely used empirical radio propagation models. It models the PL by using the logarithmic form $f + h\log_{10} d$, where f and h are functions expressing the signal frequency and the antenna heights, and d is the distance between the Tx and Rx antennas. This is a quite useful form as it enables simple mathematical manipulations of the channel model. It is commonly used for frequency bands between 150 MHz and 1.5 GHz. It considers frequency, BS antenna height, and the distance between the Tx and Rx for calculating PL. The model calculates the median PL in dB by using the following formulas, for urban and suburban environments [33], [34]:

$$\begin{aligned} L_{OH}^{Urban} &= 69.55 + 26.16\log f - 13.82\log h_B \\ &\quad - C_m + [44.9 - 6.55\log h_B]\log d, \quad (9)\end{aligned}$$

$$L_{OH}^{Suburban} = L_{OH}^{Urban} - 2\left(\log\frac{f}{28}\right)^2 - 5.4 \quad (10)$$

where f is the frequency in MHz, and C_m is a constant offset in dB.

The Okumura-Hata model is more suitable for the calculation of PL in urban and suburban areas. The model lacks accuracy in rural environments and mountain areas and also has frequency limitations such as the assumption of static antenna height and omnidirectional antennas.

D. COST 231

COST 231 [3] is the extension of the Hata model developed by European Cooperation in the Field of Scientific and Technical Research (COST) specifically for European markets. The COST 231 model is particularly effective in the frequency range of 500 MHz to 2 GHz, making it especially well-suited for applications such as GSM1800 in urban environments. COST 231 model calculates PL in dB by using the formula [35], [36]:

$$L_{Cost231} = 46.3 + 33.9\log f - 13.82\log h_B - a(h_R) + (44.9 - 6.55\log h_B)\log d + C \quad (11)$$

C is the constant, C=0 for medium cities and suburban areas, C=3 for metropolitan areas. $a(h_R)$ is the mobile station antenna height correction factor as described in the Hata Model for urban areas.

E. Walfisch-Ikegami Model

Walfisch-Ikegami model is useful in calculating PL in urban and suburban environments. It takes into account street orientation, building heights, street width, frequency of operation, and antenna heights for greater accuracy. For LOS settings, the model [37] is expressed as

$$L_{WI}^{LOS} = 42.6 + 26\log d + 20\log f. \quad (12)$$

For Non-Line-of-Sight (NLOS) settings, the model combines several loss components:

$$L_{WI}^{NLOS} = \begin{cases} L_{FS} + L_{\text{rts}} + L_{\text{ms}}, & L_0 + L_{\text{rts}} + L_{\text{ms}} > 0 \\ 0, & L_{FS} + L_{\text{rts}} + L_{\text{ms}} \leq 0 \end{cases} \quad (13)$$

where L_{FS} represents free-space loss, L_{rts} is rooftop-to-street diffraction and scatter loss, and L_{ms} is the loss due the multi-screen diffraction [38]. These transmission losses are given by

$$L_{FS} = 32.4 + 20\log d + 20\log f, \quad (14)$$

$$L_{\text{rts}} = -16.9 - 10\log w + 10\log f + 20\log \Delta h_{\text{roof}} + L_{\text{ori}}, \quad \text{and} \quad (15)$$

$$L_{\text{ms}} = L_{\text{bsh}} + k_a + k_d \log d + k_f \log f + 9\log b, \quad (16)$$

where w is the width of the roads in meters, b is the building separation distance in meters, d is the distance in km, and $\Delta h_{\text{roof}} = h_{\text{roof}} - h_{\text{m}}$. As shown in Fig. 3, h_{m} and h_{roof} are, respectively, the height (in meters) of the mobile and the roof height of the building that reflects the signal. L_{bsh} is the shadowing gain and it is calculated as

$$L_{bsh} = \begin{cases} -18\log(1 + \Delta h_{\text{base}}), & h_{\text{base}} > h_{\text{roof}} \\ 0, & h_{\text{base}} \leq h_{\text{roof}} \end{cases} \quad (17)$$

where $\Delta h_{\text{base}} = h_{\text{base}} - h_{\text{roof}}$. The factor k_a signifies the additional PL encountered by BS antennas when situated below the rooftops of nearby buildings and it is represented as

$$k_a = \begin{cases} 54, & h_{\text{base}} > h_{\text{roof}} \\ 54 + 0.8\Delta h_{\text{base}}, & d \geq 0.5 \text{ and } h_{\text{roof}} \leq h_{\text{base}} \\ 54 + \frac{0.8 d \Delta h_{\text{base}}}{0.5}, & d < 0.5 \text{ and } h_{\text{roof}} < h_{\text{base}}. \end{cases} \quad (18)$$

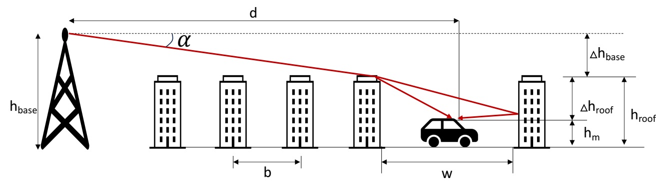

FIGURE 3. Walfisch-Ikegami model parameters

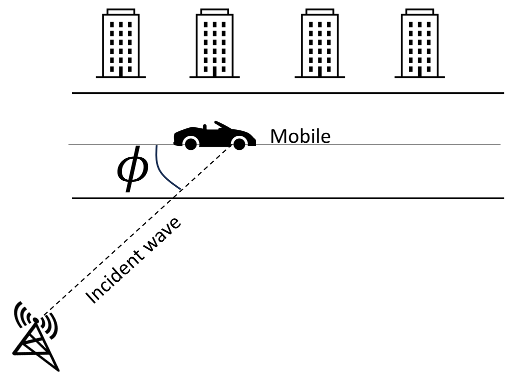

FIGURE 4. Walfisch-Ikegami model's street orientation parameter

The parameters k_d and k_f respectively govern the relationship between multi-screen diffraction loss and distance, and radio frequency. They are calculated as follows:

$$k_d = \begin{cases} 18 - 15\left(\frac{\Delta h_{\text{base}}}{h_{\text{roof}}}\right), & h_{\text{base}} \leq h_{\text{roof}} \\ 18, & h_{\text{base}} > h_{\text{roof}} \end{cases} \quad (19)$$

$$k_f = -4 + \begin{cases} 0.7\left(\frac{f}{925} - 1\right), & \text{medium-size city or suburb} \\ 1.5\left(\frac{f}{925} - 1\right), & \text{metropolitan center.} \end{cases} \quad (20)$$

L_{ori} is the orientation factor that captures the loss due to street orientation with respect to the BS, as shown in Fig. 4. It is represented as

$$L_{\text{ori}} = \begin{cases} -10 + 0.354\phi, & 0^\circ \leq \phi < 35^\circ \\ 2.5 + 0.075(\phi - 35^\circ), & 35^\circ \leq \phi < 55^\circ \\ 4.0 - 0.114(\phi - 55^\circ), & 55^\circ \leq \phi < 90^\circ. \end{cases} \quad (21)$$

The Walfisch-Ikegami model is considered to be highly accurate for urban environments as it captures even the street orientation. Since the model focuses on capturing an urban environment's geometry, it boils down to the FSPL model when such components do not exist in the environment, and hence, performs poorly for rural environments. It is computationally complex as it considers environment's geometry to calculate the L_{rts} and L_{ms} components. Due to its high computation requirements, it has limited use in real-time applications.

F. Lee Model

Lee model is one of the most accurate models and it is designed for microcellular environments. Lee PL model is used to model a flat terrain and operates around 900

MHz. It accounts for both outdoor and indoor propagation characteristics, and can be used for various terrains like urban, suburban, and open areas. The Lee model calculates the PL by using the following formula [39]–[43]:

$$P_r = P_{r_0} - \gamma \log\left(\frac{r}{r_0}\right) + G_{\text{effh}}(h_e) - L - A_f + \alpha \quad (22)$$

where P_r is the received power, P_{r_0} received power at the intercept point r_0 in dBm, γ is the PL slope, r is the distance between Tx and the Rx in kilometers, r_0 is the distance between the Tx and the intercept point in kilometers, α is a factor of antenna heights, h_e is the effective antenna height, $G_{\text{eff}}(h_e)$ is gain from effective antenna height, L is actual antenna height at the Tx, and A_f is the maximum effective antenna height.

G. Egli Model

The Egli model was developed by John Egli in 1957. It is a terrain-dependent model, mainly developed for forest environments. It takes into consideration trees and uneven terrain. This prediction model is applicable in the frequency range of 40-900 MHz and with a distance range of less than 60 km. The Egli model has limited applicability outside forest areas. The median PL in dB according to the Egli model [44]–[46] is

$$L_{Egli} = 117 + 40\log d + 20\log f - 20\log(h_B - h_R). \quad (23)$$

H. ITU-R P.1546 Model

ITU-R P.1546 Model was developed by the International Telecommunication Union (ITU). It is suitable for predicting PL in rural, urban and suburban environments. ITU-R P.1546 model is useful for outdoor short-range propagation with the frequency range from 30 MHz to 4,000 MHz [47]. This model takes into account crucial factors such as the effective height of the transmitting antenna, corrections based on the receiving antenna height, and adjustments related to the Terrain Clearance Angle (TCA) [48], [49]. TCA correction may be added to increase the prediction accuracy, enabling consideration of obstacles close to the Rx site. This correction is determined by the TCA in degrees, expressed as the difference between the measured TCA and the reference TCA. The model takes into account the angular differences of the positions of the Tx and the Rx and considers field strengths of different paths the signal can take. These considerations contribute to a comprehensive and accurate prediction of the radio wave propagation in diverse outdoor scenarios.

I. Log-Distance Path Loss (LDPL)

LDPL is an extension of the FSPL model. It considers the shadowing effect that can be caused by objects in dense environments such as buildings. So, it is useful to estimate PL for a wide range of environments mainly for urban or densely populated locales, traversing significant distances within buildings. LDPL calculates PL in decibel by using the following formula [32], [50]:

$$L_{LDPL} = PL_0 + 10\rho\log_{10}\frac{d}{d_0} + X \quad (24)$$

where d_0 is the reference distance, PL_0 is the PL at the reference distance d_0, ρ is the PLE, X is a zero-mean Gaussian distributed variable. LDPL has limited applicability with rural and hilly terrain regions but accuracy can be improved with certain modifications [51].

III. Radio Data Preparation for ML

Collecting data, representing real scenarios, and preparing it by removing irrelevant parts are important steps in ML-based modelling since a large amount of data is needed to train the ML models. Data preparation starts with raw data collection, data preprocessing and feature extraction. Then, these data are used for training and performance evaluation (testing) for the ML methods.

A. Raw Radio Data

Radio measurement plays a key role in RSS prediction and radio map construction. RSS field measurements can be gathered in a few different ways. The most used methods are data collection through measurement and simulation. Data collected through physical measurement typically includes various signal measurements such as location, RSS, frequency, and phase. We give a brief description of the common radio data gathering methods below:

Fixed-point Method: The fixed-point method requires setting reference points in the entire survey location. RSS and coordinates are recorded for each reference point. Skilled workers and special types of equipment are required to collect data. This process is expensive since it requires a lot of manpower and time to create the database.

Walking Method: Instead of using reference points, landmarks and predesigned pathways are used for position reference in the walking method. Surveyors first label the landmark for the position reference and walk at a constant speed in the predesignated path. [52], for example, used a robotic vacuum cleaner to move in predesigned pathways to collect data and construct a radio map.

Drive Test: Drive test involves physically driving on a designated route while measuring various network parameters to evaluate coverage, signal strength, quality, and other network performance metrics. Drive tests provide valuable insights into the real-world performance of wireless networks and help network operators optimize their systems for better coverage, quality, and user experience. One of the major downsides of the conventional drive test is its high operational expenditure. To overcome this challenge the 3rd Generation Partnership Project (3GPP) community have come up with the solution named Minimization of Drive Tests (MDT) in their Release 10 (Rel-10) [53]. MDT enables operators to monitor network performance in real time, detect issues promptly, and optimize measurement coverage more comprehensively. It enables operation, administration and maintenance to collect radio measurements and location information from User Equipment (UE) even when it is idle. This minimizes the need for manual drive tests, reducing costs and improving efficiency. Later in 3GPP Release 11, MDT is enhanced to provide the complete view of network performance [54].

Crowdsourcing: In the crowdsourcing method, the data set is built with the help of a large group of people. There are several publicly available crowdsourcing databases. OpenCellID [55] and OpenBmap [56] are a few of them. Volunteers install apps on their mobile phones and the app collects the signal quality information when the user travels. The crowdsourced data collection method holds its advantages as well as disadvantages. Crowd-sourced data is very cost-effective and less time-consuming. It collects raw signal measurements and this information is considered more reliable than coverage maps provided by the service provider. Also, it provides information about the serving BS locations. A key challenge is the difficulty in differentiating indoor vs. outdoor measurements. Indoor data is more sensitive less reliable than outdoor data as the coverage map changes even for changes made in wall decorations and furniture etc. So, periodic updates of the radio frequency coverage map are required to maintain the quality. Further, this data collection method depends on volunteers and devising a sustainable framework to regularly update the radio data is challenging due to the need to incentivize volunteers.

Simulation: Simulated data collection is another way of collecting data, where network topology is generated with ray tracing software. A few of the commonly used simulations are Wireless Insite SignalPro® by EDX Wireless, Inc. [57], [58], WinProp software (Altair HyperworksTM) [59], and Remcom Wireless Insite [60]. Different types of ray tracing methods and algorithms are used to generate data based on the need. In general, raw data consist of information such as location, height, azimuth, tilt, Tx power, frequency, and antenna type of the BS.

Obtaining clean measurements are crucial in radio propagation modelling. To achieve clean measurements, researchers have used data preprocessing strategies (to be detailed next). De-noising, averaging or smoothing algorithms are employed to remove random fluctuations [61], and data labelling [62], [63] and cross-validation [64], [65] are performed by identifying and discarding unreliable measurements. Standard measurements such as Reference Signal Strength Indicator (RSSI) or Reference Signal Strength Power (RSSP) are commonly utilized, especially in the crowdsourcing method. RSSI and RSSP can vary in their interpretation and may not always account for interference. Interference, whether from nearby devices or environmental factors, can significantly impact the accuracy of measurements. Techniques like signal filtering can effectively mitigate the interference effects [66]. Indeed, ML algorithms can understand patterns in the collected data, enabling them to adapt and compensate for interference dynamically [67].

B. Data Preprocessing

Data preprocessing is considered one of the important steps in data processing since raw data has a lot of redundancies and errors. Cleaning the raw data at this stage can be quite complicated and may require the use of ML, e.g., the authors of [68] used manual marking of the data and the NN toolbox of MATLAB for data preprocessing. In general, data preprocessing consists of the following major tasks: Data cleaning, dimensionality reduction, scaling, data griddling and data partitioning.

Data cleaning [63] is used to improve the quality of data by removing missing values, duplicates, and irrelevant data. *Dimensionality reduction* is used to omit the less important dimensions of data which in turn reduces the complexity of data. *Data augmentation* is used to increase the size of the training dataset by generating new samples through various modifications applied to the existing data. For example, in [64], Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) is used to address the issue of imbalanced datasets in regression problems. It creates synthetic samples for the minority class by interpolating between existing minority samples. Additionally, it introduces Gaussian noise to both minority and majority samples to enhance the diversity of the dataset. *Scaling* data is used to normalize the data to similar or same ranges. *Data griddling* technique is a two-fold process, that maps all UE measurements into unique spatial bins to handle positioning errors in the measurements, and then averages the measurements inside each spatial bin to offset random noise from RSS. For example in [63], they use the data griddling method to average out the values based on data characteristics. *Data partitioning* is splitting the data into groups and processing them based on their characteristics. A common approach is to group data based on the Euclidean distances, e.g., between the sample location and the BS location [69].

C. Feature Extraction

Feature extraction helps in preparing input data by identifying the most relevant information from the raw data and transforming it into a format suitable for the ML algorithm. Input parameters obtained from the raw data can be classified as system-dependent parameters and environment-dependent parameters.

System-dependent parameters refer to the physical properties of the transmission system, including the carrier frequency, transmit power, antenna height and orientation, and Rx sensitivity. These parameters affect the propagation mechanism and can have a significant impact on PL prediction.

Environment-dependent parameters are related to the physical characteristics of the propagation environment, such

TABLE 2. Common Parameters for ML-Based Radio Propagation Modeling

Parameters	**System or Environment**	**Unit**
Tx Power	System	Watts
Height of Tx	System	Meters
Height of Rx	System	Meters
Distance	System	Meters
Angle of Tx	System	Degrees
Angle of Rx	System	Degrees
Antenna tilt	System	Degrees
Antenna directivity loss	System	Decibels
Antenna gain	System	Decibels
Bandwidth	System	Hz
Horizontal angle	System	Degrees
Vertical angle	System	Degrees
Antenna loss	System	Decibels
RSS data	System	Decibels
Vertical diffraction	Environment	Meters
Horizontal diffraction	Environment	Meters
First Diffraction Point	Environment	Meters
Last Diffraction Point	Environment	Meters
Coordinates	Environment	N/A
Line-of-Sight (LoS) angle	Environment	Degrees
Terrain database	Environment	N/A
Terrain clearance angle	Environment	Degrees
Terrain map	Environment	Meters
Coordinates	Environment	N/A
Building map	Environment	N/A
Street widths	Environment	Meters
Building heights	Environment	Meters
Building separation distance	Environment	Meters
Number of Building Penetrations	Environment	N/A
Indoor Distance	Environment	Meters
Outdoor Distance	Environment	Meters
Receiver Clutter Type	Environment	N/A
Path visibility	Environment	N/A
Vegetation type and density	Environment	N/A
Temperature	Environment	Celsius
Humidity	Environment	Percentage
Precipitation	Environment	Millimeters

as the terrain, building conditions, and vegetation. These parameters can important particularly in urban and suburban environments where multipath reflections and diffractions are prevalent. Environmental information can be obtained from 3D digital maps or topographic databases. The weather conditions, including temperature, humidity, and precipitation, are also considered environment-dependent parameters. The propagation environment holds its importance in making design choices of ML-based propagation modeling regarding what input features need to be used. Classification of parameters and their units are shown in Table 2. Recent studies in ML-based propagation modeling for PL prediction explored different types of environmental settings such as urban [57]–[60], [63]–[65], [70]–[86], suburban [58], [60], [70], [71], [77], [83], [87], rural [88], [89], and indoor [68], [74], [90]. Table 3 indicates which environmental parameters are important based on the propagation environment.

TABLE 3. Importance of Environmental Parameters for ML-based Propagation Modeling

Environmental Parameters	**Urban**	**Suburban**	**Rural**	**Indoor**
Building Conditions	Yes	Yes	Yes	No
Vegetation Conditions	Yes	Yes	Yes	No
Temperature	No	No	No	Yes
Humidity	No	No	No	Yes
Precipitation	No	No	No	No

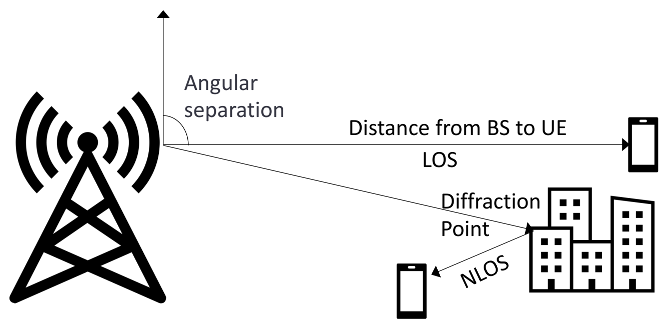

FIGURE 5. Features in radio propagation

D. Feature Selection and Construction

Many input features can be extracted from raw radio data, and selecting the best set of features can notably reduce the complexity of ML-based radio propagation modeling [91]–[94]. Generally speaking, there is a well-established tradeoff in this selection, i.e., considering a greater number of features leads to an increase in the model accuracy but also an increase in the complexity of the model. The predictive power of each feature is not equal. Hence, the goal should be selecting the features which contribute most to learning accuracy. Features are selected based on the design needs and the type of ML algorithm used for modeling. The few most important features selected in the ML-based radio propagation modeling are *propagation distance*, *angular separation*, *LOS state*, *carrier frequency*, *Tx power*, *indoor distance*, *outdoor distance*, *clutter type*, *building penetration diffraction point*, and *antenna optimization*. Some of these important features are illustrated in Fig. 5.

- *Propagation distance* refers to the distance between the Tx and the Rx, a crucial factor that influences the signal's strength and quality. As the propagation distance increases, the signal strength decreases due to factors like interference, scattering, and attenuation that impact the signal's ability to maintain its strength and quality over long distances.

- *Angular separation* refers to the angle between the angular bore height to the direction of the LOS path to the Rx. Both horizontal and vertical *angular separation* are considered in the modeling. The horizontal angular separation represents the angle between the LOS path and the horizontal plane, while the vertical angular separation represents the angle between the LOS path and the vertical plane. Angular separation plays an important role in determining PL and signal strength, especially in the presence of obstacles.
- *LOS state* refers to a clear and unobstructed path between the Tx and Rx antennas. Knowing whether or not a signal is LOS or NLOS plays a vital role in determining the channel state and the best use of the channel [95], [96], particularly in the emerging super-6 GHz bands [97], [98].
- *Carrier frequency* refers to the frequency of the electromagnetic wave used to transmit information.
- *Tx power* is the power level at which the Tx is radiating or emitting electromagnetic waves.
- *Indoor/outdoor distance* refers to the direct path between the Tx and Rx antennas. The indoor distance is the path that passes through buildings or other structures, while the outdoor distance is the path that is in the open air.
- *Clutter type* refers to the type of obstacles that can affect the signal path. Common clutter types include open areas, dense buildings, sparse buildings, trees, and water bodies. The clutter type plays a significant role in designing [99] and optimizing [100] wireless communication systems.
- *Building penetration* refers to the number of buildings penetrated by the signal in its direct path between the Tx and Rx.
- *Diffraction point* refers to the horizontal distance from the Tx to the first point of diffraction in the propagation path between a Tx and Rx.
- *Antenna optimization* plays a crucial role in determining RSS in wireless communication systems. Choosing the right antenna type, height, gain, polarization, placement, are essential factors in optimizing RSS for reliable communication.

Feature construction is the immediate next step after feature selection. It aims to construct a new feature based on the features obtained from the input data and can be performed through different strategies. Finding a new way of constructing new features can change the performance of the ML algorithm significantly. The literature includes a variety of feature construction methods for radio propagation modeling. In [101], mathematical transformations are first applied to the coordinate values of Tx and Rx. This leads to the extraction of several features such as the horizontal distance between Tx and Rx, elevation angle, distance across the buildings, and clutter types. In [57], input features were used to construct the mean value of building height, the standard deviation of building height, the normalized mean value of building distance, the normalized standard deviation of building distance, and building density. In [87], the height ratio of the Tx to Rx antenna feature is constructed. In [76], Tx antenna directivity loss and Tx antenna gain were calculated from the input features. Beyond these innovative heuristics, strategies to build self-learning algorithms are also used for feature construction. In [72], a fully connected Feedforward Neural Network (FNN) is used to learn and construct input features. In [102], different objects in the direct path, and the type of object are constructed from the input features. In [63], clutter type, number of building penetrations and indoor distance in each clutter type, and outdoor distance in each clutter type are constructed.

IV. Modeling Algorithms

ML-based techniques hold value in PL propagation modeling due to their higher accuracy and lower complexity rate. The performance of the propagation model mainly depends on the type of ML algorithm used in the modeling along with the amount of input data used for training as well as feature selection. Since propagation modeling is considered a regression problem, mostly supervised ML techniques are used to address the problem. In supervised learning, the algorithm aims to establish a relationship (mapping) between input data to the output. Considering input data as a and b as a corresponding label, the algorithm learns from this labelled dataset to understand the relationship between a and b. We can represent this relationship in functional form:

$$B = f(A, \theta) \tag{25}$$

where A is the input data vector such as transmitted beam's power, Rx sensitivity, antenna height, and terrain features; B is RSS (i.e., the output label); $f()$ is the mapping function that the ML algorithm learns during training; and θ is the set of tunable parameters for the modeling function $f()$. Supervised learning approximates mapping function $f()$ from the given dataset containing input-output pairs (a, b). Once the model is trained, it can then predict RSS values from any new input a'. The training process in supervised learning aims to minimize the difference between predicted output b' to the actual output b using a cost function $C()$ and adjusting the parameters, θ, of the model. The goal of the supervised learning can be expressed as

$$\arg\min_{\theta} C(b', b) \tag{26}$$

where $C(b', b)$ is the cost function that measures the difference between the predicted label b' and the actual label b. In its simplest form $C()$ can be just the difference of the two output labels, i.e., $C(b', b) = |b' - b|$. However, more complex and customized cost functions can be devised, e.g., a weighted sum of differences of various output metrics/labels such as RSS and AOA.

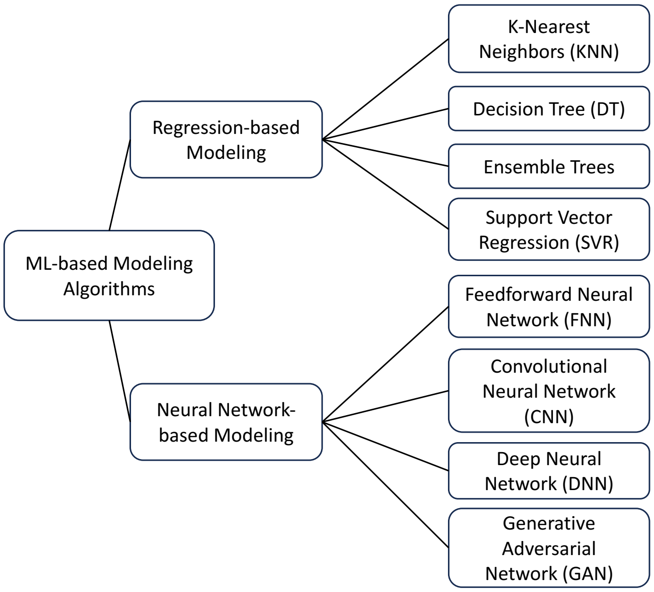

FIGURE 6. **ML models used for radio propagation modeling**

In general, a more complex ML model, with large training data gives better accuracy and a simple ML-based model performs tasks more quickly. A better model is considered to give better performance in less complex models with limited data sets. In the rest of this section, we discuss ML-based propagation modeling methods in terms of two dimensions: (1) Their requirements and types of input data and (2) the use of NNs in their methods (Figure 6).

A. Input Data Types and Requirements

Based on the type of data used for training we can classify ML-based propagation modeling methods into the ones using ray tracing, measured data, or a combination of ray tracing and measured data. Ray tracing-based models depend on the data generated from ray tracing software, which computes individual rays' propagation paths, accounting for interactions such as reflection, diffraction, and scattering. The data generated by the software consists of coordinates of ray paths, signal strengths, time taken to travel, and reception angle. Ray tracing-based models offer detailed insights into propagation mechanisms but may be computationally intensive and require accurate environmental data. Measured data-based models use the data measured in real world environments. They rely on a statistical analysis of measured data to derive PL. Training involves collecting significant measurement data in different environmental settings and analyzing it to develop models. The models that involve both ray tracing software and measured data obtain data from both the ray tracing software as well as measurements from real world scenarios. These integrated models aim to leverage the strengths of both approaches to improve prediction accuracy across various scenarios. Tables 5-7 show the comprehensive overview of papers we surveyed

TABLE 4. **Input Requirements of ML-based Modeling Methods**

	ML Method	Selection of Features	Order of Features	Functional Form	Environmental Data
Regression-based	**Linear Regression**	R	NR	R	NR
	Decision Tree	R	R	R	NR
	Ensemble Trees	R	R	R	NR
	KNN	R	NR	R	NR
	SVR	R	NR	R	NR
NN-based	**CNN**	NR	NR	NR	R
	DNN	NR	NR	NR	NR
	FNN	NR	NR	NR	NR
	GAN	NR	NR	NR	R

R: Required, NR: Not Required

in terms of the type of data used for training, also marked with key features used for modelling, ML algorithms and training method. Understanding these components is crucial for designing accurate and reliable predicting models.

In terms of the input data required by ML-based modeling methods, there may be four types of requirements: 1) *Selection of features*, indicating which independent variables are necessary for the ML method, 2) *Order of features*, which is a rank ordered list of the features, 3) *Functional form*, which is a functional form that uses the features as parameters to determine the outcome of the PL model, and 4) *Environmental data*, that gives information about the environment of the radio propagation. Table 4 summarizes the input data requirements of the ML-based PL modeling methods. Regression-based methods require the selection of features and at least one functional form while the NN-based methods do not require these input data. Further, tree-based regression methods require the order of features as well. While NN-based methods do not suffer from requiring feature selection or a particular functional form between the independent and dependent variables, some of them (e.g., CNN and GAN) may require environmental data to be available as input.

B. Regression-based Modeling

Regression-based models are the simplest form of ML models for radio propagation characterization. The approach feeds input data into a model with a pre-configured structure between the dependent d and independent variables of the radio propagation d'. A dependent variable is known as the response variable and the independent variable is known as the predictor variable. Considering unknown function to be mapped as g, we can represent the relationship as

$$g : d \rightarrow d'.$$

The training of the model involves minimizing the error in the capability of the independent variables to predict

TABLE 5. **ML-based Propagation Models: Using Ray Tracing Data***

Papers	Frequency	Environment	Features		Modelling algorithm	Training method
			System Dependent	*Environment Dependent*		
[72]	Unspecified	Urban	Distance, Angle of Tx	Building map	DNN	90% Training and 10% Validation
[73]	2.4 GHz	Urban	Distance, Height of Tx, Height of Rx	LOS angle, Terrain elevation	Random Forest and KNN	84% Training and 16% Validation
[65]	2.1 GHz	Urban	Height of Tx	Coordinates	KNN, SVR, Random Forest and AdaBoost	5-fold cross-validation
[60]	811 MHz and 2,630 MHz	Urban, Suburban	Distance, Height of Tx, Height of Rx	Terrain elevation, Terrain databases	DNN	75% Training and 25% Validation
[59]	2.3 GHz	Urban	Distance, Frequency, Height of Tx, Height of Rx	LOS angle, Coordinates	SVR, Random Forest, KNN	k-fold cross-validation
[78]	900 MHz and 1.8 GHz	Urban	Distance	LOS angle, Building heights	Random Forest, ANN	80% training and 20% Validation
[80]	900 MHz	Urban	Distance, Height of Tx	LOS angle	ANN (MLP)	80% training and 20% Validation
[81]	28 GHz	Urban	Frequency, Height of Tx	Vertical diffraction, Horizontal diffraction	CNN	70% training, 15% validation, and 15% test
[64]	3.5 GHz	Urban	Distance, Frequency, Height of Tx	Coordinates	Tree-based prediction models (Random Forest, ensemble)	k-fold cross-validation
[58]	900 MHz	Urban, Suburban		Terrain database, Terrain map	GAN	90% Training and 10% Validation
[79]	Unspecified	Urban	Height of Tx, Height of Rx, Distance		ANN (MLP)	Combination of DE approach and the Levenberg-Marquardt for training
[82]	Unspecified	Urban		Terrain database	ANN (KGNN)	
[103]	900 MHz		Height of Tx, Height of Rx	Terrain database	ANN (MLP)	Training and Validation
[85]	Unspecified	Urban		Building heights, Building widths, Street widths	ANN (MLP)	Training and Validation
[104]	Unspecified	Urban	Distance, Angle of Tx, Angle of Rx	Coordinates, Terrain elevation	GLM, KNN, MLP and DNN	3-fold cross-validation

* Data generated from ray tracing software for model training and validation

TABLE 6. **ML-Based Propagation Models: Using Both Ray Tracing and Measured Data***

Papers	Frequency	Environment	Features		Modelling algorithm	Training method
			System Dependent	*Environment Dependent*		
[105]	800 MHz	Urban	Frequency, Angle of Tx, Angle of Rx, Antenna tilt, Antenna gain, Bandwidth	Coordinates	DNN, Unets	
[106]	Unspecified	Urban	Distance, Height of Tx, Height of Rx	Coordinates, Building map	CNN, UNET, Radio UNET, UNET with Strided Convolutions and Inception	Training and Validation
[107]	Unspecified	Urban	Height of Tx, Heigh of Rx, Tx power, Antenna gain, Frequency		Random Forest	
[77]	3.5 and 28 GHz, 3.7 and 26 GHz	Urban, Suburban	Distance, Frequency		SVR, Random Forest	10-fold cross validation

*Ray tracing software for modelling, and measured data for validation

the dependent variables. During this depreciation, the coefficients in the pre-configured structure are tuned, which is essentially the training of the model. In most cases of PL modeling, this pre-configured structure includes several $\log$ functions preceded with scalar coefficients, as in (4) for FSPL and (8) for TRGR. This sum of $\log$s structure of the PL model function g arises from the dB unit for the PL and conveniently serves the regression-based ML methods as they simply try to learn the coefficients. Though learning PL in dB works well for regression-based methods, NN-based modeling does not necessitate such pre-configured structure and can work with any input features and output metrics as we will later discuss in Section IV.C.

As the baseline, Linear Regression describes the association between the dependent and independent variables as a linear relationship, e.g., as a polynomial of an arbitrary degree. Linear regression cannot fully model most real-world phenomena as they involve non-linear and mutually dependent features [63]. All dynamics of radio propagation, too, cannot be properly modeled with linear relationships as some features of the radio propagation environment are heavily correlated.

Regression-based ML algorithms are used in radio propagation modeling due to their ability to handle different types of data [70], [99], [112], [113] and the automatic selection of relevant features from a large set of input variables [104], [114]–[118]. Moreover, they are effective in solving complex non-linear relationships between environmental variables and signal propagation. They are the most suitable in scenarios where the radio environment is interpretable and/or simple. In particular, when there is limited data for training DL models, the regression-based ML methods are the best alternative.

TABLE 7. ML-based Propagation Models: Using Measured Data

Papers	Frequency	Environment	Features		Modelling algorithm	Training method
			System Dependent	*Environment Dependent*		
[74]	2,021.4 MHz	Urban	Frequency, Height of Tx, Height of Rx, Angle of Tx, Angle of Rx, Vertical angle	Coordinates	ANN (BPNN)	80% Training and 20% Test
[88]	881.52 MHz	Rural	Distance, Height of Tx, Height of Rx	Terrain clearance angle, Terrain database, Vegetation type and density	ANN	Training and Validation
[75]	Unspecified	Urban	Distance, Angle of Tx, Angle of Rx	LOS angle	3D-CNN	
[89]	3.7 GHz	Rural	Distance, Height of Tx, Height of Rx	Path visibility, Coordinates	SVR, Random Forest, ANN, B-kNN	Training and Validation
[76]	2.1 GHz	Urban	Distance, Height of Tx, Angle of Tx, Antenna tilt, Antenna directivity loss, Antenna gain, Tx power, Angle of Rx		CNN, FNN	90% Training and 10% Validation
[68]	1,800 MHz	Urban, Suburban, Rural	Distance, Height of Tx, Height of Rx, Antenna gain, Antenna loss, RSS data	Terrain map, Coordinates	ANN-BP and ELM	50% training and 50% Validation
[87]	450 MHz, and 1.45 and 2.3 GHz	Suburban	Distance, Frequency, Height of Tx, Height of Rx,		ANN, GPR	5-fold cross validation
[83]	1,890 MHz	Urban, Suburban	Distance, Height of Tx, Height of Rx,	Street width, Building height, Building separation distance	ANN (MLP) (generalized RBF-NN)	Training and Validation
[86]	853.71 MHz.	Urban	Distance, Horizontal angle, Vertical angle, Antenna loss	Terrain elevation, Coordinates	SVR	Cross-Validation
[63]	2.1 GHz	Urban	Distance, Horizontal Angle, Vertical Angle	LoS angle, First Diffraction Point, Last Diffraction Point, Number of Building Penetrations, Indoor Distance, Outdoor Distance, Receiver Clutter Type,	Decision Tree, KNN, Linear Regression, MLP	
[57]	Unspecified	Urban		Building map	GLM, KNN, MLP and DNN.	3-fold cross validation
[83]	1,890 MHz	Urban, Suburban	Distance	Street widths, Building heights	ANN (MLP) (generalized RBF-NN for path loss prediction)	Training and Validation
[84]	1,140 MHz	Urban	Antenna tilt, Antenna gain, Angle of Tx, Angle of Rx, Frequency, Tx power	Coordinates	ANN (MLP)	Training and Validation
[69]	Unspecified	Urban, Suburban	Distance, Angle of Tx, Angle of Rx, RSS data	Coordinates	ANN (MLP)	50% Training and 50% Validation
[108]	Unspecified	Urban, Suburban		Number of building penetrations	DNN	Mini-batch training
[109]	2.1 GHz	Urban, Suburban	Distance, Frequency, Height of Tx, Height of Rx, Antenna gain, Angle of Tx, Angle of Rx,	Coordinates	DNN (ConvNet)	80% Training and 20% Validation
[110]	Unspecified	Urban		Building map, Coordinates	DNN (Radio UNET)	80% Training and 20% Validation
[111]	Unspecified	Urban, Suburban	Distance, Frequency, Height of Tx, Height of Rx, Antenna tilt	LOS angle	SVM, RT, ET, LR, ANN, and GPR	10-fold cross validation

Next, we cover the most established regression-based radio modeling methods.

1) K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm used for solving classification and regression problems, where K represents the number of nearest neighbors considered for making predictions. First, we need to train the model with a measured dataset with corresponding variables/features such as the ones in Table 2. Once the model gets trained with a measured dataset, it can be used to predict the PL for new input variables or queries by finding the K nearest data points in the training set to the new input variables and taking the mean of their corresponding PL values. This is illustrated in Fig. 7. KNN methods are very transparent and explainable approaches. In general, the more variables/features they use the less explainable they are. But, since they require a certain set of variables to be defined before training, the outcome is explained by the significance of the match between the input query's values and the trained features. A study [59] predicts PL in an urban environment for cellular networks with the help of ML methods, including SVR, RF and KNN algorithms, and compares them with COST-Walfisch-Ikegami (CWI) empirical model [119]. Results show that ML algorithms used in the survey performed better than CWI providing much lower statistical errors. While comparing performance between the ML algorithms, KNN performed better in terms of preciseness of the PL predictions.

2) Decision Tree (DT)

DT is a versatile tree-based algorithm, with a flow chart-like structure that can be used for PL prediction modeling. It can be constructed by selecting relevant features that affect PL and recursively splitting the data into more homogeneous subsets until the predicted PL values are represented by the leaf nodes of the DT. An example DT for radio propagation

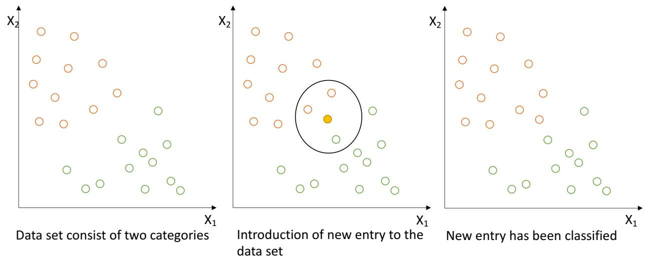

FIGURE 7. K-Nearest Neighbors (KNN)

modeling is shown in Fig. 8. [63] uses a DT algorithm for PL prediction, where the tree is derived from a set of independent variables and training data, and the leaf nodes represent the predicted PL values for the corresponding subset of data. The study shows that the DT algorithm predicts PL with good accuracy, as long as the features used in constructing the tree are relevant and the tree is not overfitting the training data. A nice characteristic of DT methods is their full explainability. To form a DT, one has to determine which feature is more important than others. For the model in Fig. 8, for example, distance is a more important feature than frequency. Yet, this clarity comes at a cost of being unable to learn models in settings for which we do not have a clear idea of which feature is more important.

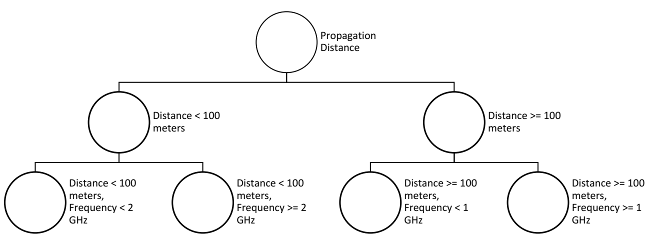

FIGURE 8. Decision Trees

3) Ensemble Trees

The principle behind the ensemble tree is a group of weak learners combined to gather to form a strong learner. In comparison to DT, ensemble trees are less explainable but offer more flexibility in exploring the possibilities of feature importance. They allow combining various feature importance possibilities, learned by weak learners, into a common model. Depending on how these weak learners are combined, there are a few different techniques in ensemble learning, the most common among them are *Bagging*, *RF*, and *Boosting*. The bagging technique is used to reduce the variance of the DT. The illustration for the bagging technique is shown in Fig. 9. [89] is one of the best examples using the bagging technique.

RF is one of the commonly used ML methods which employs a DT and applies a bootstrap aggregation for the selection of training samples as an ensemble member and training on these members. Unlike traditional DTs, RF introduces a random selection of features in the training process. The final result is obtained by averaging the predictions of all the ensemble members. A study [73], performed on air-to-air scenarios, based on RF and KNN builds prediction models and evaluates them by using the data generated by ray-tracing software and compares them with the Stanford University Interim [120] and CWI [119] models. The results show that ML-based models perform better than the empirical models, and further that, RF has better prediction performance. Moreover, RF has considerable advantages in handling a large number of input features and sorting the important features. Also, it is easy to implement and can do parallel computing.

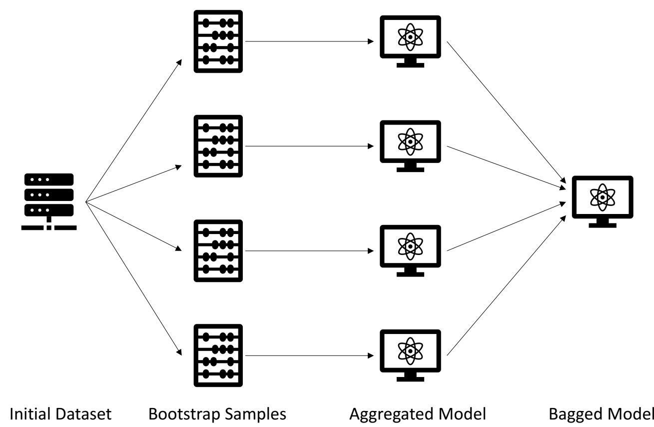

FIGURE 9. Ensemble Trees: Bagging technique

Boosting models mainly focus on reducing bias, often have low variance [121], and cannot be done in parallel without introducing approximations [122]. In baseline Boosting technique, trees are built one-by-one by fitting a simple model of data and summed sequentially. At each step, the model's net error is analyzed and the eventual aim of Boosting is to solve for net error from the prior

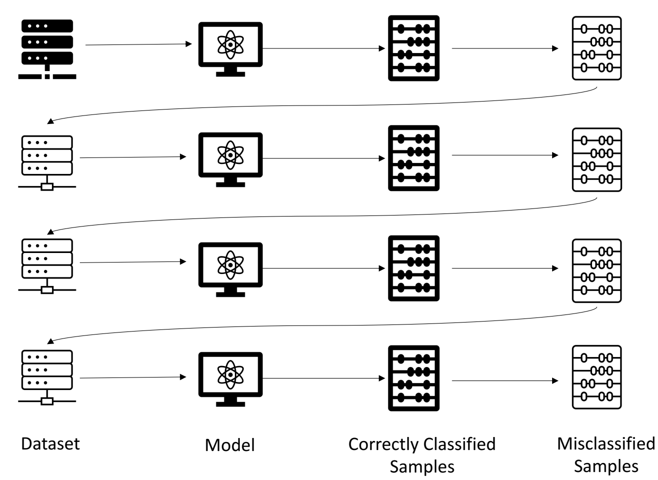

FIGURE 10. Ensemble Trees: Boosting technique

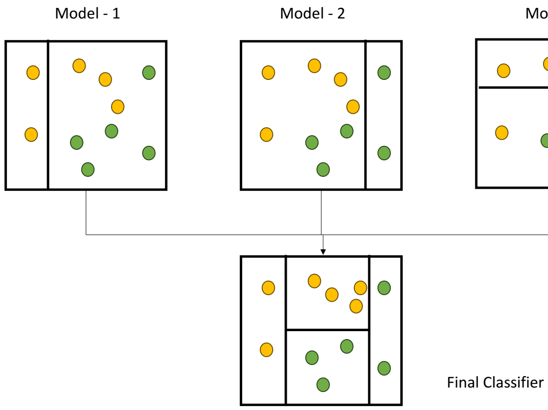

FIGURE 11. **Ensemble Trees: AdaBoost**

tree. The illustration for the boosting technique is shown in Fig. 10. Adaptive Boosting (AdaBoost) is one of the examples of ensemble trees that uses the Boosting technique. An example of AdaBoost is shown in Fig. 11. [65] studies modeling methodology on predicting PL from a flying BS in an urban environment using different ML methods such as KNN, SVR, RF and AdaBoost. Results show that AdaBoost has better accuracy in prediction. Extreme Gradient Boosting (XGBoost) [64], [123] is another Boosting technique used for radio propagation modeling. XGBoost starts with the simple DT and grows sequentially to correct mistakes from the previous ones. Each new tree focuses on the incorrect prediction from the previous tree, and the sequential growth of the tree continues until it reaches error-free results. XGBoost works by combining DT and gradient boosting. XGBoost's ability to handle large datasets, feature identification (to aid in optimizing signal coverage) and model interpretability make it well-suited for radio propagation modeling added that XGBoost is known for its execution speed and model performance.

4) SVR

SVR is a powerful statistical technique commonly used for regression analysis. It is an extension of the popular Support Vector Machine (SVM) algorithm and is particularly well-suited to solving complex nonlinear regression problems. SVR, in essence, is a more advanced version of KNN. As illustrated in Fig. 12, SVR categorizes training data to classes by finding the best hyper-planes that maximally separate the classes with the maximal margin. Hence, in addition to minimizing the classification error (typically defined as the sum of distances of data-points from the centers of their classes) just like KNN, SVR also attempts to maximize the difference between classes. Similar to KNN, SVR is also a transparent method as the variables have to be defined before training. However, they are computationally more complex.

In the context of radio propagation modeling, the SVR algorithm has been widely used for modeling PL in various studies [59], [65], [77], [89], [124]. In [74], [89], the authors choose to use SVR with the Gaussian kernel function since it is more suitable for small feature dimensions and lack of prior knowledge. Further, the Gaussian kernel function allows for the mapping of input data from a low-dimensional space to a high-dimensional space using non-linear functions. This, combined with the ability of SVR to search for an optimal hyperplane in the high-dimensional feature space that maximally separates the samples, makes it a particularly useful tool for solving such problems.

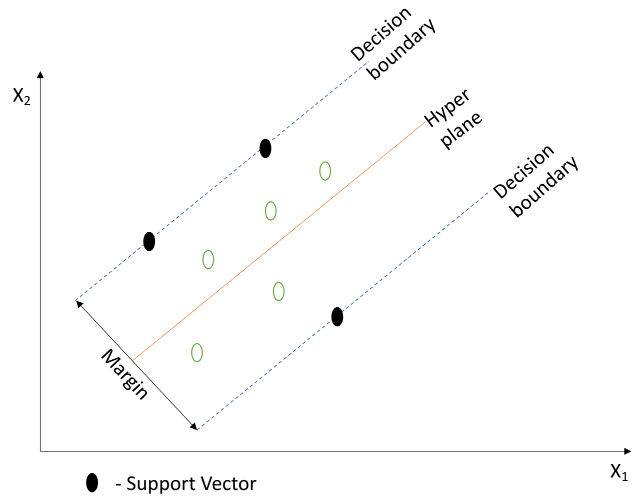

FIGURE 12. **Support Vector Regression**

C. Neural Network (NN)-based modeling

When the relationship between the independent and dependent variables is not expressible in (closed form) functional forms or is heavily dependent on the specific setting, the regression-based learning can only attain approximations of the actual relationship. Typically, NNs are used in such cases to better capture the model [125]. NNs are structured like the human brain, consisting of nodes, called artificial neurons. These nodes are connected via links, each of which has a weight. Similar to the tuning of the coefficients in regression-based models, an ANN gathers knowledge from samples through the training process and tune the weights to minimize the error in their predictive capability. Once training is complete, ANNs are then used to make predictions based on the weights learned obtained during training. This approach can be applied to any possible set of input and output variables of a model, which offers a great flexibility. Further, if a larger network of neurons is utilized in between the input and output variables, the ANN can learn better and attain higher modeling accuracy. Yet, this flexibility and accuracy in modeling comes at the cost of being less explainable. Adding more neurons to the ANN makes the learned patterns less transparent.

A simple ANN consists of the input layer, hidden layers, and output layer. Data is given in the form of input to each node. The input data to each node is multiplied with

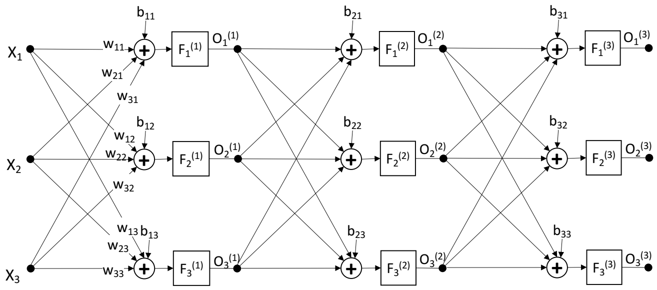

FIGURE 13. Three-layer FNN

a random weight and then passed through a bias, where a transfer function is applied. Finally, an activation function is applied before being given to the output. Various transfer and activation functions are used for modeling purposes. In general, more hidden layers in ANN offer deeper learning capability, hence the name Deep Learning (DL). In this survey, we discuss the DL approaches that are most commonly used for propagation modeling: FNNs, CNNs, Deep Neural Networks (DNNs), and Generative Adversarial Network (GAN).

1) Feedforward Neural Network (FNN)

FNNs are composed of multiple layers of neurons where each neuron in a layer receives inputs only from the previous layer and sends outputs only to the next layer in the network. The data in these networks flow only in one direction, from the input layer to the output layer, without any feedback loops. As illustrated in Fig. 13, FNNs utilize linear operations of weights and biases at each layer. In addition to regular weights on each link between neurons, FNNs use biases at each neuron to gain more flexibility. These biases are also tuned during learning. The use of FNN for PL prediction in rural macrocell environments is investigated in [88] with the ANN inputs such as antenna-separation distance, transmitting antenna height, TCA, land usage, vegetation type, and vegetation density. Terrain parameters are derived from a Digital Elevation Model and while vegetation type and density are obtained from measurement datasets. Transmitting antenna height, TCA, and land usage/vegetation information are identified as the most significant parameters. The study analyzes different ANN sizes and incorporates faster training algorithms to reduce training time while maintaining accuracy. The results show that ANN models outperform other common PL models, such as Recommendation ITU-R P.1546 and Okumura-Hata [126], of similar complexity.

FNNs offer unique strengths over other ML methods in the context of radio propagation modeling, including non-linearity, feature learning, generalization, scalability, and parallel processing. In radio propagation modeling, FNNs excel at learning hierarchical representations of features through their hidden layers. One of the key advantages of FNNs is their ability to perform feature learning. This enables FNNs to extract meaningful and discriminative representations from the input data, capturing the essential factors that influence radio wave propagation.

MLP belongs to the FNN class. All layers of an MLP are fully connected and there are no connections between neurons within the same layer. MLPs are used for predicting PL in urban environments. In [79], the approach involves providing detailed yet small information about the propagation environment to the ANN. Two different ANN design cases are presented and trained using a hybrid Differential Evolution-Levenberg Marquardt method, which is found to be more efficient than the classical Levenberg Marquardt algorithm in terms of weight optimization. The approach demonstrates satisfactory accuracy compared to the ray-tracing model. In a more specific outdoor setting [68], a single-hidden-layer MLP, Artificial Neural Network Backpropagation (ANN-BP) with a maximum of 50 hidden neurons, is employed using the normalized distance between the base and mobile stations as input. The models are trained with 50 different models trained 20 times each [127]. The ELM models have an architecture similar to ANN-BP, but with random input weights and biases generated based on a learning algorithm. In an ELM study [128], four different activation functions (sigmoid, sine, triangular basis, and radial basis) are investigated and the output weights are calculated using the Moore-Penrose pseudo-inverse matrix. The study evaluates the performance of empirical, ANN-BP, and ELM models for PL predictions in outdoor propagation environments. ELM models are found to be 140 times faster to train than ANN-BP models and produce the lowest Root-Mean-Square Error (RMSE). ELM models also show good generalization ability when tested with new input data.

Again for an urban setting, the study in [80] aims to synthesize ANNs that accurately predict PL with minimal input data using an MLP architecture with one hidden layer of 10 nodes and a linear activation function for output. They are trained using input-output data vectors and evaluated with mean absolute error, RMSE, and mean absolute percentage error. ANNs using both site-specific and LOS data trained for restricted coverage areas show higher accuracy.

When site-specific information is available MLPs show notably better performance since it becomes possible to utilize ray-tracing in addition to ML. [87] proposes a method for PL prediction in wireless sensor networks using a combination of dimensionality reduction, ANN-MLP, and Gaussian Process. This method results in a more generalized model with lower training time and better PL prediction accuracy than conventional models. Furthermore, the proposed method is highly reliable for site-specific wireless sensor network design. The proposed approach in [87] utilizes a simplified ray-tracing tool and two ANNs to predict radio propagation for indoor and outdoor scenarios. The design aims to achieve similar prediction accuracy to more complex methods but with reduced computational load. Accurate predictions depend on the selection of training sets based on either dominant path types or a route-oriented strategy. Results demonstrate errors below 7dB, although the model has limitations in predicting sharp variations between consecutive points.

The performance of MLP-ANNs and RF in predicting PL are compared in [78] for two Narrowband-Internet-of-Things frequency bands at 900 MHz and 1,800 MHz. The results indicate that both methods performed similarly well across three different input data types. But, the study highlights that the quality of the input data is critical in predicting PL through ML approaches. It is also found that LOS information is more important than street scene information in cases where the Tx is situated above building rooftops, but using both types of information led to even better results.

MLP has its unique strengths such as flexibility and being able to handle non-linear propagation effects and multi-dimensional data. MLPs can effectively model these relationships and capture the intricate dependencies between these dimensions. By considering the non-linear interactions of different parameters, MLPs can capture the complexities of radio wave propagation more effectively than linear models.

2) Convolutional Neural Network (CNN)

CNN is a type of ANN primarily used in image processing and object detection tasks. It is designed to take an input image and extract relevant features from it, which are then used to make predictions about the image. A typical CNN is comprised of two main parts: the feature extraction and the classification. The feature extraction part consists of multiple hidden layers responsible for identifying features from the input image. An activation function is applied after each convolutional layer, introducing non-linearity into the network and enabling it to learn more complex representations of the input data. Some of the common activation functions used in radio propagation modeling are described in Table 8. A fully connected layer in the classification part is responsible for identifying and classifying images. The output of the feature extraction part is flattened and passed into one or more fully connected layers, which output a probability distribution for the object classes. The class with the highest probability is then predicted for the input image.

A study was made to predict radio wave propagation [76] with a single image. The authors employ a CNN with four convolutional layers and two pooling layers in their CNN architecture to extract relevant features. The resulting feature values, combined with Tx and Rx parameters, are flattened and input into a fully connected NN for additional processing. The approach is illustrated in Fig. 14. The flattened feature values and input parameters are then used to train an FNN, which proved to be effective for regression-based prediction. The authors aim to leverage the strengths of both CNNs and FNNs to optimize feature extraction and achieve accurate predictions of radio propagation. By utilizing multiple convolutional and pooling layers, CNN extracts increasingly complex features from the input data.

TABLE 8. Activation functions

Activation Function	Description
Linear	No transformation, produce the same value as input value
Sigmoid	Maps the input value to a value between 0 and 1
Sinc	Maps the input value to between -1 and 1, resembling a sinusoidal wave with amplitude fading as the input goes to +/-∞
Cosine	Maps the input value to between -1 to 1
Triangular Basis	Produces isosceles triangular shaped waves with height 1 and base 2
Radial Basis	Bell curve shaped function that assigns higher values to inputs closer to the center and lower values to distant inputs
tanh	Maps the input value to between -1 and 1
ReLU	Sets negative input values to 0, Leaves positive value
Leaky ReLU	Similar to ReLU, but it allows a small negative slope for negative values
PReLU	Parametric ReLU has ability to choose best slope in the negative region
ELU (Exponential Linear Unit)	Similar to Leaky ReLU, but uses an exponential function for negative input values (smooth curve)
Softmax	Maps the input value to between 0 and 1

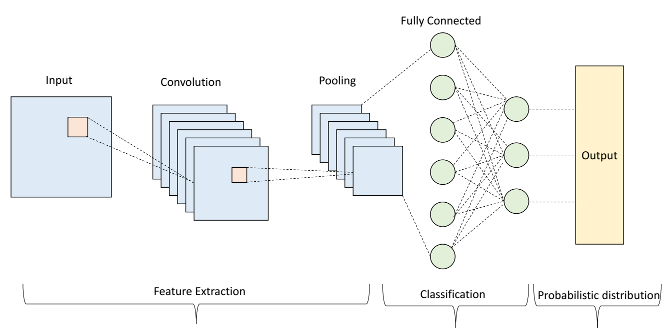

FIGURE 14. Convolutional Neural Network (CNN)

The model in [60] combines both regular NNs and CNNs to process satellite images and engineered features. The architecture has two NNs and one CNN. The CNN is used to analyze the satellite images while NN1 manages positional locators and engineered features. A middle layer is added to combine the outputs from the CNN and NN1. NN2 is then sequentially added with an activation function, such as Rectified Linear Unit (ReLU), to enable latent features to be a function of weighted positional locators and image features. Finally, NN2 is connected to the output to predict the received power of mobile communication systems. The model inputs include local coordinates, satellite images, and a PL model, and use a ReLU activation function to extract features. The results show that the DNN model improves PL prediction at unseen locations by $\approx$1 dB for 811 MHz and $\approx$4.7 dB for 2,630 MHz when compared to traditional modeling techniques such as ray-tracing and empirical models.

Research has been conducted to leverage CNNs to predict the distribution of PL directly from 2D satellite images without any additional data [71]. A CNN with multiple convolutional layers [129] is used for feature learning and VGG-16 [130] for predicting PL distribution. This approach enables real-time inference without the need for a 3D model of the area and utilizes transfer learning and fine-tuning with a pre-trained VGG-16 network. The network outputs are mapped to probabilities using softmax and optimized using the cross-entropy loss function. The results of the study indicate accurate PL distribution prediction for different communication frequencies and Tx heights, making this approach particularly advantageous when the dataset size is limited.

CNN is preferred over other ML techniques due to its unique advantages in learning spatial features and its capability to capture environment-dependent features. More importantly, when capturing spatiotemporal environmental features, CNNs can do well with less data [131]. Since radio propagation modeling typically involves data scarcity, CNNs offer better solutions as they can produce efficient results with a small set of data and can effectively extract important features such as the frequency with the help of convolutional layers.

3) Deep Neural Network (DNN)

DNN is a type of ANN that includes multiple hidden layers between the input and output layers. Compared to a traditional ANN, a DNN's additional deeper layers allow it to perform more complex tasks. However, due to the increased number of layers, DNNs are more susceptible to overfitting. Regularization methods, such as dropout or L2 regularization [132], [133], should be applied to reduce the risk of overfitting. DNNs excel in capturing complex non-linear relationships and performs well even with sparse training data, with lower prediction errors compared to other algorithms such as Linear Regression, KNN, and DT. Further, they exhibit [63] a 25% increase in prediction accuracy compared to empirical propagation models and a 12-fold decrease in prediction time compared to ray tracing-based commercial tools.

A data-driven, DNN-based PL model, named the Ubiquitous Satellite-Aided Radio Propagation (USARP) model [70], was developed to enhance the geographical generalization capabilities of empirical PL models. USARP uses satellite images to make PL predictions. DNN-based ResNet50s are used as a feature extractor for processing the input images. ResNet50 is trained using Bootstrap Your Own Latent [134] with extensive satellite images using self-supervised learning to extract information from ROI-filtered satellite images. Three Single-Layer Perceptrons, where each layer includes batch normalization, and a non-linear activation (e.g., PReLU) are used to make a prediction. The results show that USARP attains an RMSE of 12.34 dB which is 1 dB lower than linear regression, 3 dB and 2 dB lower than the SVR and the RF-based models. Extended studies are made to check the performance of USARP in multiple radio environments, and the results show that it achieves a higher prediction accuracy than linear regression models. Also, satellite-based inputs improve the RMSE of the PL predictions by more than 3 dB on the validation dataset, and around 1 dB in the generalization dataset when comparing with [135].

For an urban propagation environment, a DNN using 3D map of a city is studied for PL modeling [57]. The proposed model combines LOS and non-LOS propagation scenarios considering the path profile, Tx height, and distance between the Tx and Rx as input data. The model architecture includes an input layer of length Q+2 (where Q is cell radius), three hidden layers, and an output layer with one neuron predicting PL. The results show that DNN model outperforms traditional ML models such as the alpha–beta–gamma and close-in [136] models due to its ability to learn complex nonlinear relationships and perform well with sparse training data.

DNN is preferred over other ML algorithms due to its unique strengths such as flexibility and the advantage of capturing non-local dependencies. DNNs can handle various types of input data beyond spatial information. Radio propagation often involves analyzing time-varying signals and sequences of events. DNNs can effectively capture temporal dynamics and hence are well-suited for sequential and temporal modeling of radio propagation.

4) Generative Adversarial Network (GAN)

GANs are DL-based generative models. A GAN consists of two components: a generator and a discriminator. The goal of the generator is to generate synthetic data based on topology modeling and the goal of the discriminator is to learn from the synthetically generated data. GAN models learn much more quickly than CNN models. [58] uses a GAN technique to predict PL from satellite images. The generator consists of a U-shaped Encoder-Decoder Network (UNET) structure with skip connections to allow deeper architectures as shown in Fig. 15. Patch discriminator [137] is used to decide whether the output of the generator is true or false. The study finds that height map images provide more informative results than satellite images and that a GAN model can estimate PL values in real-time, providing an alternative to computationally complex ray-tracing simulations. The authors suggest increasing the dataset size for further improvements, indicating the need for further research in this area.

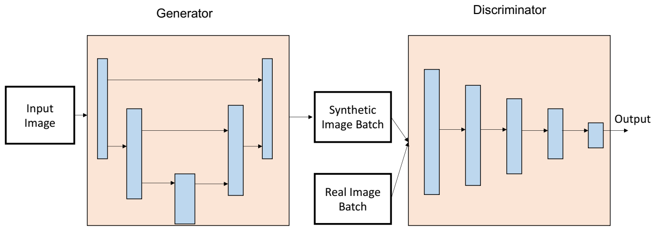

FIGURE 15. **GAN architecture**

GAN is known for its high-quality generations. The adversarial training framework of GANs encourages the generator to produce samples that are indistinguishable from real data. This makes GANs particularly effective in radio propagation modeling. GANs learn the underlying data distribution directly from the training data. Instead of explicitly modeling the probability distribution, GANs implicitly capture the patterns and features of the training data. This allows GANs to generate new samples that follow the same distribution as the training data, enabling them to generate diverse and novel samples. While GANs offer unique advantages, they also come with challenges such as training instability, mode collapse, and evaluation metrics. GANs require careful hyperparameter tuning, architectural design, and advanced training techniques to achieve optimal performance.

V. ML-based RSS Prediction Methods

Measuring RSS has been of high interest as it has direct ramifications for cellular network providers' business goals. Cellular providers regularly measure their RSS maps and advertise them to illustrate the quality of their service. Yet, it has also been a challenge to verify the providers' claims and make independent predictions of RSS from the service being provided to the users. Further, provisioning seamless wireless connectivity indoors requires a thorough understanding of the RSS across the building in question. Beyond connectivity, accurate indoor localization depends on the prediction accuracy of PL models. Based on the prediction methodology, we can classify RSS prediction approaches into three categories as shown in Figure 16, namely model-based, data-driven, or hybrid prediction methods.

Model-based prediction uses mathematical models and physical system information to make the prediction. Model-based approaches depend on physical models and data measurements, such as an Radio Environment Map (REM), to forecast environmental characteristics, offering cost-effective and swift prediction capabilities. These approaches also provide researchers with the chance to investigate diverse scenarios and extend predictions beyond available measured data. However, the accuracy of model-based predictions hinges on the quality of the underlying models and the input parameters employed.

In *data-driven prediction*, some of the measured data is given as input to train the model. The trained model is expected to have the capability of predicting the environmental features and physical propagation characteristics of the environment. Based on the model learned from measured data as well as signal processing approaches, PL can be predicted in non-measured locations. Data-driven prediction relies on real-world measured data to make predictions so it is considered more reliable. With sufficient training, data-driven models can easily adapt to the change in network conditions. Also, they can handle the complexity of interactions in intricate wireless networks while considering other factors that influence performance. While data-driven approaches offer these advantages, their challenges are dependence on data quality and sufficient training scenarios, and the risk of overfitting.

The *hybrid model* utilizes physics-based prediction models and measured data to formulate predictions. Depending on the specific requirements, various physics-based prediction models are employed as baseline. Hybrid models combine the strengths of both model-based prediction and data-driven prediction methodologies which in turn gives more accurate, adaptive, and cost-effective solutions than both models individually. Their key benefit is to correct errors in the measured data by using insights from the physics-based models.

In the next two sections, we survey the literature on RSS prediction for outdoors and indoors scenarios in the three categories.

VI. RSS Prediction in Outdoor Environments

A. Model-based Prediction

Based on how radio signals propagated in the environment and how they are affected by various factors, and how they can be predicted, the model-based approaches for outdoors can be classified into Propagation Models, REMs, PLE Models, and Terrain Profiling Models. We will next cover these types of model-based prediction approaches and the recent ML literature pertaining to them.

1) Propagation Models

Early propagation models give mathematical expressions for how the radio signals travel through the atmosphere and interact with the environment. Among the most successful propagation models for outdoor environments are the Okumura-Hata and LDPL. The Okumura-Hata model [2] has been a significant step in simplifying the prediction of PL in urban and suburban areas. It has been used as a benchmark for subsequent models and has greatly contributed to the planning of wireless communication networks. For long ranges, as applicable to lower radio frequencies, the LDPL [138] is a widely used empirical propagation model. LDPL is based on the logarithmic relationship between the distance travelled by the signal and its received power. It provides a quick and practical way to estimate signal strength over moderate to long distances in open areas.

To estimate signal strength, ML-based approaches [57], [63] complement empirical and ray tracing-based models,

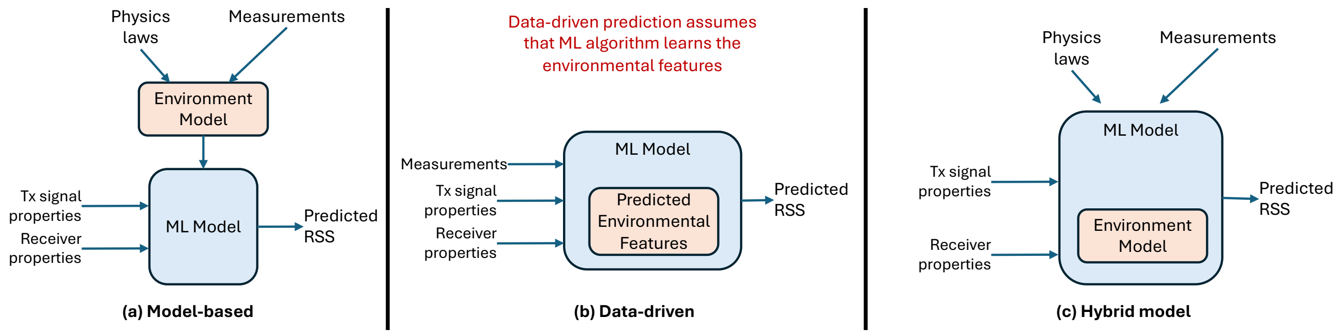

FIGURE 16. ML-based approaches to RSS prediction

incorporating both measured data and environmental information into a model of the environment. This model of the environment is fed to the ML algorithm along with various system features. This modern solution enhances the accuracy of signal strength predictions by utilizing intelligent predictors derived from these data sources. This ML-based approach is particularly valuable in environments where mathematical models become impractical to capture geometrical complexities such as in urban settings. To make more realistic predictions, [63], for example, utilizes measured data, BS information, and geographical data from three distinct datasets: Digital Terrain Model (DTM), Digital Height Model, and Digital Land Use Map. These measurements of the environment and fundamental physics laws (e.g., attenuation through a particular clutter type) are used to form predictors that can strongly model the environment. This requires judicially identifying critical components for the environment model to be effective. After this step, using both the environmental model and system properties, the prediction is accomplished using various ML algorithms, such as Linear Regression, KNN, DT, and DNN. The study assesses the performance of these algorithms, particularly focusing on the capability of DNNs to capture intricate channel characteristics.

In large urban environments, the importance of synthetic data becomes more pronounced for the radio propoagation model. [106] discusses the effectiveness of CNN and UNET models for radio signal prediction in such settings. The study uses the Wireless Insite ray-tracing software to generate the synthetic data for modeling, and compares the effectiveness of the model by using the RadioMapSeer dataset against Levie et al.'s [110] findings. [58] is another example of a model-based prediction method. This study uses satellite images in addition to the data generated from the InSite software [139] as input. It utilizes the PlaceMaker [140] extension for Google SketchUp4 [141] to obtain the 3D models along with the satellite images. From the 3D model, the height map image is extracted. Then 3D models are imported and merged with the InSite ray-tracing simulation environment. Using the ray-tracing results, excess PL values at the Rxs are calculated. PL values are treated on the grid of Rxs as an image which enables the utilization of image synthesis methods for PL prediction. Finally, a GAN-based technique is used to produce a PL image. Since the networks can produce real-time inference, this technique proves to be a viable alternative to ray tracing simulations, which have high computational complexity.

2) Radio Environment Maps REMs

An REM is a spatial database that stores information about the radio environment. REM consists of information such as signal strengths, interference levels, spectrum usage, antenna data, terrain data, and building and infrastructure data. REMs play a crucial role in predicting signal behavior based on real-world data. They utilize empirical values from measurements and offer data for the most realistic model-based prediction discourse.

ML techniques are heavily used for building and utilizing REMs. In [76], along with measured data obtained from Tokyo metropolitan area in the 2.1 GHz band, aerial photographs and building map images are used for radio propagation prediction. The study focuses on enhancement of predictive accuracy by incorporating building map images into ML. In particular, the proposed ML methodology employs a CNN-FNN model where the CNN extracts spatial features from the building map and the FNN incorporates them with system parameters to forecast received signal power. The outcomes of the investigation reveal that the proposed ML methodology consistently achieves superior predictive accuracy when compared with conventional techniques, even with a single image.

In a similar vein, enlarging the REM-based RSS prediction to a city level requires more capable approaches. [110] utilizes the Dominant Path Model [142] method and Intelligent Ray Tracing [143] to generate simulation, RadioMapSeer [144] dataset and city maps from OpenStreetMap [145] (Ankara, Berlin, Glasgow, Ljubljana, London, and Tel Aviv) as inputs to build an REM. The study proposes to use UNETs [146] for REM estimation. Two versions of RadioUNet are designed, one with no input measurements (RadioUNetC) and the other with input measurements (RadioUNetS). RadioUNetC uses only Tx location, PL measurements, and city map as its input (geometry of the environmental features are extracted). This method is categorized as a model-based simulation as a model is learned from training data. In that way, it does not have a physically interpretable formulation and its execution time (for the trained network) is much faster than existing model-based tools. RadioUNetS uses some measurements of the PL as one of its inputs along with the inputs given to RadioUNetC. So, this model can be categorized as a model-based fitting method.

Incorporating terrain and building occlusion effects into the REM-based prediction requires meticulous design of environmental parameters in the ML model. In an experimental study, [101] chooses four RSS-related features from readily available geographic data: 1) Horizontal distance between the antenna and receive terminal, 2) Elevation angle from receive terminal to the antenna, 3) Distances across different clutter types, and 4) Distance across buildings. Existing ML-based RSS models such as MLP, RF, XGBoost, and LightGBM (i.e., an improved version of XGBoost) [147] are testified with the reliable drive test data obtained from the outdoor scenario at 3.5 GHz for a 5G network in an urban environment. The results show high accuracy in RSS prediction with an efficient and extensible design.

3) PLE Models

PLE models are very commonly used mathematical representations to express the reduction in a radio signal's strength. The PLE model predicts that the radio signal's power decreases exponentially as it travels over a distance. Depending on various factors in the radio environment, the PLE can be different and time-varying. Hence, estimating PLE for a specific radio environment has been the focal point in this type of radio propagation modeling.

ML-based techniques have proved to be useful for improving the accuracy of PLE estimation. For long-range or omni-directional radio propagation models, DL-based methods are successful in estimating the PLE. For example, [148] utilizes CNNs to estimate the PLE by directly analyzing 2D satellite images. When the radio signal operates at short ranges or is heavily directional, more granular datasets or rigorous capturing of the 3D environment becomes important. In [1], the authors introduce an algorithm that leverages DL techniques to predict the PLE for outdoor millimeter-wave band channels. This algorithm incorporates a 3D radio ray tracing tool to generate comprehensive wireless channel data, which is subsequently used to train an NN. By directly learning from the data, the algorithm obviates the necessity for manual feature extraction. The NN is trained on diverse channel data from various environments, ensuring its ability to generalize across different scenarios. The study identifies optimal hyperparameters that enhance prediction precision, and importantly, the algorithm's performance remains consistent regardless of the number of or distance between buildings in the environment, highlighting its robustness.

4) Terrain Profiling Models

Analyzing the impact of the surrounding terrain on signal propagation has been a critical component for modeling radio propagation, particularly in short ranges or urban settings. However, given the massive terrain data, such analyses can be prohibitive in terms of computational complexity, for which ML-based approaches can be quite useful in developing a terrain profile. [108] introduces an innovative approach that integrates DL techniques with top-view geographical images. By using DL, the model gains the ability to discern complex patterns and relationships within the data. The geographical images serve as a source of terrain profiling data, allowing the model to implicitly learn the impact of terrain on signal propagation. As a result, it offers a more holistic and accurate understanding of how signal strength behaves in various environments.

Most terrain profiling studies in radio propagation modeling involved 3D maps or satellite image data. [105] uses DNN for modeling where 3D maps are given as input and the accuracy of the model is verified using the dataset generated by the wireless ray-tracing software for different environmental settings. Similarly, [57] uses SVR, RF, and KNN algorithms to predict PL. Training and testing are carried out from simulated results considering the LTE network utilizing a DTM.

For more scalability, some studies have focused on predicting a terrain's field strength from small samples of map or imagery data. The end goal of these approaches is still to make RSS prediction once such field strengths are modeled. For instance, [149] designs a DL-based model that uses smaller samples of geographical information and satellite data. The proposed model consists of two pre-trained network models, ResNet18 [150] and NN, to optimize the parameters to be learned. The output from ResNet18 and NN are then concatenated with another NN model. Performance comparison with an existing CNN-based model [135] shows that the proposed model can predict field strength even with a small amount of data.

B. Data-driven Prediction

Based on the type of ML algorithm used to make predictions, data-driven RSS prediction methods for outdoors can be regression-based or DL-based, which we will delve into next.

1) Regression-based Prediction

Most regression-based studies have used RF approach for predicting the Reference Signal Received Power (RSRP). [102] shows that the RF-based predictor significantly improves the tradeoff between prediction error and the number of measurements needed. In particular, the proposed framework can predict RSRP with 80% less measurement and with the same accuracy on real-world datasets under different scenarios of radio deployment, e.g., small and dense, large, and sparser.

Augmenting the ML approach with a physical propagation model has proved to be useful in attaining better RSRP prediction. [107] chooses the RF algorithm due to its ability to operate with smaller data requirements and offer improved organization. The proposed model leverages geographical information of the selected area alongside ray tracing software to extract multi-path information, which, along with measured data, is utilized for training the model and predicting RSRP. Notably, this approach employs a physical approach to calculate RSRP. When a new location is introduced, the ray tracing software generates multi-path information, which is then used by the predictor to estimate RSRP. The proposed physical propagation model, rooted in measured data and multi-path information, is considered more accurate than the ray tracing model and requires fewer data compared to an RF-based approach, all while achieving superior performance.

In a more lightweight approach, [151] aims to find the location details of the cell towers by using RSS measurements from a crowd-sourced dataset. The framework uses the dataset to build the RSS fingerprint database, and uses a weighted K-means clustering algorithm to predict the locations of cell towers from crowd-sourced data. The trade-off between the light computational complexity of data-driven regression approaches as in [151] and the accuracy of RSS prediction is an open research direction that requires further exploration.

2) DL-based Prediction

Empowering the DL-based methods with real data enables notable improvement in RSS prediction. Datasets for these methods come from various sources ranging from very high-level measurements (e.g., satellite images) to detailed UE-based measurement campaigns. Utilization of satellite images offers a scalable method to prepare datasets. [70] introduces an approach to estimate signal strength through the satellite data presented as a GeoTIFF file [152]. The technique incorporates mixed urban and suburban settings, strategically employed for self-supervised learning to achieve robust representations of the radio environment. To attain supervised learning, the study collects drive test data from 23 BSs, operating at 2.6 GHz. This dataset encompasses essential information like coordinates, RSRP, and Physical Cell Identity, which are utilized for both model training and evaluation. Performance evaluation is carried out with different ML models leveraging satellite imagery alongside DNN algorithms offer a prospect to amplify the geographic generalization of data-driven RSS models. Using a small dataset, [104] performs a comparative study of different ML models, including regression-based as well as DL-based models. The results show that DNN models perform better than the others even with limited data.

Using datasets from UE-based measurement campaigns significantly improves the accuracy of RSS prediction. A mobility handover prediction framework [125] has been proposed to assist handover decision-making. It uses NNs for signal strength prediction and handling uncertainty of the estimation. The model is trained with the RSRP samples from real-world drive tests. Later, subsequent handover probabilities to each cell are derived. In a comparative study, [111] evaluates the performance of different ML algorithms to develop an RSS prediction model for mobile networks. Gaussian Process Regression (GPR) is compared against other ML algorithms. Data is collected from a measurement campaign conducted in the location which contains 5G NR network deployment. Cloud RF planning tool is utilized in this study, that integrates 3D buildings information from OpenStreetMap and clutter data with 10 m resolution. Showing the additional benefit of using real measurements, the results show that the GPR model is the most accurate model for signal strength prediction.

Integrating UE-based measurements with satellite imagery data can further improve the prediction accuracy. In [109], drive test data along with satellite map and cellular configuration are used to form a multimodal model. DNN is used for modeling and training data is obtained from measurements taken in the Cairo and Giza regions. Transfer learning is used to leverage the parallel training processes and converge to the optimum weights that achieve the best predictions.

C. Hybrid Model

CWI models [3] are well suited to make a prediction in urban and suburban regions since they account for both diffraction and reflection effects, and also consider antenna heights. [153] is a good example for CWI-based hybrid models, where PL data from the CWI model and measured data are given as input to train an ANN. A correction-based ANN model is used to the calculate error in predicting PL. The prediction error is, then, considered for predicting the actual PL values. Experimental results show that the proposed model outperformed the data-driven model [154] as well as other CWI models on similar setups.

The ITU has widely recognized and accepted models that are comprehensive in capturing various propagation scenarios. These models, such as ITU-R.526's Cascade Knife Edge [155] and Delta-Bullington models [156], are good at incorporating spatial data and are widely used to provide general coverage estimates. Combining these physics-based models with data-driven learning capabilities of ANNs has proven to significantly improve the signal strength predictions in various environments [84]. In [107], the authors develop a data-assisted physical propagation modeling framework that uses multipath information from ray tracing simulations and real-world measurement data. The model outperforms an RF-based model under a non-uniformly partitioned dataset even with a small dataset. In another study, Adaptive Knowledge-Guided Neural Network (KGNN) propagation model is proposed to make efficient use of the knowledge database from the Hata model [2] to train a low-complexity ANN.

Experimental results show that the KGNN model offers the least RMSE and convergence time when compared with the high complexity ANN-based models [157].

The LDPL model is one of the simpler models to obtain quick results with limited data. It can be customized to fit specific environments by adjusting the PLE. [102] is one of the examples of a hybrid model that utilizes the LDPL physics-based model as the baseline. The authors make an extensive study on data-driven prediction using an RF-based predictor and conduct experiments to validate the results. The study show that the RF-based predictor performs well in places which has enough data to make predictions, but in the places with less data, the model under-performs compared to an LDPL-KNN model [158], [159]. To improve the prediction accuracy in the areas with less data, a hybrid model is introduced. Both data-driven RF-based predictor and LDPL-KNN are combined by using Stacking regression [160]. Experimental results show that the proposed RF-based predictor requires 80% less data than the state-of-the-art data-driven predictors [158], [161] for the same prediction accuracy.

VII. RSS Prediction in Indoor Environments

Attaining accurate RSS prediction indoors has been a major area of research. Being able to accurately localize indoors can be life-saving during an emergency. Yet, due to the difficulty of indoor RSS prediction caused by signal blockage or excessive multi-path effects, traditional positioning systems such as GPS struggle to attain accurate enough location predictions [162]. Hence, most practical indoor positioning systems rely on additional sources of information such as Wi-Fi [163], [164], RFID devices [165], ZigBee [166], [167], visible light [168], acoustics [169], UWB [170], Bluetooth [171] and even magnetic fields to triangulate positions [172], [173].

Developing accurate indoor models often involves collecting data within the building, which also involves collecting the building's structural plan and selecting the positions to place sensors to collect data. Inferring indoor positions based on these sensors can be done by predicting TOA [174], [175], TDOA [176], AOA [177], Channel State Information [178] or RSS [179]. In this survey, we focus on research papers using RSS measurements for estimating approximate distances of deployed sensor nodes. ML algorithms can then be used to analyze this data and create models that reflect the physics of indoor environments.

A. Model-based Prediction

In the model-based prediction method, PL models and geographical data are used to predict RSS and construct a radio map. PL models can be used for indoor localization using RSS data and ML methods by leveraging the relationships between RSS values and the physical distances between devices and Access Points (APs) or beacons. Most of the model-based approaches utilize regression-based prediction with the aim of fast outcome.

Consideration of multi-path effects play a crucial role in model-based RSS prediction in indoor settings. [180] presents a framework for constructing a radio map for indoor localization using the Multiple Path Loss Model (M-PLM). M-PLM reduces positioning errors by employing an interpolation-based smoothing of the wireless signal fluctuations and inherent noise from the signal paths. Then, KNN algorithm is used to predict indoor locations. The proposed M-PLM approach attains 44% more accuracy than conventional interpolation techniques, including single PL models, inverse distance weighting, and Kriging. [181] proposes an indoor positioning system based on PL modeling for a multi-floor environment. The radio map is constructed from the samples selected on the floor and the floor number is localized in the online phase by searching the radio map. The mobile device location is then estimated by using the PL model. In an effort to utilize the knowledge about the indoor environment, [182] constructs radio maps based on a multi-wall PL model along with the features of the indoor environment. Similarly, the authors use the KNN algorithm to verify the accuracy of the indoor positioning system.

Model-based prediction presents an advanced strategy for indoor RSS prediction and localization, showcasing the potential to provide improved accuracy, flexibility, and intricacy. This approach proves particularly valuable in scenarios where achieving precision and comprehending the fundamental physical mechanisms play a pivotal role in the success of the application.

B. Data-driven Prediction

For indoors, data-driven RSS prediction heavily relies on datasets that have an accurate representation of the environment. This is because the eventual radio propagation model is significantly impacted by the geometry and physical features of the indoor environment. To attain such datasets, fingerprinting is used. Fingerprinting involves creating a repository of RSS measurements (a.k.a. fingerprints) collected from precise indoor locations. Utilizing measured RSS values from surveyed sites enables generation of a baseline radio map. The fingerprinting measurements embody traits like resolution and accuracy. Resolution pertains to distinguishing signals based on their unique attributes, while accuracy hinges on Signal-to-Noise Ratio and signal waveform properties. Notably, measurements are singular and fingerprinting doesn't consider post-measurement changes, e.g., décor shifts, furniture repositioning, or foot traffic alterations.

Fingerprinting often pairs with ML techniques like KNN or classification algorithms for enhanced predictions. Inline with this, [183] and [184] explore efficient methods for constructing radio maps through crowd-sourced data, aimed at reducing time and manpower requirements. Crowd-sourced databases and trajectory-matching algorithms are used to avoid tedious and time-consuming data collection. Indoor

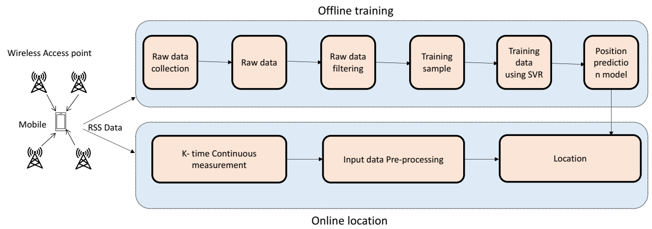

FIGURE 17. Framework of the indoor location algorithm

map processing methods are used to extract the information from the map. Depth-first traversal method is used to extract all possible routes. The trajectory optimization is used to limit the starting and ending points to reduce matching errors. Then, a shape context algorithm is used for matching candidate routes to the optimized trajectories. A radio map is eventually generated from the fingerprint database.

In the realm of indoor RSS-based localization, data-driven prediction presents the benefit of adjusting to real-world scenarios, adeptly managing intricate environments, and harnessing the power of ML for precise and responsive forecasts. This approach tackles the constraints posed by purely model-based or theoretical methods, thereby enhancing the efficacy of indoor localization systems. We, next, delve into regression-based and DL-based efforts in data-driven indoor RSS prediction.

1) Regression-Based Prediction

Regression-based models are helpful for RSS prediction in indoor environments due to their ability to model the relationship between independent variables (factors affecting signal strength) and the dependent variable, i.e., RSS. [185] is a good example of using SVR-based prediction to achieve higher positioning accuracy in indoor localization. The framework of the indoor location algorithm using SVR is shown in 17. To handle complex indoor environmental conditions such as obstruction and changes in node communication range, data filtering rules and K-times continuous measurement method are introduced.

Conversely, *tree-based algorithms* such as DT, RF, and Gradient Boosting excel in capturing complex relationships, handling noisy data, and balancing accuracy and interpretability. When applied to data-driven RSS prediction in indoor environments, they can significantly contribute to more accurate positioning and navigation. A comparison study is made on DT-based indoor positioning techniques [186]. The study focuses on comparing the typical DT and Gradient Boosted Tree algorithms. The experimental results demonstrate that the Gradient Boosted algorithm outperforms the typical DT. [187] introduces a new adaptation of the RF technique for indoor localization, leveraging RSS data from various Wi-Fi sources. It evaluates two localization models: one employs a basic RF approach, while the other fuses individual models crafted for specific APs within the network, using the same RF framework.

Clustering algorithms enhance data-driven RSS prediction in indoor environments by segmenting the data into meaningful clusters, adapting to changes, handling outliers, and enabling more accurate and efficient predictive modeling. [188] proposes the Strongest Access Point information-based clustering algorithm for its ability to accurately represent signal coverage areas within indoor environments. By incorporating real-time signal strength data and considering the most influential APs, this approach enhances the accuracy of indoor positioning, making it a suitable choice for the proposed Wi-Fi RSS fingerprint positioning method.

2) Neural Network (NN) Approaches

NN-based RSS prediction approaches are well suited for indoors due to their ability to learn complex patterns, capture non-linear relationships, and adapt to varying conditions. Research work has been carried out to find the best indoor localization technique using RSS data without using the radio PL model or comparing the radio maps [189]. To improve location accuracy, the authors use a multi-layer NN system to integrate RSS signals and network-boosting techniques. The experimental results show higher location accuracy when compared with Maximum Likelihood Estimator [190], Generalized Regression NN [191], and fingerprinting methods [192]. Another example is [193], which uses a deep Gaussian process model for radio map construction and indoor localization from the sparse training data. The model architecture consists of two stages: An offline training stage and an online estimation stage. In the offline training stage, a measurement database collected from the university hall is utilized. The offline training stage is a twofold process. First, to find the relationship between RSS samples and location, a deep Gaussian process model is used. Next, the Bayesian training method is used to optimize model parameters and construct a radio map. The online localization stage consists of unknown location RSS data. A Bayesian fusion location estimation algorithm is utilized to estimate location.

C. Hybrid model

Indoor RSS prediction benefits from hybrid models that amalgamate the advantages of empirical data-driven and physics-based model-driven methodologies, resulting in improved precision, flexibility, and resilience. These hybrid models prove especially advantageous in situations demanding a comprehensive and dependable prediction framework to bolster indoor localization and its associated applications.

1) Model-Data Fusion

The most common hybrid method is model-data fusion, which aims to capitalize on the advantages of both model-based and data-driven approaches using various merging methods. An effective example of this is to use interpolation. [194] proposes fast radio map construction by using an adaptive PL model interpolation. RSS fingerprints are obtained from crowd-sourced data with sparse AP data (only 15% of APs). To calculate RSS, the PL model for sparsely distributed APs is built by using the least squares algorithm. Parameters obtained from PL calculation such as random noise, RSS loss and threshold restraint are taken into consideration to calculate RSS and construct a radio map.

When the measurement data is annotated or limited, the model-data fusion method can be more sophisticated. [195] addresses the challenge of constructing an accurate Power Spectral Density map from the limited distributed measurements collected from crowd-sourced data. The study blends the advantages of model-based techniques, utilizing mathematical models for accuracy, and data-driven methods, capitalizing on real-world measurements for adaptability.

In a similar approach, [196] utilizes crowd-sourced data to construct the radio map. The study addresses challenges associated with the crowd-sourced measurement database, such as inaccurate sample annotation, unequal sample dimensionality, measurement device diversity, and nonuniform spatial distribution. To mitigate these challenges, the study employs four distinct algorithms. A grid fingerprint is established to demarcate the area, and for grids with sufficient samples, a density-based clustering algorithm is introduced to eliminate outliers. To select vital signals, a threshold-based selection approach is adopted. For grids with limited samples, a fingerprint interpolation algorithm is used to create device-specific fingerprints. Further enhancing the process, a device calibration algorithm is introduced to derive a unified grid fingerprint from different device-specific fingerprints. Finally, an improved nearest-neighbour algorithm is presented for refined online positioning.

An effective fusion method is to dynamically switch between different prediction models based on contextual information [197], such as device mobility, environment changes, or available signal sources. [198] presents a self-calibrating and self-adaptive localization method for a Wi-Fi network. FSPL and ITU models are used as propagation models for simulation. Localization is carried out in three stages: data acquisition, PL modeling, and propagation simulation. Data is acquired by sending periodic queries to the Wi-Fi router from a server. LDPL model is used for propagation parameter estimation. Finally, ITU indoor propagation model is used for propagation simulation.

2) Manifold Alignment

A frugal hybrid approach is the manifold alignment method which utilizes both the model-based and data-driven methods in parallel and combines the predictions from them. This approach uses mathematical rules and learns from real-world data patterns. For example, in [199], instead of building a complete fingerprinted map of the indoor environment, simultaneous construction of radio maps from localized RSS information is proposed. A manifold alignment scheme is used to transfer knowledge from the source dataset to the destination dataset by learning the underlying relationship between source and destination datasets in a low dimensional space. Radio map construction involves an offline deployment phase and an online localization phase. In the proposed scheme, the algorithm collects the environmental information and builds the grid point system on the floor plan to build the coordinates. Deployment load is reduced since it removes the need to know the positions of APs.

VIII. Challenges and Open Research

Even though radio propagation modeling and RSS prediction are heavily studied topics, many research challenges remain and several major open research directions persist. We, next, discuss some of the notable open research areas in this space.

- **More and Diverse Measurements**. One of the biggest challenges in RSS prediction is data collection. Manual data collection is an expensive and time-consuming process. Proper planning and expertise are needed to collect data to fit the specific requirements. Outdoor data collection requires lots of driving through selected paths to collect the desirable quantity of data. Considering the importance of environmentally dependent features in the studies, data should be collected in different seasons and weather conditions. This will make the data collection process quite lengthy and challenging. More research is needed to study how environmental variables affect RSS prediction. An alternate option to data collection is using publicly available crowdsourced data collected from the websites such as OpenCellID [55] and OpenBmap [56]. Obtaining a considerable quantity with a wide variety of features is quite challenging in both indoor as well as outdoor environments. The most common problems with crowdsourced data are a lack of volunteers in the planned survey area and a lack of data from private areas. Scalable and low-cost methods of radio data collection are very much needed as studies clearly showed that more real data measurements make RSS predictions more accurate.
- **Better Models for Indoor Settings.** Multipath effects are major challenges in indoor RSS prediction. Due to multipath effects, obtaining a single LOS signal and estimating the distance between the Tx and the Rx is challenging. To tackle these issues, we need to utilize complex signal processing techniques that can identify the LOS signal and minimize the effects of multipath signals. To assist with better indoor localization, more research work is needed in effective multipath reduction and noise suppressing algorithms. Further, consideration of physical properties of an indoor environment is needed when building a PL model. For exact accuracy, this requires development of a specific model for every indoor environment separately, which is too costly.

A challenge is to build models that can be used as a baseline for various common indoor settings. For example, ML-based models can be trained for a typical office, living room, conference room.

- **Efficient Solutions with Sparse Data**. Given the difficulties in collecting radio data, there is always a need for more accurate and efficient models for radio propagation. More research is required to find efficient ways to obtain better accuracy with missing or sparse data. Fundamental insights into the boundaries on radio propagation model accuracy, particularly on the emerging super-6 GHz spectrum technologies, are much needed. Answers to questions such as "What is the maximum error on the RSS prediction for a mmWave link?" will be crucial to our understanding of the limits of what is possible with sparse data.
- **Privacy-Preserving Data Sharing**. Indoor data collection requires lots of walking on the desired path. To get a complete set of data, the data collector may require accessing all the areas (including private places) in the building or organization. Complying with privacy regulations and obtaining proper permission from the building authorities are quite challenging. More research is needed on privacy-preserving radio data measurements and sharing.
- **Data Integration to Improve Accuracy**. To reduce the need for more radio data collection, techniques that utilize non-radio sources are needed. Integrating measured data with satellite image/geographic information systems is a complex process. But, one can achieve better accuracy while integrating different forms of data [74], [200], [201]. More research is needed to reduce complexity of these integrative methods and improvise learning as part of the integration.
- **Real-Time Data Processing in Indoor and Urban Settings**. The indoor scenarios as well as some urban settings are volatile when compared to other settings. Changes in the furniture set, movement of a parked vehicle, adding decors in the walls, or walking traffic patterns may cause changes in the radio data. These changes require the radio measurement datasets and RSS predictions based on those datasets to be updated in a timely manner. Solutions that can swiftly handle these dynamics in the indoor settings are needed. More research is needed to study how indoor changes affect the radio propagation model and develop RSS prediction mechanisms that adapt to real-time changes.
- **Comprehensive DL Models**. Many of the available ML algorithms are not specifically tailored for RSS prediction. Finding and capturing complex relationships, limited data availability, and model complexity and scalability are notable challenges in using existing ML algorithms for PL modeling. There are a few studies [70], [105], [106], [109], [110], we already reviewed above, that used novel DL architectures specifically for RSS prediction. Developing comprehensive DL architectures, for RSS prediction modeling, which can effectively capture both system- and environment-dependent features from the input data would be a better solution for future readiness.
- **Transfer Learning**. A major challenge that comes with RSS prediction is the lack of data availability. Transferring knowledge learned from one scenario to another different scenario will be an efficient way to improve model performance but, at the same time, it is challenging. There is a need for better capturing of which scenarios are similar to each other so that radio data from measured scenarios can be utilized for learning the models for a new scenario. Exploring transfer learning would be a worthy research direction to improve the scalability of RSS prediction by reducing the need for data measurements.
- **Automated Hyperparameter Tuning**. Hyperparameter setting plays a crucial role in the performance of the ML models. Though we choose the better algorithm to make PL prediction, the performance of the ML algorithm is decided based on the hyperparameter setting. For instance, the hyperparameter setting includes the number of neighbours in the KNN algorithm, the size of the tree and depth of the tree in tree-based modeling, and regularization coefficients and type kernel function in SVR-based modeling. Further, in ANN-based modeling, the number of hidden layers and the number of neurons in each layer decide the performance of the algorithm. In RSS prediction, tuning hyperparameters is one of the challenging areas to work on. Exploring efficient automated hyperparameter settings is one of the open areas to research.
- **ML-Integrated Models**. Integrating ML models with empirical radio models improves the accuracy and generalization in RSS prediction since it combines the fundamental principles of empirical modeling and the data-driven capability of ML models. Integrating ML models with empirical models is quite challenging because those two models might have different data requirements. For example, empirical models might rely on some specific measurements that might not perfectly align with the data need for ML models. Further, as summarized in Figure 16, how to reconcile environmental features with the ML model needs to be investigated. ML model can naturally learn the environmental feature but those insights may conflict with the physical features of the environment due to errors in signal measurements. Weeding out the measurement errors while keeping the influence of the critical environmental features is balancing act. Capturing complex environmental relationships and leveraging data-driven insights are a few of the open research areas that come with ML integration.

IX. Summary

We surveyed the application of ML in PL modeling and RSS prediction both in outdoor and indoor scenarios. We outlined different types of raw radio data used as parameters for radio propagation modeling. We made a detailed analysis of the PL modeling procedure using ML algorithms. A comprehensive review of various models and methodologies demonstrated that ML algorithms can efficiently predict PL with higher accuracy compared to traditional methods. Our survey provides valuable insights into the current state of ML-based RSS prediction, offering guidance for researchers and practitioners in the field. Summarizing the insights from the survey, we also offered several new and existing directions of research that will be needed for radio propagation modeling.

We discussed RSS prediction methodologies for both outdoor and indoor scenarios. Throughout our survey, it became evident that the success of ML-based PL prediction heavily depends on several factors such as the size and quality of the dataset, the choice of features, and the appropriate selection of the ML algorithm. By leveraging the power of ML in radio propagation modeling, more efficient and reliable wireless communication networks will be possible in the future. While various models have shown promising results, there is no one-size-fits-all approach, and researchers should carefully consider the specific characteristics of their wireless environment when choosing the most suitable prediction method.

We focused on the sub-6 GHz bands in this survey. Due to increased recent activity in super-6 GHz bands, a sizable literature is forming on channel sounding and propagation modeling in mid-band (also known as FR3) ($\approx$7-24 GHz), mmWave ($\approx$30-300 GHz), Terahertz ($\approx$0.1-100 GHz), and optical ($\approx$300 GHz - 3,000 THz) bands. These bands have distinct characteristics in terms of the types of attenuation they experience, the range they can reach, and the antenna designs they entertain. A major impediment has been that empirical studies of these bands are limited. Upper mid-band, for example, has recently become available for civilian access and the measurement studies of these bands in open literature are very few, providing little opportunity to use ML-based modeling on them. Similarly, mmWave and Terahertz bands have recently gotten the attention of the wireless community and the antenna systems to access them have not been widely available, limiting the possibility of channel sounding or PL measurement experiments. Once more empirical literature on these emerging frequency bands become available, a survey focusing on the propagation models and the use of ML for propagation modeling on them will be a worthy effort.

Finally, the influence of antenna parameters on the channel performance is a worthy topic for another survey. In terms of radio propagation modeling, antenna parameters are captured by gain or directivity of the antennas involved. In this paper, we explained how antenna directivity relates to typical gain parameters of a radio propagation model. However, surveying the influence of antennas on channel performance would be a worthy future work. There are many different types of antennas even in the sub-6 GHz frequencies. Investigation of antenna parameters and their interaction with attained channel performance would be enlightening.

REFERENCES

[1] J.-Y. Lee, M. Y. Kang, and S.-C. Kim, “Path loss exponent prediction for outdoor millimeter wave channels through deep learning,” in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–5.

[2] M. Hata, “Empirical formula for propagation loss in land mobile radio services,” *IEEE Transactions on Vehicular Communications*, vol. 29, no. 3, pp. 317–325, Aug. 1980.

[3] P. E. Mogensen and J. Wigard, “Cost action 231: Digital mobile radio towards future generation system, final report.” in *Section 5.2: On antenna and frequency diversity in GSM. Section 5.3: Capacity study of frequency hopping GSM network*, 1999.

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

[5] ——, “Ray tracing for radio propagation modeling: Principles and applications,” *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[6] M. Yang, R. He, B. Ai, C. Huang, C. Wang, Y. Zhang, and Z. Zhong, “AI-enabled data-driven channel modeling for future communications,” *IEEE Communications Magazine*, pp. 1–7, 2023.

[7] B. Matthiesen, A. Zappone, E. Jorswieck, and m. Debbah, “Deep learning for optimal energy-efficient power control in wireless interference networks,” 12 2018.

[8] 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 Vehicular Technology Magazine*, vol. 14, no. 3, pp. 60–69, 2019.

[9] P. Tilghman, “Will rule the airwaves: A DARPA grand challenge seeks autonomous radios to manage the wireless spectrum,” *IEEE Spectrum*, vol. 56, no. 6, pp. 28–33, 2019.

[10] S. Mustafa, S. Seth, M. Yuksel, and M. Rahman, “Cellular service with settlement-free peering,” in *2021 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2021, pp. 153–162.

[11] F. A. P. d. Figueiredo, D. Stojadinovic, P. Maddala, R. Mennes, I. Jabandžić, X. Jiao, and I. Moerman, “Scatter phy: A physical layer for the DARPA spectrum collaboration challenge,” in *Proceedings of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–6.

[12] C. Bowyer, D. Greene, T. Ward, M. Menendez, J. Shea, and T. Wong, “Reinforcement learning for mixed cooperative/competitive dynamic spectrum access,” in *Proceedings of IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–6.

[13] J. M. Shea and T. F. Wong, “A deep Q-Learning dynamic spectrum sharing experiment,” in *Proceedings of IEEE International Conference on Communications (ICC)*, 2021, pp. 1–6.

[14] R. Mennes, J. Struye, C. Donato, M. Camelo, I. Jabandžić, S. Giannoulis, I. Moerman, and S. Latré, “Collaborative flow control in the DARPA spectrum collaboration challenge,” *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2024–2038, 2020.

[15] M. Taye, “Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions,” *Computers*, vol. 12, p. 91, 04 2023.

[16] C. Phillips, D. Sicker, and D. Grunwald, “A survey of wireless path loss prediction and coverage mapping methods,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 255–270, 2013.

[17] A. Hrovat, G. Kandus, and T. Javornik, “A survey of radio propagation modeling for tunnels,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 658–669, 2014.

[18] M. E. Diago-Mosquera, A. Aragón-Zavala, and G. Castañón, “Bringing it indoors: A review of narrowband radio propagation modeling for enclosed spaces,” *IEEE Access*, vol. 8, pp. 103 875–103 899, 2020.

[19] A. Al-Saman, M. Cheffena, O. Elijah, Y. A. Al-Gumaei, S. K. Abdul Rahim, and T. Al-Hadhrami, "Survey of millimeter-wave propagation measurements and models in indoor environments," *Electronics*, vol. 10, no. 14, p. 1653, July 2021.

[20] H. Chiroma, P. Nickolas, N. Faruk, E. Alozie, I.-F. Y. Olayinka, K. S. Adewole, A. Abdulkarim, A. A. Oloyede, O. A. Sowande, S. Garba, A. D. Usman, L. S. Taura, and Y. A. Adediran, "Large scale survey for radio propagation in developing machine learning model for path losses in communication systems," *Scientific African*, vol. 19, p. e01550, 2023.

[21] J. Mladenović, A. Nešković, and N. Nešković, "An overview of propagation models based on deep learning techniques," *International Journal of Electrical Engineering and Computing*, vol. 6, no. 1, pp. 18–25, 2022.

[22] 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 I: Channel characterization and antenna-channel optimization," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 6, pp. 3939–3954, 2022.

[23] ——, "Artificial intelligence enabled radio propagation for communications—part II: Scenario identification and channel modeling," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 6, pp. 3955–3969, 2022.

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

[25] S. M. Aldossari and K. Chen, "Machine learning for wireless communication channel modeling: An overview," *Wireless Personal Communications*, vol. 106, p. 41–70, 2019.

[26] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[27] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[28] P. Roy and C. Chowdhury, "A survey of machine learning techniques for indoor localization and navigation systems," *J. Intell. Robotics Syst.*, vol. 101, no. 3, mar 2021. [Online]. Available: https://doi.org/10.1007/s10846-021-01327-z

[29] A. Nessa, B. Adhikari, F. Hussain, and X. N. Fernando, "A survey of machine learning for indoor positioning," *IEEE Access*, vol. 8, pp. 214 945–214 965, 2020.

[30] H. T. Friis, "A note on a simple transmission formula," *IRE Proc.*, pp. 254–256, May 1946.

[31] M. Viswanathan, *Wireless Communication Systems in Matlab*. Mathuranathan Viswanathan, 2020.

[32] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[33] A. Neskovic, N. Neskovic, and G. Paunovic, "Modern approaches in modeling of mobile radio systems propagation environment," *IEEE Communications Surveys & Tutorials*, vol. 3, no. 3, pp. 2–12, 2000.

[34] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Transactions on Vehicular Technology*, vol. 29, no. 3, pp. 317–325, 1980.

[35] R. V. Akhpashev and A. V. Andreev, "Cost 231 hata adaptation model for urban conditions in lte networks," in *2016 17th International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices (EDM)*, 2016, pp. 64–66.

[36] A. Deme, D. Dajab, and D. Nyap, "Computer analysis of the cost 231 hata model and least squares approximation for path loss estimation at 900mhz on the mountain terrains of the jos-plateau, nigeria," *Computer Engineering and Intelligent Systems*, vol. 4, no. 9, pp. 39–48, 2013.

[37] Y. A. Alqudah, "On the performance of cost 231 walfisch ikegami model in deployed 3.5 ghz network," in *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, 2013, pp. 524–527.

[38] "COST Action 231 – digital mobile radio towards future generation systems: Final report," European Commission and Directorate-General for the Information Society and Media, 1999.

[39] R. Lopez and K. Vlahodimitropulos, *Adaptive Propagation Prediction using Lee's Model in a Non-Homogeneous Environment*, ser. The Springer International Series in Engineering and Computer Science. Boston, MA: Springer US, 1995, vol. 309, pp. 179–184. [Online]. Available: https://doi.org/10.1007/978-1-4757-2368-7_13

[40] W. Lee. (2024) Microcell prediction model. Scholarpedia. [Online]. Available: $http : //www.scholarpedia.org/article/User_{t}alk :$ $William_{C \cdot Y \cdot L}ee/Proposed/Lee_{m}icrocell_{p}rediction_{m}odel$

[41] W. C. Lee, *Coverage and Antennas. In: 3rd Ed. Wireless & Cellular Telecommunication*. McGraw-Hill, 2005.

[42] W. C. Y. Lee, *Prediction of Propagation Loss*, 2nd ed. John Wiley & Sons, Inc., 1993.

[43] W. Lee, *Prediction of Propagation Loss. In: 2nd Ed. Mobile Communications Design Fundamentals*. McGraw-Hill, 1997.

[44] A. Akanni and N. Anie, "Validation of egli model and estimation of pathloss exponent of a radio signal at vhf band in hilly terrain," 10 2020.

[45] J. J. Egli, "Radio propagation above 40 mc over irregular terrain," *Proceedings of the IRE*, vol. 45, no. 10, pp. 1383–1391, 1957.

[46] Wikipedia contributors, "Egli model - wikipedia," https://en.wikipedia.org/wiki/Egli_model, Accessed: 2024.

[47] International Telecommunication Union, "ITU-R P.1546-6: Propagation by Diffraction," https://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1546-6-201908-I!!PDF-E.pdf, 2019.

[48] E. ÖstlinOstlin, H. Suzuki, and H.-J. Zepernick, "Evaluation of the propagation model recommendation itu-r p.1546 for mobile services in rural australia," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 1, pp. 38–51, 2008.

[49] E. Ostlin, H. Suzuki, and H.-J. Zepernick, "Comparison and evaluation of itu-r recommendation p.1546 versions," in *2006 IEEE 63rd Vehicular Technology Conference*, vol. 6, 2006, pp. 2896–2900.

[50] D. P. Grabowsky, J. M. Conrad, and A. F. Browne, "Limited log-distance path loss model path loss exponent estimation using deep deterministic policy gradient," in *SoutheastCon 2021*, 2021, pp. 1–6.

[51] Z. El Khaled, W. Ajib, and H. Mcheick, "Log distance path loss model: Application and improvement for sub 5 ghz rural fixed wireless networks," *IEEE Access*, vol. 10, pp. 52 020–52 029, 2022.

[52] L.-W. Yeh, M.-S. Hsu, Y.-F. Lee, and Y.-C. Tseng, "Indoor localization: Automatically constructing today's radio map by irobot and rfids," in *SENSORS, 2009 IEEE*, 2009, pp. 1463–1466.

[53] W. A. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, and B. Sebire, "Minimization of drive tests solution in 3gpp," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 28–36, 2012.

[54] J. Johansson, W. A. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3gpp release 11," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 36–43, 2012.

[55] "Opencellid," https://www.opencellid.org/.

[56] "Openbmap," https://www.openbmap.org/, n.d., accessed: July 10, 2023.

[57] R.-T. Juang, J.-Q. Lin, and H.-P. Lin, "Machine learning-based path loss modeling in urban propagation environments," in *2021 30th Wireless and Optical Communications Conference (WOCC)*, 2021, pp. 291–292.

[58] A. Marey, M. Bal, H. F. Ates, and B. K. Gunturk, "Pl-gan: Path loss prediction using generative adversarial networks," *IEEE Access*, vol. 10, pp. 90 474–90 480, 2022.

[59] N. Moraitis, L. Tsipi, and D. Vouyioukas, "Machine learning-based methods for path loss prediction in urban environment for lte networks," in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2020, pp. 1–6.

[60] J. Thrane, D. Zibar, and H. L. Christiansen, "Model-aided deep learning method for path loss prediction in mobile communication systems at 2.6 ghz," *Ieee Access*, vol. 8, pp. 7925–7936, 2020.

[61] L. Peng, S. Fang, Y. Fan, M. Wang, and Z. Ma, "A method of noise reduction for radio communication signal based on ragan," *Sensors*, vol. 23, p. 475, 01 2023.

[62] K. Ahmad and S. Hussain, "Machine learning approaches for radio propagation modeling in urban vehicular channels," *IEEE Access*, vol. 10, pp. 113 690–113 698, 2022.

[63] U. Masood, H. Farooq, and A. Imran, "A machine learning based 3d propagation model for intelligent future cellular networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[64] S. P. Sotiroudis, G. Athanasiadou, G. V. Tsoulos, C. Christodoulou, and S. K. Goudos, "Ensemble learning for 5g flying base station path loss modelling," in *2022 16th European Conference on Antennas and Propagation (EuCAP)*, 2022, pp. 1–4.
[65] G. Vergos, S. P. Sotiroudis, G. Athanasiadou, G. V. Tsoulos, and S. K. Goudos, "Comparing machine learning methods for air-to-ground path loss prediction," in *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2021, pp. 1–4.
[66] T. Oyedare, V. K. Shah, D. J. Jakubisin, and J. H. Reed, "Interference suppression using deep learning: Current approaches and open challenges," *IEEE Access*, vol. 10, pp. 66 238–66 266, 2022.
[67] A. Sonny, A. Kumar, and L. R. Cenkeramaddi, "A survey of application of machine learning in wireless indoor positioning systems," 2024.
[68] S. I. Popoola, S. Misra, and A. A. Atayero, "Outdoor path loss predictions based on extreme learning machine," *Wireless Personal Communications*, vol. 99, no. 1, pp. 441–460, 2018.
[69] R. Azoulay, E. Edery, Y. Haddad, and O. Rozenblit, "Machine learning techniques for received signal strength indicator prediction," 07 2023.
[70] M. Sousa, P. Vieira, M. P. Queluz, and A. Rodrigues, "An ubiquitous 2.6 ghz radio propagation model for wireless networks using self-supervised learning from satellite images," *IEEE Access*, vol. 10, pp. 78 597–78 615, 2022.
[71] O. Ahmadien, H. F. Ates, T. Baykas, and B. K. Gunturk, "Predicting path loss distribution of an area from satellite images using deep learning," *IEEE Access*, vol. 8, pp. 64 982–64 991, 2020.
[72] T. Imai, K. Kitao, and M. Inomata, "Radio propagation prediction model using convolutional neural networks by deep learning," in *2019 13th European Conference on Antennas and Propagation (EuCAP)*, 2019, pp. 1–5.
[73] Y. Zhang, J. Wen, G. Yang, Z. He, and X. Luo, "Air-to-air path loss prediction based on machine learning methods in urban environments," *Wireless Communications and Mobile Computing*, vol. 2018, 2018.
[74] Y. Zhang, J. Wen, G. Yang, Z. He, and J. Wang, "Path loss prediction based on machine learning: Principle, method, and data expansion," *Applied Sciences*, vol. 9, no. 9, p. 1908, 2019.
[75] H. Kim, W. Jin, and H. Lee, "mmwave path loss modeling for urban scenarios based on 3d-convolutional neural networks," in *2022 International Conference on Information Networking (ICOIN)*. IEEE, 2022, pp. 441–445.
[76] K. Inoue, K. Ichige, T. Nagao, and T. Hayashi, "On the building map for radio propagation prediction using machine learning," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 842–847.
[77] 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. 101 787–101 808, 2021.
[78] S. P. Sotiroudis, S. K. Goudos, and K. Siakavara, "Neural networks and random forests: A comparison regarding prediction of propagation path loss for nb-iot networks," in *2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2019, pp. 1–4.
[79] S. Sotiroudis, S. Goudos, K. Gotsis, K. Siakavara, and J. Sahalos, "Modeling by optimal artificial neural networks the prediction of propagation path loss in urban environments," in *2013 IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC)*. IEEE, 2013, pp. 599–602.
[80] S. P. Sotiroudis and K. Siakavara, "Mobile radio propagation path loss prediction using artificial neural networks with optimal input information for urban environments," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 10, pp. 1453–1463, 2015.
[81] J.-Y. Lee, M. Y. Kang, and S.-C. Kim, "Path loss exponent prediction for outdoor millimeter wave channels through deep learning," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–5.
[82] G. Panda, R. K. Mishra, and S. S. Palai, "A novel site adaptive propagation model," 2005.
[83] I. Popescu, A. Kanstas, E. Angelou, L. Nafornita, and P. Constantinou, "Applications of generalized rbf-nn for path loss prediction," in *The 13th IEEE international symposium on personal, indoor and mobile radio communications*, vol. 1. IEEE, 2002, pp. 484–488.
[84] G. P. Ferreira, L. J. Matos, and J. M. Silva, "Improvement of outdoor signal strength prediction in uhf band by artificial neural network," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 12, pp. 5404–5410, 2016.
[85] S. P. Sotiroudis, K. Siakavara, and J. N. Sahalos, "A neural network approach to the prediction of the propagation path-loss for mobile communications systems in urban environments," *Piers Online*, vol. 3, no. 8, pp. 1175–1179, 2007.
[86] R. D. Timoteo, D. C. Cunha, and G. D. Cavalcanti, "A proposal for path loss prediction in urban environments using support vector regression," in *Proc. Advanced Int. Conf. Telecommun*, 2014, pp. 1–5.
[87] 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, 2020.
[88] E. Ostlin, H.-J. Zepernick, and H. Suzuki, "Macrocell path-loss prediction using artificial neural networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 2735–2747, 2010.
[89] 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 Networks*, vol. 27, no. 6, pp. 4169–4188, 2021.
[90] I. Fernández Anitzine, J. A. Romo Argota, and F. P. Fontán, "Influence of training set selection in artificial neural network-based propagation path loss predictions," *International Journal of Antennas and Propagation*, vol. 2012, 2012.
[91] 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, 2020.
[92] J. Zhang, L. Liu, Y. Fan, L. Zhuang, T. Zhou, and Z. Piao, "Wireless channel propagation scenarios identification: A perspective of machine learning," *IEEE Access*, vol. 8, pp. 47 797–47 806, 2020.
[93] S. Bhandari, A. K. Kukreja, A. Lazar, A. Sim, and K. Wu, "Feature selection improves tree-based classification for wireless intrusion detection," in *Proceedings of ACM International Workshop on Systems and Network Telemetry and Analytics*, 2020, p. 19–26.
[94] L. Tao, J. Lu, X. Wang, and L. Qian, "Feature engineering based intelligent wireless propagation model for RSRP prediction," *IOP Conference Series: Materials Science and Engineering*, vol. 768, no. 4, p. 042038, mar 2020.
[95] Q. Zheng, R. He, B. Ai, C. Huang, W. Chen, Z. Zhong, and H. Zhang, "Channel non-line-of-sight identification based on convolutional neural networks," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1500–1504, 2020.
[96] Z. Xiao, H. Wen, A. Markham, N. Trigoni, P. Blunsom, and J. Frolik, "Non-line-of-sight identification and mitigation using received signal strength," *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1689–1702, 2015.
[97] Y. Wu, J. Kokkoniemi, C. Han, and M. Juntti, "Interference and coverage analysis for terahertz networks with indoor blockage effects and line-of-sight access point association," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 1472–1486, 2021.
[98] Y. Xue, X. Zheng, and V. K. N. Lau, "Line-of-sight MIMO for high capacity millimeter wave backhaul in fdd systems," *Journal of Communications and Information Networks*, vol. 5, no. 2, pp. 177–193, 2020.
[99] C. R. Anderson, "An integrated terrain and clutter propagation model for 1.7 GHz and 3.5 GHz spectrum sharing," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 7, pp. 5804–5818, 2022.
[100] F. Wang and H. Li, "Power allocation for coexisting multicarrier radar and communication systems in cluttered environments," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1603–1613, 2021.
[101] Y. Liu, J. Dong, W. Huangfu, J. Liu, and K. Long, "3.5 ghz outdoor radio signal strength prediction with machine learning based on low-cost geographic features," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 6, pp. 4155–4170, 2022.
[102] E. Alimpertis, A. Markopoulou, C. Butts, and K. Psounis, "City-wide signal strength maps: Prediction with random forests," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2536–2542. [Online]. Available: https://doi.org/10.1145/3308558.3313726

[103] G. Cerri, M. Cinalli, F. Michetti, and P. Russo, "Feed forward neural networks for path loss prediction in urban environment," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 11, pp. 3137–3139, 2004.

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

[105] E. Krijestorac, S. Hanna, and D. Cabric, "Spatial signal strength prediction using 3d maps and deep learning." [Online]. Available: https://arxiv.org/abs/2011.03597

[106] O. Ozyegen, S. Mohammadjafari, K. E. mokhtari, M. Cevik, J. Ethier, and A. Basar, "An empirical study on using cnns for fast radio signal prediction," 2020. [Online]. Available: https://arxiv.org/abs/2006.09245

[107] Y. Jiang, S. Zhou, X. Zhong, Y. Jiang, and J. Li, "Data-assisted physical propagation model for signal strength prediction," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1242–1248.

[108] J. Thrane, B. Sliwa, C. Wietfeld, and H. L. Christiansen, "Deep learning-based signal strength prediction using geographical images and expert knowledge," *CoRR*, vol. abs/2008.07747, 2020. [Online]. Available: https://arxiv.org/abs/2008.07747

[109] M. T. Waheed, Y. Fahmy, and A. Khattab, "Deepchannel: Robust multimodal outdoor channel model prediction in lte networks using deep learning," *IEEE Access*, vol. 10, pp. 79 289–79 300, 2022.

[110] R. Levie, C. Yapar, G. Kutyniok, and G. Caire, "Radiounet: Fast radio map estimation with convolutional neural networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 6, pp. 4001–4015, 2021.

[111] M. F. Ahmad 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. 55 782–55 793, 2022.

[112] S. Tadik, M. A. Varner, F. Mitchell, and G. D. Durgin, "Augmented RF propagation modeling," *IEEE Journal of Radio Frequency Identification*, vol. 7, pp. 211–221, 2023.

[113] K. J. Jang, S. Park, J. Kim, Y. Yoon, C.-S. Kim, Y.-J. Chong, and G. Hwang, "Path loss model based on machine learning using multi-dimensional gaussian process regression," *IEEE Access*, vol. 10, pp. 115 061–115 073, 2022.

[114] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, "Combining multiple feature-ranking techniques and clustering of variables for feature selection," *IEEE Access*, vol. 7, pp. 151 482–151 492, 2019.

[115] M. Vogel and C. J. Reddy, "5G network planning optimization using machine learning," in *Proceedings of IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (USNC-URSI)*, 2023, pp. 1185–1186.

[116] T. Peng, Y. Guo, Y. Wang, G. Chen, F. Yang, and W. Chen, "An interference-oriented 5G radio resource allocation framework for ultradense networks," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 618–22 630, 2022.

[117] Y. Guo, C. Hu, T. Peng, H. Wang, and X. Guo, "Regression-based uplink interference identification and SINR prediction for 5G ultra-dense network," in *Proceedings of IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[118] L. L. Oliveira, L. A. O. Jr., G. W. A. Silva, R. D. A. Timoteo, and D. C. Cunha, "An RSS-based regression model for user equipment location in cellular networks using machine learning," *Wireless Networks*, vol. 25, p. 4839–4848, 2019.

[119] European Communities, "Digital mobile radio towards future generation systems," European Communities, Tech. Rep., 1999.

[120] V. Abhayawardhana, I. Wassell, D. Crosby, M. Sellars, and M. Brown, "Comparison of empirical propagation path loss models for fixed wireless access systems," in *Proceedings of the 2005 IEEE 61st Vehicular Technology Conference VTC'05-Spring*, Stockholm, Sweden, May-June 2005, pp. 73–77.

[121] A. Name. (Year, Month) Title of the article. [Online]. Available: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

[122] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin, "Parallel boosted regression trees for web search ranking," in *Proceedings of ACM International Conference on World Wide Web*, 2011, p. 387–396.

[123] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, 2016, pp. 785–794.

[124] V. Priyashman and W. Ismail, "Signal strength and read rate prediction modeling using machine learning algorithms for vehicular access control and identification," *IEEE Sensors Journal*, vol. 19, no. 4, pp. 1400–1411, 2019.

[125] S. Wu, J. Ren, T. Zhao, and Y. Wang, "Machine learning based signal strength and uncertainty prediction for mec mobility management," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 1–5.

[126] E. Stlin, H. Suzuki, and H.-J. Zepernick, "Evaluation of the propagation model recommendation itu-r p.1546 for mobile services in rural australia," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 38–51, Jan. 2008.

[127] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural Network Design*. PWS Pub. Co., 1996.

[128] R. Penrose, "A generalized inverse for matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, 1955.

[129] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[130] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[131] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[132] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[133] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[134] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P.-H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *ArXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2006.07733

[135] J. Thrane, D. Zibar, and H. Christiansen, "Model-aided deep learning method for path loss prediction in mobile communication systems at 2.6 ghz," *IEEE Access*, vol. 8, pp. 7925–7936, 2020.

[136] S. Sun, T. Rappaport, T. Thomas, A. Ghosh, H. Nguyen, I. Kovács, and et al., "Investigation of prediction accuracy sensitivity and parameter stability of large-scale propagation path loss models for 5g wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 2843–2860, 2016.

[137] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 702–716.

[138] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, N.J.: Prentice Hall PTR, 2002.

[139] Wireless insite em propagation software. [Online]. Available: https://www.remcom.com/wireless-insite-em-propagation-software

[140] Suplacemaker. [Online]. Available: https://www.suplacemaker.com

[141] Sketchup. [Online]. Available: https://www.sketchup.com

[142] R. Wahl, G. Wölfle, P. Wildbolz, and F. Landstorfer, "Dominant path prediction model for urban scenarios," in *Proc. IST Mobile Wireless Commun.*, 2005, pp. 1–5.

[143] T. Rautiainen, G. Wölfle, and R. Hoppe, "Verifying path loss and delay spread predictions of a 3d ray tracing propagation model in urban environment," in *Proc. IEEE 56th Veh. Technol. Conf.*, vol. 4, Dec. 2002, pp. 2470–2474.

[144] Radiomapseer. [Online]. Available: https://RadioMapSeer.github.io

[145] (2017) Planet dump. [Online]. Available: https://planet.osm.org

[146] O. Ronneberger, O. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[147] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157.

[148] H. F. Ates, S. M. Hashir, T. Baykas, and B. K. Gunturk, "Path loss exponent and shadowing factor prediction from satellite images using deep learning," *IEEE Access*, vol. 7, pp. 101 366–101 375, 2019.

[149] M. Zhou, W. Shao, Y. Liu, and X. Yang, "Field strength prediction based on deep learning under small sample data," *Electronics Letters*, vol. 58, no. 23, pp. 857–859, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ell2.12631

[150] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[151] M. Rahman, M. A. Hossain, and M. Yuksel, "Multi-operator cell tower locations prediction from crowdsourced data," in *2021 International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2021, pp. 1–9.

[152] S. S. Mahammad and R. Ramakrishnan. (2017) Geotiff - a standard image file format for gis applications. Retrieved on 2017-02-01. [Online]. Available: https://example.com/geotiff

[153] S. Cheerla, V. Ratnam, and H. Borra, "Neural network-based path loss model for cellular mobile networks at 800 and 1800 mhz bands," *AEU - International Journal of Electronics and Communications*, vol. 94, pp. 179–186, 2018.

[154] S. Salous, *Radio propagation measurement and channel modelling*. John Wiley & Sons, 2013.

[155] "Itu-r recommendation P.526 - propagation by diffraction," ITU-R, Geneva, Switzerland, 2011, rec. ITU-R P.526-12.

[156] A. of the Discussion Paper, "Discussion paper concerning recommendation ITU-R P.1812—the bullington diffraction model and its correction," ITU-R, United Kingdom of Great Britain and Northern Ireland, May 2008, discussion Paper.

[157] K. E. Stocker and F. M. Landstorfer, "Empirical prediction of radio wave propagation by neural network simulator," *Electronics Letters*, vol. 28, no. 8, pp. 724–726, Apr. 1992.

[158] A. Chakraborty, M. S. Rahman, H. Gupta, and S. R. Das, "Specsense: Crowdsensing for efficient querying of spectrum occupancy," in *Proc. of the IEEE INFOCOM*, Atlanta, Georgia, USA, 2017.

[159] E. Alimpertis, N. Fasarakis-Hilliard, and A. Bletsas, "Community rf sensing for source localization," *IEEE Wireless Communications Letters*, vol. 3, no. 4, pp. 393–396, 2014.

[160] L. Breiman, "Stacked regressions," *Springer Machine Learning*, vol. 24, no. 1, pp. 49–64, 1996.

[161] M. Molinari, M.-R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," in *Proc. of the ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big Internet Data (C2BID)*, London, UK, 2015, pp. 33–38.

[162] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[163] N. Singh, S. Choe, and R. Punmiya, "Machine learning based indoor localization using wi-fi rssi fingerprints: An overview," *IEEE Access*, vol. 9, pp. 127 150–127 174, 2021.

[164] S. Fudickar and M. Valentin, "Most accurate algorithms for rss-based wi-fi indoor localisation," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014, pp. 38–47.

[165] M. Merenda, L. Catarinucci, R. Colella, D. Iero, F. G. D. Corte, and R. Carotenuto, "Rfid-based indoor positioning using edge machine learning," *IEEE Journal of Radio Frequency Identification*, vol. 6, pp. 573–582, 2022.

[166] V. Bianchi, P. Ciampolini, and I. De Munari, "Rssi-based indoor localization and identification for zigbee wireless sensor networks in smart homes," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 2, pp. 566–575, 2019.

[167] H. S. Fahama, K. Ansari-Asl, Y. S. Kavian, and M. N. Soorki, "An experimental comparison of rssi-based indoor localization techniques using zigbee technology," *IEEE Access*, vol. 11, pp. 87 985–87 996, 2023.

[168] A. M. Ghonim, W. M. Salama, A. A. Khalaf, and H. M. Shalaby, "Indoor localization based on visible light communication and machine learning algorithms," *Opto-Electronics Review*, pp. e140 858–e140 858, 2022.

[169] L. Phillips, C. B. Porter, N. Kottege, M. D'Souza, and M. Ros, "Machine learning based acoustic sensing for indoor room localisation using mobile phones," in *2015 9th International Conference on Sensing Technology (ICST)*. IEEE, 2015, pp. 456–460.

[170] Y.-M. Lu, J.-P. Sheu, and Y.-C. Kuo, "Deep learning for ultra-wideband indoor positioning," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021, pp. 1260–1266.

[171] J. Lovon-Melgarejo, M. Castillo-Cara, O. Huarcaya-Canal, L. Orozco-Barbosa, and I. Garcia-Varea, "Comparative study of supervised learning and metaheuristic algorithms for the development of bluetooth-based indoor localization mechanisms," *IEEE Access*, vol. 7, pp. 26 123–26 135, 2019.

[172] A. Nessa, B. Adhikari, F. Hussain, and X. N. Fernando, "A survey of machine learning for indoor positioning," *IEEE Access*, vol. 8, pp. 214 945–214 965, 2020.

[173] T.-H. Chiang, Z.-H. Sun, H.-R. Shiu, K. C.-J. Lin, and Y.-C. Tseng, "Magnetic field-based localization in factories using neural network with robotic sampling," *IEEE Sensors Journal*, vol. 20, no. 21, pp. 13 110–13 118, 2020.

[174] Z. Liu, L. Chen, X. Zhou, Z. Jiao, G. Guo, and R. Chen, "Machine learning for time-of-arrival estimation with 5g signals in indoor positioning," *IEEE Internet of Things Journal*, 2023.

[175] I. Sharp and K. Yu, "Indoor toa error measurement, modeling, and analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2129–2144, 2014.

[176] S. M. Sheikh, H. M. Asif, K. Raahemifar, and F. Al-Turjman, "Time difference of arrival based indoor positioning system using visible light communication," *IEEE Access*, vol. 9, pp. 52 113–52 124, 2021.

[177] A. Alteneiji, U. Ahmad, K. Poon, N. Ali, and N. Almoosa, "Angle of arrival estimation in indoor environment using machine learning," in *2021 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 2021, pp. 1–6.

[178] R. Zhou, M. Hao, X. Lu, M. Tang, and Y. Fu, "Device-free localization based on csi fingerprints and deep neural networks," 06 2018, pp. 1–9.

[179] C.-H. Hsieh, J.-Y. Chen, and B.-H. Nien, "Deep learning-based indoor localization using received signal strength and channel state information," *IEEE access*, vol. 7, pp. 33 256–33 267, 2019.

[180] Y. Yong, C. Tan, I. Tan, and S.-W. Tan, "Robust fingerprint construction based on multiple path loss model (m-plm) for indoor localization," *Computers, Materials and Continua*, vol. 74, no. 1, pp. 1801–1818, Sep. 2022, funding Information: Funding Statement: This research was funded by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS) with grant number FRGS/1/2019/ICT02/MMU/02/1. Publisher Copyright: © 2023 Tech Science Press. All rights reserved.

[181] H.-H. Liu and Y.-N. Yang, "Wifi-based indoor positioning for multi-floor environment," in *TENCON 2011-2011 IEEE Region 10 Conference*. IEEE, 2011, pp. 597–601.

[182] I. Alshami, N. Ahmad, and S. Sahibuddin, "Automatic wlan fingerprint radio map generation for accurate indoor positioning based on signal path loss model," pp. 17 930–17 936, 01 2015.

[183] X. Du, X. Liao, M. Liu, and Z. Gao, "Crcloc: A crowdsourcing-based radio map construction method for wifi fingerprinting localization," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 12 364–12 377, 2022.

[184] F. Meneses, A. Moreira, A. Costa, and M. J. Nicolau, "Radio maps for fingerprinting in indoor positioning," in *Intelligent Data-Centric Systems: Geographical and Fingerprinting Data to Create Systems for Indoor Positioning and Indoor/Outdoor Navigation*, J. Conesa, A. Pérez-Navarro, J. Torres-Sospedra, and R. Montoliu, Eds. Academic Press, 2019, pp. 69–95. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128131893000046

[185] K. Shi, Z. Ma, R. Zhang, W. Hu, and H. Chen, "Support vector regression based indoor location in ieee 802.11 environments," *Mobile Information Systems*, vol. 2015, 09 2015.

[186] L. Chanama and O. Wongwirat, "A comparison of decision tree based techniques for indoor positioning system," in *2018 International Conference on Information Networking (ICOIN)*, 2018, pp. 732–737.

[187] M. Luckner and R. Górak, "Modified random forest algorithm for wi–fi indoor localization system," 09 2016.

[188] B. Ezhumalai, M. Song, and K. Park, “An efficient indoor positioning method based on wi-fi rss fingerprint and classification algorithm,” *Sensors*, vol. 21, no. 10, p. 3418, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/10/3418

[189] J. Xu, H. Dai, and W.-h. Ying, “Multi-layer neural network for received signal strength-based indoor localisation,” *IET Communications*, vol. 10, 03 2016.

[190] N. Patwari, O. H. I. Alfred, M. Perkins, and et al., “Relative location estimation in wireless sensor networks,” *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2137–2148, 2003.

[191] M. S. Rahman, Y. Park, and K. D. Kim, “Rss-based indoor localization algorithm for wireless sensor network using generalized regression neural network,” *Arabian Journal for Science and Engineering*, vol. 37, no. 4, pp. 1043–1053, 2012.

[192] K. Kamol and K. Prashant, “Modeling of indoor positioning systems based on location fingerprinting,” in *Proceedings of IEEE INFOCOM*, vol. 2, 2005, pp. 1012–1022.

[193] X. Wang, X. Wang, S. Mao, J. Zhang, S. C. G. Periaswamy, and J. Patton, “Indoor radio map construction and localization with deep gaussian processes,” *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 238–11 249, 2020.

[194] J. Bi, Y. Wang, Z. li, S. Xu, J. Zhou, M. Sun, and M. Si, “Fast radio map construction by using adaptive path loss model interpolation in large-scale building,” *Sensors*, vol. 19, p. 712, 02 2019.

[195] S. Bi, J. Lyu, Z. Ding, and R. Zhang, “Engineering radio maps for wireless resource management,” *IEEE Wireless Communications*, vol. 26, no. 2, pp. 133–141, 2019.

[196] Y. Ye and B. Wang, “Rmapcs: Radio map construction from crowdsourced samples for indoor localization,” *IEEE Access*, vol. 6, pp. 24 224–24 238, 2018.

[197] H. Wymeersch, J. He, B. Denis, A. Clemente, and M. Juntti, “Radio localization and mapping with reconfigurable intelligent surfaces: Challenges, opportunities, and research directions,” *IEEE Vehicular Technology Magazine*, vol. 15, no. 4, pp. 52–61, 2020.

[198] J. Tuta and M. B. Juric, “A self-adaptive model-based wi-fi indoor localization method,” *Sensors*, vol. 16, no. 12, p. 2074, 2016.

[199] S. Sorour, Y. Lostanlen, S. Valaee, and K. Majeed, “Joint indoor localization and radio map construction with limited deployment load,” *IEEE Transactions on Mobile Computing*, vol. 14, no. 5, pp. 1031–1043, 2015.

[200] M. Milijić, Z. Stanković, and I. Milovanović, “Hybrid-empirical neural model for indoor/outdoor path loss calculation,” in *Proceedings of the 2011 10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS)*, Nis, Serbia, October 2011, pp. 548–551. [Online]. Available: https://scholar.google.com/scholar?cluster=12345678901234567890

[201] I. Popescu, D. Nikitopoulos, and P. Constantinou, “Comparison of ann based models for path loss prediction in indoor environment,” in *Proceedings of the IEEE Vehicular Technology Conference*, Montreal, QC, Canada, September 2006, pp. 1–5. [Online]. Available: https://scholar.google.com/scholar?cluster=09876543210987654321