Pontifícia Universidade Católica
do Rio de Janeiro

Yoiz Eleduvith Nunez Ruiz

On Machine Learning Techniques Toward Path Loss Modeling in 5G and Beyond Wireless Systems

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica, do Departamento de Engenharia Elétrica da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica.

Advisor : Prof. Luiz Alencar Reis da Silva Mello
Co-advisor: Prof. Lisandro Lovisolo

Rio de Janeiro
August 2023

Pontifícia Universidade Católica
do Rio de Janeiro

Yoiz Eleduvith Nunez Ruiz

On Machine Learning Techniques Toward Path Loss Modeling in 5G and Beyond Wireless Systems

Thesis presented to the Programa de Pós–graduação em Engenharia Elétrica da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Engenharia Elétrica. Approved by the Examination Committee:

Prof. Luiz Alencar Reis da Silva Mello
Advisor
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Lisandro Lovisolo
Departamento de Eletrônica e Telecomunicações – UERJ

Prof. Marley Maria Vellasco
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Natalia Castro Fernandes
Departamento de Engenharia de Telecomunicações – UFF

Prof. José Cândido Silveira Santos Filho
Departamento de Comunicações – UNICAMP

Prof. Marcio Eisencraft
Departamento de Engenharia de Telecomunicações e Controle – EPUSP

Prof. Marco Antonio Grivet Mattoso
Departamento de Engenharia Elétrica – PUC-Rio

Rio de Janeiro, August the 29th, 2023

Yoiz Eleduvith Nunez Ruiz

Received the B.S. degree in electronic engineering from the Technological University of Honduras, in 2012, the M.S. degree in project management from the Central American Technology University, Honduras, in 2016, and the M.S. degree in electrical engineering from the Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, in 2019. Worked for Claro, Honduras for five years, from 2012 to 2017, at the network operation center and network quality engineering department. Her research interests include interpretability of machine learning techniques and radio propagation channel modeling for next-generation wireless systems.

Bibliographic data

Nunez Ruiz,Yoiz Eleduvith

On Machine Learning Techniques Toward Path Loss Modeling in 5G and Beyond Wireless Systems / Yoiz Eleduvith Nunez Ruiz; advisor: Luiz Alencar Reis da Silva Mello; co-advisor: Lisandro Lovisolo. – 2023.

161 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2023.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Perda de percurso. 3. Aprendizado de máquinas. 4. Interpretabilidade. 5. Ondas milimétricas. 6. Sub-6 GHz. 7. Comunicação vehicular. I. da Silva Mello, Luiz Alencar Reis. II. Lovisolo, Lisandro . III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

To my parents, for their support
and encouragement.

Acknowledgments

Foremost, I convey a special thanks to almighty God. He has given me strength and encouragement to complete this journey. I am truly grateful for His mercy and grace.

I am remarkably grateful to my advisor, Professor Luiz Alencar da Silva Mello, for his unconditional support and guidance. His practical advice and experience gave me confidence in challenging moments.

I would like to express my sincere gratitude to my co-advisor, Professor Lisandro Lovisolo, for his insightful suggestions, enthusiasm, and always being there when I needed his support. This work would not have been possible without his guidance.

I would like to acknowledge and thank my family, my parents, my brother, and his wife. They always encouraged me and believed in me.

I would like to thank my colleagues at PUC-Rio and CETUC, especially Johnes Ricardo, for their sincere friendship and assistance. A special appreciation to Prof. Guilherme Temporão, Marcelo Balisteri, and Marta Pudwell for their valuable support in the final stage of this journey.

I would like to thank my colleagues from the Radio-Propagation Laboratory, CETUC, Carlos Orihuela, Marcelo Molina, Carlos Rodriguez, and Professor Glaucio Lopes, for helping me process the datasets from the measurement campaigns used in this research and always being there when I needed their support.

A special appreciation and dedication to my husband Jorge for all the encouragement, love, patience, source of inspiration, and technical support he gave me during this journey. For always looking after my well-being. Thank you from the bottom of my heart!

This work was supported in part by the Brazilian Agency CNPq under Grant 142449/2019-9 and in part by the Brazilian Agency FAPERJ under Grant 204.143/2022.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Abstract

Nunez Ruiz,Yoiz Eleduvith; da Silva Mello, Luiz Alencar Reis (Advisor); Lovisolo, Lisandro (Co-Advisor). **On Machine Learning Techniques Toward Path Loss Modeling in 5G and Beyond Wireless Systems**. Rio de Janeiro, 2023. 161p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Path loss (PL) is an essential parameter in propagation models and critical in determining mobile systems' coverage area. Machine learning (ML) methods have become promising tools for radio propagation prediction. However, there are still some challenges for its full deployment, concerning to selection of the most significant model's inputs, understanding their contributions to the model's predictions, and a further evaluation of the generalization capacity for unknown samples. This thesis aims to design optimized ML-based PL models for different applications for the 5G and beyond technologies. These applications encompass millimeter wave (mmWave) links for indoor and outdoor environments in the frequency band from 26.5 to 40 GHz, macrocell coverage in the sub-6 GHz spectrum, and vehicular communications using measurements campaign carried out by the Laboratory of Radio-propagation, CETUC, in Rio de Janeiro, Brazil. Several ML algorithms are exploited, such as artificial neural network (ANN), support vector regression (SVR), random forest (RF), and gradient tree boosting (GTB). Furthermore, we have extended two empirical models for mmWave with improved PL prediction. We proposes a methodology for robust ML model selection and a methodology to select the most suitable predictors for the machines considered based on performance improvement and the model's interpretability. In adittion, for the vehicle-to-vehicle (V2V) channel, a convolutional neural network (CNN) technique is also proposed using a transfer learning approach to deal with small datasets. The generalization tests proposed shows the ability of the ML models to learn the pattern between the model's inputs and PL, even in more challenging environments and scenarios of unknown samples.

Keywords

Path loss; Machine learning; Interpretability; Millimeter waves; Sub-6 GHz; Vehicular communication.

Resumo

Nunez Ruiz,Yoiz Eleduvith; da Silva Mello, Luiz Alencar Reis; Lovisolo, Lisandro . **Sobre técnicas de aprendizado de máquina em direção à modelagem de perda de propagação em sistemas sem fio 5G e além**. Rio de Janeiro, 2023. 161p. Tese de Doutorado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A perda de percurso (PL) é um parâmetro essencial em modelos de propagação e crucial na determinação da área de cobertura de sistemas móveis. Os métodos de aprendizado de máquina (ML) tornaram-se ferramentas promissoras para a previsão de propagação de rádio. No entanto, ainda existem alguns desafios para sua implantação completa, relacionados à seleção das entradas mais significativas do modelo, à compreensão de suas contribuições para as previsões do modelo e à avaliação adicional da capacidade de generalização para amostras desconhecidas. Esta tese tem como objetivo projetar modelos de PL baseados em ML otimizados para diferentes aplicações das tecnologias 5G e além. Essas aplicações abrangem links de ondas milimétricas (mmWave) para ambientes indoor e outdoor na faixa de frequência de 26,5 a 40 GHz, cobertura de macrocélulas no espectro sub-6 GHz e comunicações veiculares usando campanhas de medições desenvolvidas em CETUC, Rio de Janeiro, Brazil. Vários algoritmos de ML são explorados, como redes neurais artificiais (ANN), regressão de vetor de suporte (SVR), floresta aleatória (RF) e aumento de árvore de gradiente (GTB). Além disso, estendemos dois modelos empíricos para mmWave com previsão de PL melhorada. Propomos uma metodologia para seleção robusta de modelos de ML e uma metodologia para selecionar os preditores mais adequados para as máquinas consideradas com base na melhoria de desempenho e na interpretabilidade do modelo. Além disso, para o canal veículo-veículo (V2V), uma técnica de rede neural convolucional (CNN) também é proposta usando uma abordagem de aprendizado por transferência para lidar com conjuntos de dados pequenos. Os testes de generalização propostos mostram a capacidade dos modelos de ML de aprender o padrão entre as entradas do modelo e a PL, mesmo em ambientes e cenários mais desafiadores de amostras desconhecidas.

Palavras-chave

Perda de percurso; Aprendizado de máquinas; Interpretabilidade; Ondas milimétricas; Sub-6 GHz; Comunicação vehicular.

Table of contents

List of figures

List of tables

List of Abreviations

5G – Fifth generation mobile network

PL – Path loss

ML – Machine learning

mmWave – Millimeter wave

V2I – Vehicle-to-Infraestructure

V2V – Vehicle-to-Vehicle

ANN – Artificial neural network

SVR – Support vector regression

RF – Random forest

GTB – Gradient tree boosting

CNN – Convolutional neural network

ABG – Alpha-beta-gamma

CIF – Close-in frequency-dependent exponent

MI – Mutual information

IML – Interpretable machine learning

RMSE – Root mean squared error

MAPE – Mean absolute percentage error

CW – Continuous wave

OFDM – Orthogonal frequency-division multiplexing

Man cannot discover new oceans unless he has the courage to lose sight of the shore.

Andre Gide, .

1
Introduction

1.1
Motivation

The next generation of mobile communication systems relies on millimeter wave technology to fulfill high data rates and low latency requirements [1, 2]. However, deploying the fifth generation (5G) mobile network and beyond systems will involve the integration of heterogeneous networks with overlapping cells to meet capacity and higher data rate transmission [3, 4], as illustrated in Figure 1.1. The 5G network system is expected to support a variety of scenarios from cellular mobile broadband to vehicular communication, as well as the integration with existing technologies [1]. While the 4G cellular network enhanced mobile broadband capabilities, the advent of 5G needs to meet new and increased performance requirements; however, the resulting scenarios become more complex and dynamic than for previous technologies, with a higher demand for greater capacity and coverage, smaller latency, more flexibility, and increased efficiency exploiting the resources and meeting the operational requirements [5].

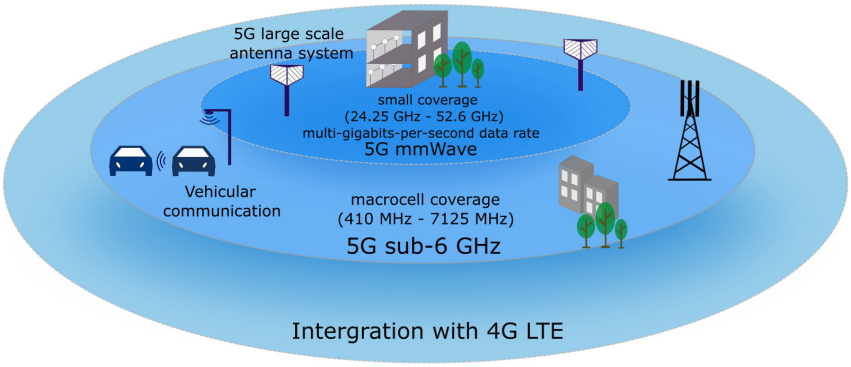

Figure 1.1: 5G research challenges [1].

The expected frequency bands of 5G are defined in Release 15-16 of the 3rd Generation Partnership Project (3GPP) [6], divided into the frequency range 1 (FR1) considering the sub-6 GHz spectrum from 410 MHz to 7125 MHz and the frequency range 2 (FR2) covering the frequency ranges from 24.25 GHz to 52.6 GHz. Employing the sub-6 GHz spectrum for macrocell coverage

is expected to increase the performance and capacity of 5G networks [7]. Meanwhile, the millimeter wave band is expected to enable high data rate communications due to the larger bandwidth available [8, 9]. Additionally, vehicular communication is a key technology for intelligent transportation systems (ITS) due to the need for safer, more efficient, and sustainable transportation [10]. ITS applications consider the exchange of data in different vehicular communication among vehicles and with road infrastructure referred to as vehicle-to-vehicle (V2V) [11, 12], vehicle-to-infrastructure (V2I) [13, 14, 15], and vehicle-to-everything (V2X) [16] to provide secure and reliable wireless communications between vehicles to road infrastructure, and among vehicles, respectively, expected to operate in the frequency band below 6 GHz [17, 18], and in millimeter wave [11], oriented to short-range communications [19].

The millimeter wave (mmWave) spectrum is identified as one of the key technologies to provide multi-gigabit-per-second (Gbps) data rates [8]. Nevertheless, due to the smaller wavelength, the propagating wave suffers from atmospheric absorption, higher attenuation loss, and blockage from obstacles [3]. In addition, the employment of array-antenna systems and beamforming offer new possibilities at mmWave [1, 8, 20]. However, the frequency band between 28 GHz and 38 GHz (26.5 GHz to 40 GHz) presents a negligible atmospheric absorption in short-range links (less than 200 m) suitable for communication within 5G small cells for indoor and outdoor coverage [20].

Currently, the 5G technology is being deployed across the world. Globally, 5G reached a coverage of 35 percent of the world population at the end of 2022, and only about 25 percent of the 4G sites have been upgraded to 5G as reported in the 2023 Ericsson Mobility Report [21]. One of the most important aspects of planning and optimizing these wireless communication systems is the development of accurate radio propagation models for the different operational frequency bands, and for indoor and outdoor environments [22, 23].

1.2 Overview of Path Loss Modeling

One important indicator for characterizing the channel of the communication system is path loss (PL). The PL is the decrease in the signal strength during propagation from the transmitter to the receiver [24]. PL plays a crucial role in estimating link budgets for coverage planning when designing wireless networks. The mmWave spectrum exhibits very different propagation conditions when compared to the microwave bands (below 6 GHz) [25]. At these frequencies, not only the propagation loss is higher but it is also highly dependent on atmospheric conditions. The result in a wider range of variation of the

propagation loss and existing channel models designed for microwave bands are inadequate at mmWave frequencies and the development of new appropriate PL models becomes necessary [25, 26].

In vehicular communication, the V2I and V2V channels differ and deviate significantly from those in cellular communications [10]. For the V2I, both the transmitter and receiver have low antenna heights, and additionally, the receiver or the transmitter can be moving [14, 17]. For the V2V case, the channel is more dynamic due to the relative movement between the vehicles' transmitter and receiver, rapidly changing the surrounding environment [18, 27]. Therefore, other objects, such as moving and stationary vehicles, buildings, and vegetation, affect the radio propagation environment [15, 28].

Traditionally, PL models have been proposed and employed for specific environments and frequencies, given the distinctive propagation characteristics in each scenario. Over the years, deterministic and empirical models have been used to predict the PL at different frequencies and distances with various degrees of complexity (number of parameters and algorithm complexity in the model) [29, 30, 31]. Deterministic models, such as ray-tracing techniques, are site-specific and accurate. However, in challenging environments, their complexity and computational requirements increase significantly [3]. They become particularly demanding in the mmWave band since detailed characterizations of foliage, building geometry and electric characteristics that influence reflections and refraction phenomena, are required for calculations [32]. Furthermore, these techniques become impractical for macrocell coverage due to the extensive area and the high number of multipath components [33].

On the other hand, in an empirical model, coefficients are adjusted using measurement data. These models have low computational complexity and require less geometric information. However, they may be simplistic and provide low accuracy leading to inherent prediction errors. The alpha-beta-gamma (ABG) [34] and the close-in frequency-dependent exponent (CIF) [22] are multi-frequency PL prediction models covering a wide range of frequencies, including the microwave and mmWave spectrum [35, 36]. For macrocell coverage, several empirical models have been proposed, such as Egli model [37], Okumura-Hata [38] and Lee model [39]. For the characterization of the vehicular channel, empirical models such as ITU-R P.1411-5, two-ray, log-distance, and street canyon have been proposed [10, 14, 15, 17, 19].

More recently, with the motivation of reducing computational demands while maintaining good accuracy, machine learning methods have emerged as promising tools for radio propagation prediction in the different 5G application scenarios and communication environments [40].

1.3 Machine Learning for Path Loss Prediction

In the last years, artificial intelligence (AI) and, particularly, machine learning (ML) algorithms have proven to be feasible solutions for a variety of problems in the wireless communication domain, addressing challenges in network planning and design, management of resources, optimization, control and operation of the network, and managed customer service system [5].

ML aims at designing systems that learn and evolve from experience/data automatically [41]. It encompasses approaches for providing approximate mappings between inputs (predictors) and output (response). ML algorithms are classified among supervised, unsupervised, and reinforcement learning [42, 43, 44]. Supervised learning refers to designing systems that predict or estimate a response using at least one predictor and examples of input and known output mappings, divided into regression and classification tasks. Unsupervised learning encompasses techniques to discover patterns and associations within the data without known responses, that is, without supervision. Reinforcement learning techniques employ exploration (of the unknown) and exploitation (of the known) to maximize the reward for dynamic scenarios [43, 44].

PL modeling based on machine learning may be tackled as a regression problem that, in principle, any supervised algorithm can solve if enough data is available. The input features for the PL prediction can be extracted from the system- and environment-dependent parameters created on tabular datasets, 2D/3D digital maps, and topographic databases [45]. For PL prediction based on tabular data, some studies have demonstrated that the prediction performance using ML models based on artificial neural networks (ANN), support vector regressor (SVR) and ensembles of trees perform reasonably well [45, 46, 47] surpassing those based on k-nearest neighbors (k-NN) [48, 49] or one single decision tree [49]. Some methods using an ensemble of trees applied for PL prediction are random forest (RF), and AdaBoost [50, 51].

On the other hand, deep learning techniques for PL prediction extract features from the environment between the transmitter and receiver using digital surface/elevation models [52], and satellite images [53] for outdoor environments, and from local area multi-scanning (LAMS) images for indoor cases [54]. In addition, deep learning models for PL prediction, in general, require a vast number of samples obtained from extensive measurement campaigns or ray-tracing simulations [52, 55]. The studies demonstrate that ML-based models improve over the traditional PL models with a trade-off between computational load and efficiency, regardless of frequency and radio propagation environment. As a result, ML-based models have emerged as a

promising alternative to model the path loss. However, many ML algorithms have been regarded as black-box systems because of the lack of knowledge of the machine's internal mechanism to output the prediction after training. Therefore, despite the several studies for PL modeling based on ML techniques, it is still unclear why the ML models improve over the traditional PL models, how to interpret them, and how well the trained model can generalize to unseen data [32, 40].

Consequently, there are some open challenges:

- The selection of the most relevant predictors to obtain a high-quality model for the radio propagation prediction problem;
- The explanation/interpretability of the ML-based PL model response, and
- Further exploitation of the trained PL model's applicability and generalization capacity for unknown samples, i.e., scenarios.

Regarding predictor selection techniques in supervised algorithms, filter, and wrapper methods are usually employed to forecast the most relevant predictors for good performance and to understand the model's response [56]. Filter methods assess the importance of the predictor based on the dataset, independently of the applied ML algorithm and its performance [57], for example, using mutual information (MI) [58]. The MI is calculated on the dataset to provide pre-interpretations, i.e., before training, aiming to measure the correlation between predictors and the output. Pre-interpretability techniques remain independent of the model since they are solely applicable to the data itself [59]. However, they can explore and provide an understanding of the data before model selection.

On the other hand, wrapper methods assess the effect of the predictor by evaluating the performance of a specific ML algorithm using sequential forward or backward greedy approaches [58]. In the forward approach, the predictor leading to the largest increase in the model's performance at each iteration is included in the subset of already selected predictors. In the backward approach, the predictor that does not significantly reduce the model's performance is removed; the process starts using all the available predictors [60, 61]. In addition, in [62], the term coalition is incorporated to refer to the union of predictors used in the input; thus, the coalition incorporates a new predictor that evaluates the contribution, ranking the importance of each predictor by its Shapely value [62].

When considering ML-based PL models, some schemes of predictor selection incorporate principal component analysis (PCA) techniques to reduce the

dimension of the dataset [63, 64]. Nevertheless, the obtained principal components (data that explain a maximal amount of variance) are less interpretable, and the PCA technique may have a low trade-off between information loss according to the number of principal components and dimension reduction [65].

Nonetheless, some of the limitations witnessed in supervised learning, particularly for wireless communications, are the lack of the capacity to interpret the model's response [66]. This lack of interpretability leads to uncertainty regarding the model's inner workings and decision-making processes [67]. An interpretability definition for ML models is introduced in [68] to refer to the degree to which a human can understand the cause of a given decision of the model. This is related to the cause-and-effect relationship between the model's inputs and outputs. Another related term with interpretability in ML models is explainability. It is associated with explaining the internal logic and mechanisms inside the machine [67].

As a result, to gain insight into the connections between the ML model's input and output, in the last years, significant contributions have emerged to advance the interpretability of ML models. These tools provide post-hoc interpretation, analyzing the models' responses after training [69]. Methods assessing interpretability [67] are based on model-specific or model-agnostic techniques that can be applied to a single model or group of models, or any model, respectively; assessing and interpreting a single prediction instance (local interpretation) or the whole model (global interpretation).

Therefore, within the context of radio propagation modeling, diverse environments result in different propagation characteristics, which influence the input-output relationship of the channel. Since these relationships vary across the different scenarios, an appropriate selection of the input features that maximize the model's performance and the interpretability of its response is required to fully deploy these radio propagation models [49, 66]. In addition, the conduction of generalization tests is necessary to provide a reliable conclusion about an ML model's generalization capacity. The term generalization capacity denotes the ability of the model to accurately predict the output value when the instance is unknown during the training of the model [40, 70].

1.4 Major Research Contributions

This thesis addresses the previously mentioned issues in ML-based PL prediction models considering different frequency bands and environments, which include mmWave frequencies for indoor and short-range outdoor scenarios and sub-6 GHz bands for urban macrocell coverage, V2I, and V2V

links. The measurement campaigns for all those environments were carried out by the Laboratory of Radio-propagation, CETUC, in Rio de Janeiro, Brazil. Accordingly, the following topics of study correspond to our contributions:

- A large-scale review of radio propagation prediction based on ML models regarding operating frequency, algorithms employed, input features, and performance indicators is presented.
- Considering the precedent works, different supervised algorithms are exploited to build PL models for several frequency bands and environments. A robust methodology for designing the ML models is employed across the studies to offer a reliable performance comparison.
- A PL model based on deep learning technique and transfer learning is proposed and studied for the V2V environment.
- Besides the ML algorithms, empirical models are also considered for each corresponding scenario leading to fair and sound comparisons. For mmWave in indoor and outdoor environments, extensions of the multi-frequency empirical models ABG and CIF, including additional environmental feature inputs, are proposed.
- We propose a predictor selection methodology to form coalitions (a significant subset of predictors) for the ML-based PL models. It uses a forward selection approach to rank the predictors according to their performance, along with interpretable machine learning (IML) techniques, to provide understandable insights into the connections between the model's inputs (predictors) and the PL (output).
- We evaluate the generalization capability of the ML-based PL models using the conventional random split dataset and considering unknown transmission/receiver positions and streets, whatever fits better the application scenario for the model.

We evaluate the proposed ML models and the methodology for predictor selection in different environments through the dataset obtained from the diverse measurement campaigns. The codes developed, including the results for Figures and Tables, are available on GitHub: https://github.com/YoizNunez/ML-based-PL-models

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 introduces the theoretical basis of the radio propagation channel, a survey on PL modeling using ML techniques, principles of supervised learning for regression, and the fundamentals of the proposed interpretability techniques. Chapter 3 considers PL models for mmWave in the indoor environment; it also presents in detail the methodology for PL model selection employed for ML techniques and empirical models. Furthermore, in this chapter, the proposed methodology to choose the predictor coalitions is explained and applied to the proposed ML models. The results are discussed and analyzed using the pre and post-hoc interpretation techniques. In addition, a generalization test is addressed by considering unknown transmitters and receivers. To reduce the contents, some results on model tuning and intrepretability are presented in Appendix A and Appendix B, respectively.

The following chapters address the proposed methodologies and post-hoc interpretability tools for ML-based PL models in different environments; the interpretability model's response analysis focuses on the ML model with the best PL prediction. Chapter 4 tackles the issue of PL modeling for the mmWave Outdoor communications, and a generalization test considering unknown receiver positions is tailored to assess the applicability of the model in real scenarios. Chapter 5 address the PL modeling for urban macrocell coverage. Generalization tests regarding unknown streets are presented. In continuation, Chapter 6 addresses PL models for vehicular communication. It is divided into two sections, addressing V2I and V2V links. For the V2V scenario, a deep learning approach is proposed. Conclusions and suggestions for future work are presented in Chapter 7.

2
Fundamental Concepts and Background

This chapter introduces the fundamental concepts and parameters used to characterize a radio- propagation channel. Furthermore, it presents a survey on path loss modeling using machine learning to identify the techniques and methodologies that have been employed in recent years. The principles of supervised machine learning employed in this thesis are also described, aiming at explaining the mechanisms behind the techniques employed. Lastly, It presents the fundamentals for the pre and post-hoc interpretability techniques employed in this thesis.

2.1
Characteristics of the Mobile Radio Channel

The mobile radio channel is the physical transmission medium between the transmitting and the receiving antennas [71, 72]. The transmitted signal reaches the receiver by means of different propagation mechanisms, such as free space propagation, reflection, diffraction, and scattering; they arise due to the different obstacles existing between the transmitter and the receiver, as illustrated in Figure 2.1 [30, 72, 73]. These propagation mechanisms are briefly described below.

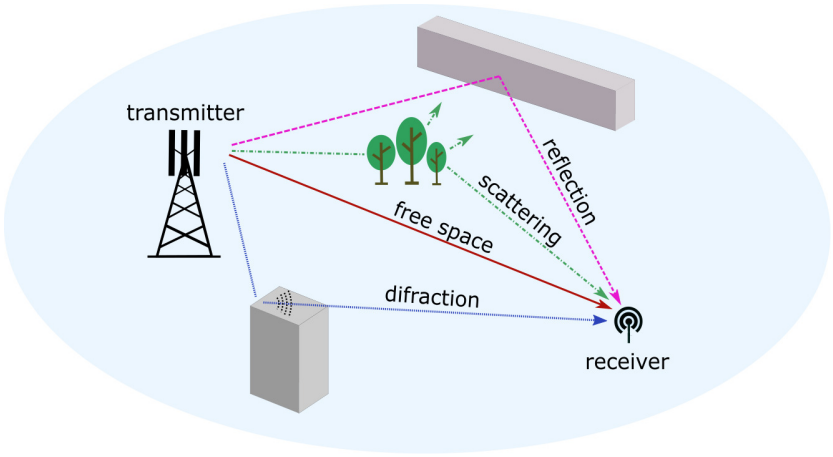

Figure 2.1: Multipath propagation effects.

– Free space propagation refers to the scenario where there is a direct, clear and unobstructed line-of-sight (LOS) path between the transmitter

and the receiver. The attenuation is simply due to the spread of the transmitted energy in space, in directions other than that of the receiver.

- Reflection occurs when the transmitted wave encounters an object with dimensions larger than the wavelength of the propagating wave, that reflect the wave in the direction of the receiver antenna.
- Scattering arises when the medium through which the wave propagates contains objects with dimensions smaller than the wavelength, such that multiple attenuated wave fronts are directed at the receiver.
- Diffraction appears when the radio path between the transmitter and receiver is obstructed by large surfaces, around which the wave "turns" to reach the receiver.

The received signal may present strength variations with the distance in large- and small-scales due to the path loss, and fading due to different effects, among which are the ones arising by multipath [30, 73].

2.1.1 Path Loss

Path loss (PL) is the average attenuation of the received power due to wave propagation between transmitter and receiver [30, 72]. This average attenuation can usually be approximately expressed as a function of the logarithm of the distance between transmitter and receiver [72], given by

$$\overline{PL}[dB] = \overline{PL}(d_o) + 10n\log_{10}\left(\frac{d}{d_o}\right), \tag{2-1}$$

where $\overline{PL}(d_o)$ is the mean path loss at the reference distance d_o, n is the path loss exponent, and d is the distance between the transmitter and receiver.

2.1.2 Fading

Fading consists of large- and small-scale fluctuations of the attenuation experimented by the propagating wave from the transmitter to the receiver. Large-scale fading (shadowing) represents the average attenuation of the receiver power due to the presence of large obstructions (buildings, foliage, walls, and furniture) [73]. On the other hand, small-scale fading refers to the rapid changes in the amplitude and phase of the propagating wave over a short period due to wave reflections on obstacles and to the receiver's motion [73]. The number and type of objects that cause large-scale fading at any receiver location are typically unknown. Hence, the attenuation due to shadowing is

modeled statistically by a log-normal distribution [72], and the PL model becomes

$$PL[dB] = \overline{PL}[dB] + X_\sigma, \tag{2-2}$$

where X_σ is a zero-mean Gaussian random variable with standard deviation σ in dB, which reflects the variation of the receiver power due to shadowing.

2.1.3 Multipath Propagation

Due to the multiple reflecting obstacles that the transmitted signal encounters when propagating, different multipath components may arrive at the receiving antenna with different delays, phase shifts, and attenuations [30, 72]. As a consequence, at the receiver, the waves from different paths can mix constructively or destructively. Thus, multipath propagation may produce severe dispersion. The expected degree of dispersion is determined by measuring the power delay profile (PDP) using wideband channel sounding techniques [74]. The PDP represents the temporal distribution of the power $P(\tau)$.

The effect of multipath propagation is commonly measured in terms of the extent of time dispersion introduced by the multipath channel, known as delay spread [74]. Delay spread can be quantified through different indicators; the most common are the root-mean-squared delay spread (RMS DS) and the mean excess delay, which can be calculated from the PDP [73]. The RMS DS measures the extent of the delay spread around the mean [75], and mathematically, it is the square root of the second central moment of the power delay profile, given by

$$\tau_{rms} = \sqrt{\frac{\sum_k (\tau_k - \overline{\tau})^2 P(\tau_k)}{\sum_k P(\tau_k)}}. \tag{2-3}$$

where $P(\tau_k)$ and τ_k are the power and excess delay of the k^{th} ray, respectively. In turn, it has been found that RMS DS is directly related to the minimum symbol length that can be used in order to avoid excessive intersymbol interference. In general, the RMS DS is modeled statistically using its probability density function (PDF) [76]. Distributions such as Lognormal, Nakagami, Weibull, Rayleigh, and Rice are usually employed in the statistical analysis of the multipath channel due to their good fit with the one constructed from empirical data [77]. On the other hand, the mean excess delay describes the mean propagation delay of the components concerning the first component to arrive at the receiver; it is the first moment of the PDP, that is, [73]

$$\overline{\tau} = \frac{\sum_k P(\tau_k)\tau_k}{\sum_k P(\tau_k)}. \tag{2-4}$$

The mentioned channel parameters can provide a basis for predicting path loss and transmitter coverage to optimize the network performance [48]. This thesis focuses on PL prediction by exploiting different machine learning techniques and empirical models. The following section presents an overview of machine learning studies to deal with PL prediction, identifying the techniques and methodologies employed.

2.2 Machine Learning for Path Loss Prediction

In recent years, several works have attempted to use machine learning (ML) for PL prediction in different radio-propagation environments, mainly in urban [47, 78, 79], and suburban areas [63]. Others have focused on vehicle-to-vehicle (V2V) communications [16, 70], prediction of the received power from unmanned aerial vehicles (UAV) [48, 51], aircraft communications [50], and also smart campus environments [80]. The dataset to design these models has been obtained by conducting measurement campaigns or trough simulation methods.

The ML algorithms employed for PL prediction encompass ANN [49], SVR [47], and RF [79], among others. Other works compare the PL prediction performance among different ML algorithms [45, 50, 51, 81], including deep neural network (DNN), decision tree, and k-nearest neighbors (k-NN) [49]. Data transferring is suggested in [45, 48] for increasing the training dataset by fusing data acquired at different frequencies and scenarios.

Table 2.1 presents some relevant aspects of works delving into PL modeling using ML. Its columns present the channel frequency and propagation environment, the ML tools employed together with the learning criterion (the so-called loss function – LF), the input features (the predictors employed) for the models, and the indices for performance evaluation. They encompass different frequencies in urban, suburban, rural, and indoor environments. It shall be noticed that the listed works use different ML models, loss functions, and predictors subsets.

Considering indoor environments, [46] and [82] propose ANN-based PL models. Meanwhile, [82] considers models when the wave must propagate through multiple walls (multi-wall) and in multiple-frequency (multi-frequency) scenarios. Body shadowing and furniture effects in an indoor area were suggested in [83] for modeling the average received power for multi-frequency models using neural networks. In [84], the authors combine ANN and a 3-D ray launching algorithm to predict the received power in an indoor scenario. Other applications employ ML techniques such as recurrent neural

network RNN and convolutional neural network (CNN) to model the environmental effect in the received power for positioning and tracking the mobile terminal in the indoor environment [85].

Table 2.1 lists some works dealing with PL prediction using ML techniques in the ultra high frequency (UHF, 300-3000 MHz). They suggest different features as input for radio-propagation channel modeling. Some works [63, 86] indicate that predictor selection techniques to reduce the dataset dimension may be worthy for PL modeling both in urban [86] and suburban [63] environments. In terms of features selection, in [79] the authors select the input features from various terrain and environmental parameters. In [87], image technologies – light detection and ranging (LiDAR) and 2D satellite images, provide volumetric data, including tree canopies and vegetation, that feed an ANN for PL prediction. Independently of the different aspects, besides model training, finding the optimal model hyperparameters is also important, as in [46] and [83], for example. The hyperparameters define the architecture of the ML model better described in Section 2.3.

Table 2.1: ML-based studies on PL prediction for urban, suburban, rural, and indoor environments.

Reference	**Freq. [MHz] | Scenario**	**ML alg. | Loss func.**	**Input features**	**Performance indicators**
[47]	853.71 | Urban	SVR | ε-insensitive loss	Transmitter-receiver distance, terrain elevation, horizontal and vertical angle, latitude, longitude, horizontal and vertical attenuation of the antenna.	Mean error and σ.
[55]	900 | Urban	CNN, ANN | N/D	Image of building height information.	MAE, MAPE, MSE, RMSE, and R^2.
[86]	947 | Urban	ANN, SVR | MSE, ε-insensitive loss	Inputs features related to the global path: distance and portion through the buildings.	MSE.
[78]	947.53 | Urban	ANN | MSE	Latitude, longitude, elevation and transmitter-receiver distance.	ME, MAE, MSE, RMSE, σ and R.
[51]	1800 | Urban	ANN, SVR, RF, AdaBoost | N/D	Transmitter-receiver distance, frequency, and relative coordinates of the receiver position.	MAE, MAPE, RMSE and R.
[49]	2100 | Urban	DNN, Tree, k-NN | N/D	Transmitter-receiver distance, horizontal and vertical angular separation, LOS/NLOS state, first and last diffraction point.	RMSE.
[45]	2021.4 | Urban	ANN, SVR, RF | MSE, N/D	Transmitter-receiver distance.	RMSE, MAPE, MAE, MaxPE and σ.
[87]	700 - 2600 | Urban	ANN | N/D	3-D image generated from LiDAR technology and 2-D satellite image.	MAPE.

Continue on the next page

Table 2.1: ML-based studies on PL prediction for urban, suburban, rural, and indoor environments. (cont.).

Reference	**Freq. [MHz] | Scenario**	**ML alg. | Loss func.**	**Input features**	**Performance indicators**
[79]	operation in UHF band | Urban	RF | MSE	Feature of the transmitter, test point location, and environment features.	RMSE and cost time.
[52]	operation in UHF band | Urban	CNN | MSE	Global information systems (GIS) layers and antenna parameters (location, height, azimuth, tilt, radiation pattern, and frequency).	RMSE.
[53]	operation in UHF band | Urban	CNN | N/D	2D satellite image. The study treats path loss prediction as an image classification problem.	Accuracy.
[63]	450, 1450, 2300 | Suburban	ANN | MSE	Transmitting and receiving antenna height, transmitting/receiving antennas heights ratio, and distance.	RMSE, MAE, MAPE, MSLE and R^2.
[88]	881.52 | Rural	ANN | MSE	Transmitter-receiver distance, antenna height, terrain clearance angle, terrain usage, vegetation type, and vegetation density.	CF.
[89]	900, 1800 | Rural	k-NN, RF | N/D	Distance, altitude, LOS/NLOS state and elevation angle.	RMSE and R^2.
[90]	3700 | Rural	ANN, RF, SVR, B-kNN | N/D	Distance, LOS/NLOS state, effective height between the transmitter and receiver.	ME, MAE, MAPE, and RMSE, σ.
[91]	operation in UHF band | Rural	SVR, RBF | N/D	Elevation, clutter heights, transmitter-receiver distance, altitude, and building to building distance.	MSE.
[46]	1890 | Indoor	ANN, RBF-NN | MSE	Position of the transmitter, gain Tx, height Tx, receiver position, type of interior, distance, number of walls and windows, accumulated loss of walls and windows.	MAE, RMSE and σ.
[83]	900, 1800, 2100, 2400 | Indoor	ANN | quadratic loss error.	Transmitter-receiver distance, frequency, number of crossed walls and floors, angle of incidence with wall and floor, furnishing index, and density of people ratio.	ME, σ and CF.
[82]	900, 1800, 2100, 2400 | Indoor	ANN | MSE	Transmitter-receiver distance, frequency, wall attenuation, and floor attenuation.	ME, σ and CF.
[84]	2400 | Indoor	ANN, RBF-NN | MSE	Relative X,Y,Z coordinates.	CPU and time cost, σ, NMSE and MAE.
[50]	2400, 3520, 5800 | Indoor	ANN, SVR, RF, AdaBoost | N/D	Distance, frequency, and relative coordinates of the receiver position.	MAE, MAPE, RMSE, and R^2.
[54]	28 GHz | Indoor	CNN | N/D	LAMS images.	RMSE.

In addition, some ML-based PL models for rural areas are also available, as pointed out in Table 2.1. In [90], the authors evaluate the PL prediction utilizing different ML algorithms at 3.7 GHz, including ANN, SVR, RF, and bagging with k-NN (B-kNN). In [91], an SVR and a radial basis function (RBF)

regression are developed for PL prediction in a rural environment. In [89], air-to-ground PL prediction models for UAV and the Internet of Things (IoT) in rural environments are carried out using k-NN and RF. In [88], different ANN architectures are employed for PL prediction for macrocell in rural areas.

Furthermore, from the literature review presented, we identify that most of the works employ the conventional random split for training and testing set allocation as described in [45, 46, 55, 63, 78, 82, 83, 84, 88, 90], usually using 80% of the dataset for training and the remaining for testing. We also evaluate the performance prediction of the PL models using random split. In addition, we perform tailored generalization tests according to the environment studied to emphasize the applicability and generalizability of the obtained ML models.

When considering ML-based PL models, some schemes of predictor selection incorporate PCA techniques to reduce the dimension of the dataset [63, 64]. We propose a predictor selection methodology to form coalitions (a significant subset of predictors) with an interpretability gain. The detailed description of the methodology is presented in Chapter 3, Section 3.9.

In addition, we also identify that most works use a single ML technique such as ANN, SVR, RF, or CNN. Some compare their performance using tabular or 2D data. In addition, most of those works consider a single frequency channel in the UHF band. This thesis employs the ANN, SVR, and RF due to their usually good performance to solve radio propagation problems. In addition, we use the GTB model as a PL prediction technique due to its outstanding performance in a wide range of applications [92]. Thus, the performance models are compared using the proposed methodologies to find the optimal ML-based PL model in each scenario and frequency band in the mmWave and sub-6 GHz spectrum. The CNN technique is also addressed with a transfer learning strategy to train the model with a small dataset.

From Table 2.1, we see that different indicators to evaluate the model performance have been used, such as the mean error (ME), mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), standard deviation (σ), mean absolute percentage error (MAPE), mean square logarithmic error (MSLE), mean square normalized error (MSNE), maximum prediction error (MaxPE), correlation factor (CF), regression coefficient (R), and coefficient of determination (R^2). We identify that most of the studies employ the indicators RMSE, MAPE, σ and R^2. We use them to compare the employed models and with the most related works.

2.3 Path loss Prediction using Supervised Learning

In a regression problem, the aim is to develop machines that map the predictors to the response. When using machine learning techniques, this problem is solved as illustrated in Figure 2.2 and described below. The training set is a collection of I input-output pairs $\left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{I}$, such that $x^{(i)} \in \mathbb{R}^P$, P is the number of predictors, and $y^{(i)} \in \mathbb{R}$ is the target. The goal is to find a function $\hat{y} = f(x)$ that minimizes the expected value of the loss function $L(y, \hat{y})$ [93]

$$\hat{y} = \underset{\hat{y}}{\operatorname{argmin}}\, E_{x,y}[L(y, \hat{y})]. \tag{2-5}$$

The loss function $L(y, \hat{y})$ quantifies how closely the model fits the training set [94]. The approximation function depends on the model parameters Θ or degrees of freedom [93], i.e., $\hat{y} = f(x; \Theta)$ [44]. User-defined parameters called hyperparameters control the learning process and the number of parameters [94], which are defined through the model selection process using cross-validation (CV) methods applied on the training set [40, 95]. The process of finding hyperparameters that optimize the learning model commonly involves techniques such as grid search, random search, or Bayesian optimization [40, 95]. The selection of method depends on the dimensionality of the data; for instance, in the case of a large dataset, grid search can become computationally demanding and may not be the most efficient approach.

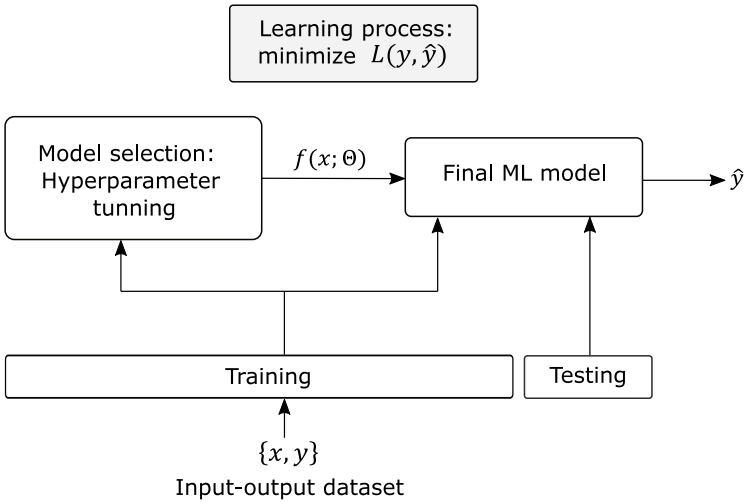

Figure 2.2: Approach for supervised learning for regression.

Once the hyperparameters are selected, the final model is obtained, and a generalization test is performed on it using the testing set. Different machine models present hyperparameters having their own definitions and meanings. In the following sections, we briefly delineate the models employed and their

hyperparameters so that later we can describe the procedures applied to obtain the final models.

2.3.1 Artificial Neural Network

The ANN is inspired by theories about human brain functioning [43]. It simulates the structure and functionalities of a biological neural network [96]. An ANN consists of interconnected basic-processing elements (artificial neurons) in a multi-layered architecture [43, 96]. Each neuron computes a weighted sum of its inputs coming from other neurons. The strength of a connection is determined by the synaptic weight. The resulting value is processed by a non-linear activation function to produce the output [96].

A feedforward ANN or multilayer perceptron (MLP) [96, 97, 98] is composed of L layers. The input layer ($l = 1$) receives the predictor vector $\left\{(x^{(i)})\right\}_{i=1}^{I}$ and, thus, has n inputs. One or more hidden layers ($l < L$) determine the mapping between the input and the output layer ($l = L$) that computes the response $\hat{y}$ of the neural network. The weights connect the outputs of the neurons in one layer to the next layer's inputs. At layer $l > 1$, there are J_l neurons. The connection from the output of the i-th neuron in layer $l-1$ to the input of the j-th neuron in layer l is the weight $w_{ji}^{(l)}$. The response of the j-th neuron in layer l is given by

$$o_j^{(l)} = \varphi_j^{(l)} \left(\sum_{i=1}^{J_{l-1}} w_{ji}^{(l)} o_i^{(l-1)} + b_j^{(l)} \right), \tag{2-6}$$

where $\varphi_j^{(i)}(\cdot)$ is the activation function, $o_i^{(l-1)}$ is the response of the i-th neuron in the previous layer, $b_j^{(l)}$ is the bias of the neuron j in layer l. Common activation functions for MLPs are the sigmoid [43], tansig [99], and ReLU [96]. The training process of an MLP consists of adjusting the weights and biases that minimize the loss function [97]. For regression, the squared error loss is commonly used

$$L(y, \hat{y}) = \sum_{i=1}^{I} \left\| y^{(i)} - \hat{y}^{(i)} \right\|^2. \tag{2-7}$$

Back-propagation is among the most employed training algorithms for MLPs [96, 97]. The algorithm initially computes the response for each neuron from the input to the output layer. Subsequently, the output error is computed $L(y, \hat{y})$ and propagated through the network in the backward direction [97]. For the gradient descent method [100], one computes the partial derivatives of the output error with respect to the weights and biases using the chain rule [100]. Then, one updates the weights and biases in the gradient descent direction of the output error [97, 100]. Consequently, the parameter updates

at iteration $t = 1, .., T$ are computed using

$$\Delta w_{ji}^{(l)}(t) = -\eta \frac{\partial E}{\partial w_{ji}^{(l)}} + \alpha \Delta w_{ji}^{(l)}(t-1), \tag{2-8}$$

$$\Delta b_{j}^{(l)}(t) = -\eta \frac{\partial E}{\partial b_{j}^{(l)}} + \alpha \Delta b_{j}^{(l)}(t-1), \tag{2-9}$$

where $0 < \eta < 1$ is the learning rate that multiplies the gradient, i.e. η is the step-size for the parameters updates [63]. Equations (2-8) and (2-9) also add the update from the previous step $(t-1)$ multiplied by the weight decay $0 < \alpha < 1$ to restrain parameter changes during training (L2-regularization) in order to reduce the model complexity and avoid over-fitting (poor generalization) [101]. Adjustment of weights and biases is performed repeatedly using batches of entries from the training dataset until some stopping criterion is met, such as a number of predefined iterations or early stopping. Other standard training algorithms are Levenberg Marquardt and Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) [102].

2.3.2 Support Vector Regression

SVR makes linear regression using [103]

$$\hat{y} = \langle w, x \rangle + b, \tag{2-10}$$

where $\langle \cdot, \cdot \rangle$ is the dot product in $\mathbb{R}^P$, w and $x \in \mathbb{R}^P$, and $b \in \mathbb{R}$. The normal vector w, the weight, and the bias b are the parameters to be learned [104]. Therefore, the SVR learning process aims to find w and b that minimize the ε-insensitive loss function [103, 105]

$$L(y^{(i)}, \hat{y}^{(i)}) = \begin{cases} |y^{(i)} - \hat{y}^{(i)}| - \varepsilon & \text{if } |y^{(i)} - \hat{y}^{(i)}| > \varepsilon \\ 0, & \text{otherwise} \end{cases}, \tag{2-11}$$

where ε represents the tolerated deviation between the target and the predicted value. The ε-insensitive function defines a tube of width ε $(\hat{y} \pm \varepsilon)$, and the datapoints whose $y^{(i)}$ value differs from $\hat{y}^{(i)}$ by more than ε are penalized [104].

Therefore, the design objective is to find a function $\hat{y}$, the hyperplane defined by w and b in Equation (2-10), such that $\hat{y}^{(i)}$ deviates from $y^{(i)}$ by less than ε. Meanwhile, the tube must be as flat as possible [105], while containing most of the data points [103]. The data points outside the tube are known as support vectors and are used to build the model [103]. A small value of ε indicates a low error tolerance, impacting the number of support vectors that are stored/learned [104].

The SVR model is obtained by minimizing

$$\min_{w,b,\xi^{(i)},\xi^{(i)*}} \frac{1}{2}||w||^2 + C\left(\sum_{i=1}^{I}(\xi^{(i)}+\xi^{(i)*})\right), \qquad \text{subject to} \begin{cases} y^{(i)} - \hat{y}^{(i)} \leq \varepsilon + \xi^{(i)}, \\ \hat{y}^{(i)} - y^{(i)} \leq \varepsilon + \xi^{(i)*}, \\ \xi^{(i)},\ \xi^{(i)*} \geq 0, \end{cases} \tag{2-12}$$

where $\hat{y}^{(i)}$ is computed using Equation (2-11), and $\xi^{(i)}$ and $\xi^{(i)*}$ indicate the differences between the target $y^{(i)}$ and the upper and lower bound in the tube, respectively [47]. C is a regularization factor that determines the trade-off between the flatness of the tube and the amount up to which deviations that are larger than ε are tolerated [104]. Given the Lagrange multipliers α_p and α_p^*, a kernel function $K(x, x_p)$, and a number of support vectors ($\mathrm{P_{sv}}$), the desired hyperplane is given by [103]

$$\hat{y} = \sum_{p=1}^{\mathrm{P_{sv}}}(\alpha_p - \alpha_p^*)K(x, x_p) + b_p, \tag{2-13}$$

$$b_p = y_p - w^T x_p - \varepsilon \text{ for } \alpha_p \in (0, C), \tag{2-14}$$

$$b_p = y_p - w^T x_p + \varepsilon \text{ for } \alpha_p^* \in (0, C). \tag{2-15}$$

When the Kernel function performs a non-linear mapping, the SVR makes the linear regression in the transformed space, allowing improved performance for problems departing from linear maps. A commonly employed kernel is the radial basis function (RBF) $K(x, x_p) = \exp(\| x - x_p \|^2 / \sigma^2)$, where σ controls the shape of the non-linear mapping [104].

2.3.3 Random Forest

A decision tree is given by [43]

$$T(x; \Theta) = \sum_{j=1}^{J} \gamma_j I(x \in R_j). \tag{2-16}$$

It is defined by the parameters $\Theta = \{(R_j, \gamma_j)\}_{j=1}^{J}$. The subscript j is the number of the leaf or terminal nodes of the tree. At each node, a partition using rules (if-then conditions) on the predictors is applied [44, 106]; g predictors randomly selected from all the existing m predictors are applied in the rule to decorrelate the tree and reduce variance [107]. The rules define regions R_j, and γ_j is $\hat{\gamma}_j = \overline{y}_j$, the average y_j for the samples that fall in the region R_j.

A widely used recursive partitioning algorithm to obtain the rules is the classification and regression tree (CART) algorithm [108]. It searches for the initial node of the tree considering s possible split nodes for every predictor. The best predictor and split value that minimizes the loss function is chosen

to split the data into two child nodes or regions [43, 109]. To find the split into the regions R_i and R_j in each tree, for the g predictors and split values s, one solves

$$\min_{g,s}\left[\min_{\overline{y}_i}\sum_{x_p\in R_i(g,s)}(y_p-\overline{y}_i)^2+\min_{\overline{y}_j}\sum_{x_p\in R_j(g,s)}(y_p-\overline{y}_j)^2\right], \tag{2-17}$$

it sums the squared errors of attributing the response value $\overline{y}_i$ to the input values belonging to R_i and $\overline{y}_j$ to the input values belonging to R_j. In each split, the response is modeled by the mean $\overline{y}$ in each region. Then, one or both child nodes are divided into two regions using the same splitting rule design criterion [43] to make the tree grow; the process continues until a stopping rule is met, such as the maximum depth tree, which defines the complexity of the model [43].

A RF is a collection of tree-based models [110]. Bootstrapping is employed to create Q independent subsets with s randomly chosen training samples with replacement from the original training set [93]. They are used to learn different regression trees $\{T(x;\Theta_q)\}$, $q=1\dots Q$, in parallel, that is, each tree is trained with a different subset of the input-output pairs. Their individual responses are averaged to compute the response [93, 110],

$$\hat{y}=\frac{1}{Q}\sum_{q=1}^{Q}T(x;\Theta_q). \tag{2-18}$$

2.3.4 Gradient Tree Boosting

The GTB algorithm obtains an ensemble of trees similarly to the RF. However, while the RF independently builds each tree from the bootstrap samples, the GTB learns the tree sequentially using the residuals (errors) between the target and predicted values by the previous trees to compensate for prediction errors [93]. At each iteration $q=1,...,Q$ [44, 93], for training the q-th tree one computes the residual values $r_q^{(i)}=y^{(i)}-\hat{y}_{q-1}^{(i)}$, where $\hat{y}_{q-1}^{(i)}$ is the predicted value obtained using up to the $q-1$ tree. This applies for all training sample $\left\{(x^{(i)},r_q^{(i)})\right\}_{i=1}^{I}$. Consequently, for the j-th terminal node of the q-th tree, the optimal predicted value is the one minimizing

$$\gamma_{jq}=\underset{\gamma}{\operatorname{argmin}}\sum_{x^{(i)}\in R_j^{(i)}}L\left(y^{(i)},\hat{y}_{q-1}^{(i)}+\gamma\right). \tag{2-19}$$

Once the residuals $r_q^{(i)}$ and the node values γ_{jq} are computed for the new tree, the model is updated using

$$\hat{y}_q = \hat{y}_{q-1} + \eta \sum_{j=1}^{J_q} \gamma_{jq} I(x \in R_{jq}), \tag{2-20}$$

where η is the learning rate ($0 < \eta < 1$), until $q = Q$.

2.3.5 Convolutional Neural Network

The CNN is a deep learning model that has shown great performance to extract features from 2D images and other data [55, 111, 112]. This ability has surpassed the manual design of features/predictor extraction for classification problems [111]. Nowadays, CNNs are employed to solve computer vision tasks, including classification, detection, and segmentation [113, 114]. Some architectures that are being widely used in computer vision are LeNet, AlexNext, U-Net, VGG, InceptionNet, and GoogleLeNet, which are usually trained with huge datasets [115, 116]. Figure 2.3 shows a basic CNN architecture with three layers, including a convolutional, pooling, and fully connected layer [117].

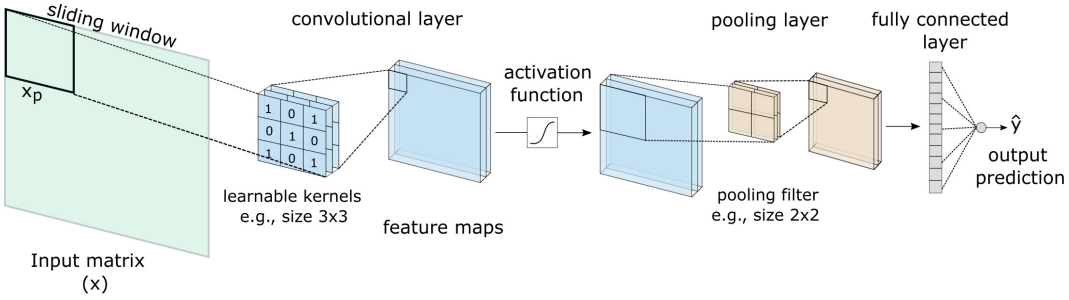

Figure 2.3: Basic architecture of a CNN model.

The convolution layer performs a linear operation to extract Q feature maps (patterns) for $q = 1, \ldots, Q$ from the input matrix; the convolution process entails sliding a set of trainable kernels or filters of an user-defined size over an input region defined by x_p. The convolution layer aims to learn the weights (k_{pq}) and biases (b_q) to detect a specific pattern, such as edges and corners [117, 118]. The summation of the convolution operations is passed through a nonlinear activation function $\varphi(\cdot)$; and the output for the q-feature map is given by [119]

$$o_q = \varphi\left(\sum_{p \in M_q} x_p * k_{pq} + b_q\right), \tag{2-21}$$

where M_q defines the selection of inputs (i.e., the number of pixels to be processed by the kernel), $*$ represents the convolution operation, k_{pq} is the trainable convolution kernel, and a bias b_q is added after the convolution operation [119]. The pooling layer performs a transformation that reduces

the size of the feature map (o_q), summarizing its outputs by calculating the maximum or average values in non-overlapping regions as illustrated in Figure 2.3 [117, 118]. Finally, the output obtained from the convolution, activation, and poling layer process is referred to as the feature/predictor map, representing the detected patterns [120], which is flattened to a vector υ that is fed to a fully connected neural network layer as described in Section 2.3.1, Equation (2-6). This last layer performs the mapping between the input υ and the final prediction $\hat{y} = \varphi(w\upsilon + b)$ adjusting their weights (w) and biases (b) during training.

Therefore, the trainable parameters of the CNN include the convolution kernels, the weights in the fully connected layer, and biases in the different layers [117]. Those parameters are learned through a recursive iteration of forward- and back-propagation. In the forward propagation, the input data is propagated forward through the CNN layer by layer to perform prediction [117]. In the back-propagation, the gradient of the loss function is calculated with respect to each parameter in the CNN (k_{pq}, b_q, w, b) using the chain rule [117].

The inclusion of deeper convolution layers allows the detection of more complex shapes [121]. During the training of the CNN model, it is possible to freeze either the entire or a fraction of the parameters within the predictor extractor for application in a different task; this is called transfer learning [106]. This is justified by the rationale that the extractor predictor learned for a given task may identify patterns from unseen images far better than a predictor extractor being trained from scratch and even better if using a small dataset [106]. The decision to freeze parameters may depend on the specific domain of the new task. If the CNN parameters are already trained to extract relevant features, only the parameters of the fully connected layer have to be trained for the task – to adjust it adequately for the new [106, 122].

2.4 Interpretability Machine Learning Techniques

This section provides the fundamental for proposed interpretability machine learning techniques with applications for radio-propagation modeling. They are based on mutual information (MI) and interpretable machine learning (IML) tools. The MI is model agnostic and computed on the dataset, providing pre interpretations. Meanwhile, IML tools provide post-hoc interpretations, analyzing the models' responses after training; they are post-hoc interpretation tools. They try to assess how the model behaves if the predictors change [69]. In this sense, we use some IML tools to examine the relations be-

tween the predictors and the path loss (response) and interpret the ML models and their complexities.

2.4.1 Mutual Information

The MI may be an analysis tool to detect relevant and redundant predictors before building the machines [123]. One may employ such prior knowledge to select the most relevant predictors for training the models ignoring less relevant ones. The MI measures linear and non-linear correlations between variables [124]. If MI is equal to zero, then the variables are independent; meanwhile, higher values indicate higher dependency [125]. Thus, assessing the MI within the dataset analyzes the relations between predictors and the response. Furthermore, the joint mutual information (JMI) [124] and conditional mutual information (CMI) [126, 127] may detect relevant and redundant predictors prior to model specification and training, and we employ them in a greedy predictor selection algorithm. The JMI $I(x_p, \ldots, x_{p+l}; y)$, $p, \ldots, p+l \in 1 \ldots P$, quantifies the information shared by the subset of predictors $\{x_p, \ldots, x_{p+l}\}$ and the target y. Meanwhile, the CMI $I(y; x_p | x_{\setminus p})$ quantifies the MI between y and x_p conditioned on the subset of other predictors $x_{\setminus p}$.

2.4.2 Interpretable Machine Learning Tools

IML techniques aim at obtaining (approximate) explanations for the model behavior [128]. Their goal is to discover the relationship between the model's inputs and outputs globally and locally. The global interpretation tools reflect the response changes to predictors changes considering the entire model behavior. Local interpretation indicates which predictors mainly influence a single model response instance [129]. This section employs them to look into the ML-based PL models to devise the more appropriate predictors coalitions to use (a significant subset of predictors).

Permutation feature importance. The PFI for a predictor is computed by randomly permuting its values in the dataset and computing the resulting change in the response. If we measure the loss in performance with the RMSE, the PFI for the predictor x_p is given by

$$\mathrm{PFI}_p = \frac{1}{D} \sum_{j=1}^{D} \mathrm{RMSE}_{p,j}^{(\mathrm{permuted})} - \mathrm{RMSE}_{p,j}^{(\mathrm{original})}, \tag{2-22}$$

the average difference between the response errors for the original and the permuted dataset, using D permutations. It tests if the permutation neutralizes the predictor influence on the response [130]. One notes that the larger the PFI is for a given predictor, the more it impacts the response. Consequently, the PFI helps ranking the predictors' influence on the model response.

Accumulated local effect. The ALE quantifies the overall individual effect (the main effect) of a predictor (x_p) in the model, ignoring the effect of all other predictors ($x_{\setminus p}$) [131, 132]. To compute the ALE, one partitions the predictor distribution using Q intervals. For each instance of the p-th predictor ($x_p^{(i)}$) belonging to the correspondent q-th interval, one computes the difference between the responses obtained when replacing $x_p^{(i)}$ by the upper and the lower limits of the interval, $z_p^{(q)}$ and $z_p^{(q-1)}$, respectively. This is accumulated over all the intervals for the p-th predictor leading to

$$\mathrm{ALE}(x_p) = \sum_{q=1}^{q(x_p)} \left\{ \frac{1}{M_p(q)} \sum_{i \mid x_p^{(i)} \in \mathcal{M}_p(q)} \left[\hat{y}(z_p^{(q)}, x_{\setminus p}^{(i)}) - \hat{y}(z_p^{(q-1)}, x_{\setminus p}^{(i)}) \right] \right\}. \quad (2\text{-}23)$$

The inner summation adds the effects of all instances within the q-th intervals, it considers the number of instances $|\mathcal{M}_p(q)|$ in each q-th interval and $\mathcal{M}_p(q)$ represents this list, while the outer sum spans all the intervals into which x_p may fall, $q(x_p)$ is the index for the last interval.

At last, ALE is centered around the average mean effect, that is,

$$\mathrm{ALE}_{\mathrm{cent}}(x_p) = \mathrm{ALE}(x_p) - \frac{1}{I} \sum_{i=1}^{I} \mathrm{ALE}(x_p^{(i)}). \quad (2\text{-}24)$$

The term on the right averages the ALE over instances I of the p-th predictor. The $\mathrm{ALE}_{\mathrm{cent}}(x_p)$ is graphically presented using a curve. This curve analyzes the effect of the predictor value (abscissa) w.r.t the average effect of the predictor [130]. It can reveal whether the predictor effect on the response is mainly linear or non-linear.

Interaction effects. One may extend the ALE to compute the interaction effects for predictor pairs. The base formulation follows the one for ALE, but instead of one-dimensional intervals, one uses rectangular cells (2D intervals) to accumulate the second-order effects [130, 132]. If the second-order ALE value is close to zero for two features, they have a low interaction effect [130].

Interaction strength. The IAS considers the overall interaction between predictors. It quantifies the extent to which the prediction function can be

approximated by the first-order effects of the predictors [131]. In [131], the IAS is estimated using a functional decomposition that employs the ALE and the prediction function, that is

$$\text{IAS} = \frac{\sum_{i=1}^{I}(\hat{y}(x^{(i)}) - \text{ALE}_{1^{\text{st}}}(x^{(i)}))^2}{\sum_{i=1}^{I}(\hat{y}(x^{(i)}) - \overline{\hat{y}})^2} = 1 - R^2, \tag{2-25}$$

where $\hat{y}_{\text{ALE}_{1^{\text{st}}}}$ is the sum of the first-order ALE effects for all the predictors, $\hat{y}(x^{(i)})$ is the prediction function, and $\overline{\hat{y}}$ is the average response. IAS values close to 0 indicate very low interaction strength between predictors. In this case, the response is mainly influenced by individual predictor effects. Meanwhile, high IAS values suggest significant interactions between predictors and their effects on the response derived from interactions.

Main effect complexity. The average MEC derives from ALE curves to quantify the non-linearity of first-order effects [131, 133]. First, one finds MEC_p the number of linear segments for approximating the ALE curve for predictor $p \in \{1, ..., P\}$ – clearly a measure of the complexity of the mapping between the predictor and the response. MEC_p is estimated if the approximation given by a linear model reaches the condition $R^2 \geq 1 - \epsilon$ (or a maximum number of segments is already employed), ϵ is the user-defined maximum approximation error; we use ϵ equal to 0.05. To obtain the mean complexity of the model, one averages the MEC_p overall predictors. Thus, the average MEC is obtained using [131]

$$\text{MEC}_p = \text{number of linear segments approximating ALE}(x_p), \tag{2-26}$$

$$\sigma_p = \frac{1}{I}\sum_{i=1}^{I}\text{ALE}(x_p^{(i)})^2, \tag{2-27}$$

$$\overline{\text{MEC}} = \frac{1}{\sum_{p=1}^{m}\sigma_p}\sum_{p=1}^{m}\sigma_p \text{MEC}_p. \tag{2-28}$$

We note that to compute the average, the weight for each MEC_p is the variance of the ALE main effect (σ_p), strengthening predictors that greatly contribute to the outcome. One readily sees that if $\overline{\text{MEC}}$ is close to unity, then the model has low complexity – one obtains a good approximation with few linear segments. Consequently, the relations among the model's inputs (predictors) and outputs (responses) are easier to interpret. IML indicators computed on the training set assess how much the model response relies on each predictor and how the model behaves upon changes in the predictor values. IML indicators computed on the testing set provide insights into the model's behavior for unseen data [128].

2.5 Summary

In this chapter, we have described the main aspects of the radio-propagation effects that are usually considered for channel modeling. The mathematical formulation presented for each supervised learning regression and IML technique gives a deeper understanding of the internal mechanisms of each machine, which are useful for hyperparameters selection and model design, as well as the fundamental characteristics of the interpretability techniques employed in the following chapters.

3
Path Loss Prediction for mmWave Indoor Communications using Machine Learning Techniques

The design of mmWave communication systems requires accurate PL prediction, which is critical to determining coverage area and system capacity. In mmWave indoor scenario, existing objects, including walls and constitutive materials, influence propagation. The wave propagates along corridors and other open areas, depending on the structure of the building [46]. All obstacles cause multiple paths through reflection, refraction, and diffraction phenomena, as well as waveguide effects in corridors [3, 36].

This chapter addresses PL models for an indoor environment at frequencies between 26.5 GHz and 40 GHz. Several supervised machine learning techniques are employed and their performances are compared with appropriate empirical PL models for the mmWave frequency band.

3.1
Related Work

Previous works of PL prediction based on ML models for indoor environments present some interesting results for comparison. The study in [46] employs ten predictors containing information about the transmitter and receiver sites, distance, and parameters for the objects in the environment; the authors report root mean squared error (RMSE) between the measured and the predicted PL of 4.23 and 4.38 dB and correspondent standard deviation (σ) values of 2.88 dB and 3.15 dB for the ANN and RBF in the 1.89 GHz frequency band, respectively.

In [83], the authors achieve a σ value of 4.4 dB using an ANN model with eight inputs, considering the body shadowing and furniture effects, in the frequency range from 900 MHz to 2400 MHz. In [84], combining ray-launching and ANN model, authors achieve a σ value of 5.96 dB in the frequency of 2.4 GHz, using relative coordinates between transmitter and receiver. In [82], the authors report a standard deviation (σ) value of 5.22 dB using four predictors, including frequency, distance, and the traversed walls and floors. Finally, in [54], a CNN model achieves RMSE values between 5.01 to 5.35 dB for different training set configurations for 28 GHz waves.

This study employs four ML algorithms: ANN, SVR, RF, and GTB. The model hyperparameters are tuned for every algorithm, thus obtaining reliable results. We compare their performances, including also the empirical PL models ABG and CIF and their extended versions incorporating the number of traversed walls. For the ML models, we present a methodology to select the predictor coalitions by examining the marginal performance and interpretability gains. The methodology allows choosing the most appropriate predictor coalition for building machines, while not overlooking the non-linear connections between predictors and the path loss. Also, pre and post-hoc interpretation tools are proposed and employed to select and analyze the different ML models for comparison. At last, the generalization capacity of the resulting models is assessed by applying the models for unknown transmitter and receiver locations, i.e., for links with characteristics ignored when obtaining the PL models.

3.2 The Measurement Campaign: Dataset Description

3.2.1 Path Loss Measurement Campaign

We employ data from a measurement campaign in an indoor environment at the Center for Telecommunication Studies in PUC-Rio (CETUC) on frequencies ranging from 26.5 GHz to 40 GHz in steps of 0.5 GHz. The measurements occurred in October 2018 and were conducted by colleagues from the radio propagation laboratory in CETUC [134]. The measurement at each point and frequency lasts between 40 to 60 seconds. The building is composed of offices, classrooms, and laboratories. A signal generator Anritsu MG3696B was used to transmit a continuous wave (CW) at 0 dBm. The receiver is a signal analyzer Anritsu MS2668C. If P_{Rx} is the received power level measured in dBm, one obtains the PL in dB from the measured received power and system parameters from

$$PL = P_{Tx} + G_{Tx} + G_{Rx} - P_{Lc} - P_{Rx}, \tag{3-1}$$

where P_{Tx} is the transmitted power in dBm, G_{Tx}, and G_{Rx} are the transmitter and receiver antenna gains in dBi, P_{Lc} is the cable loss, with maximum values of 2.40 dB at 20 GHz and 3.61 dB at 40 GHz. The transmitting and receiving antennas are WR-28 waveguide horns with a 20 dBi gain. The signal analyzer sensitivity is -101 dBm, and the transmission power is 0 dBm. Therefore, it is possible to measure path losses as low as -139 dB with a precision error in the

measurements around ±0.5 dB. This precision error value can be considered for all the measurement campaigns presented in this thesis.

Six transmitter positions were considered for full coverage of the indoor environment, and there were 40 reception points. The number of reception points perceiving each transmission point is limited by the walls, other construction aspects, and receiver sensitivity. As observed in Figure 3.1, TX1 (gray star) is perceived at 9 RX points (gray circles); TX2 (red star), at 6 RX points (red circles); TX3 (blue star), at 8 RX points (blue circles); TX4 (green star), at 7 RX points (green circles); TX5 (orange star), at 7 RX points (orange circles) and TX6 (brown star), at 3 RX points (brown circles).

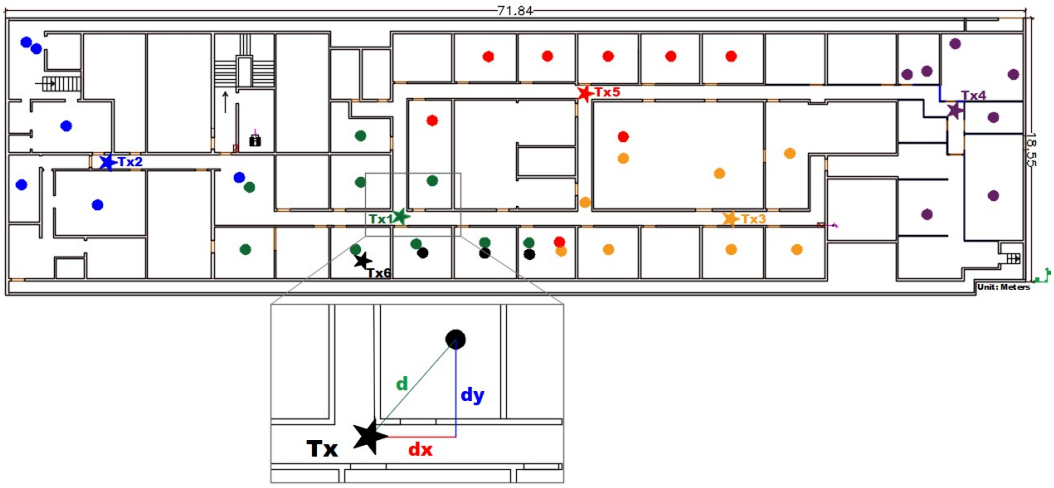

Figure 3.1: Map of the indoor measurement scenario, PUC-Rio, CETUC, and schematic view of the distance predictors. Each star represents a transmission position, while the circles with the same color indicate the correspondent positions where receivers were placed for measuring the path loss.

The campaign produces a dataset of 1,120 measured PL instances through Equation (3-1), considering the six transmissions versus the 40 possible reception points and the 28 frequencies. All the 1,120 measurements correspond to non-line-of-sight (NLOS) cases. Each sample is associated with four numerical attributes: carrier frequency (f) in GHz, the Tx-Rx distance (d) in meters, and its decomposition into vertical (d_y) and horizontal (d_x) components, together with a category attribute: the number of obstructing walls (n_w) between Tx and the Rx.

3.3 Empirical Path Loss Models

Empirical PL models fit curves or analytical expressions to the mapping between physical parameters (mainly frequency and distance) of the link and the measured path loss [30, 31, 135, 136]. In practice, the parameters of the

empirical model are adjusted using non-linear regression techniques. Some common PL models for indoor environments at mmWave are the ABG and CIF models [22]. The ABG model is a multi-frequency PL model commonly used for a broad spectrum of frequencies [34]. The model is based on a regression fit that includes frequency and distance dependence and is given by

$$\mathrm{PL}^{\mathrm{ABG}}[\mathrm{dB}] = 10\alpha \log_{10}(d) + \beta + 10\gamma \log_{10}(f) + X_{\sigma}^{\mathrm{ABG}}, \tag{3-2}$$

where d is the Tx-Rx distance in meters, f is the frequency in GHz, α and γ represent the PL coefficients that characterize the dependence on distance and frequency, respectively, and β indicates the PL offset. In the expression, $X_{\sigma}^{\mathrm{ABG}}$ is a zero-mean Gaussian random variable (in dB) describing the large-scale shadowing [63].

The 3GPP proposed the CIF model [35] as an extension of the close-in (CI) model. It includes a frequency-distance combined dependence term using [35]

$$\mathrm{PL}^{\mathrm{CIF}}[\mathrm{dB}] = \mathrm{FSPL}[\mathrm{dB}] + 10n \log_{10}(d) \left(1 + b\left(\frac{f - f_o}{f_o}\right)\right) + X_{\sigma}^{\mathrm{CIF}}, \tag{3-3}$$

where FSPL denotes the free space PL model ($20 \log_{10}\left(\frac{4\pi\ d\ f}{c}\right)$). The coefficients n and b are the distance and the linear frequency PL coefficients. The reference frequency $f_o = \frac{\sum_{j=1}^{J} f_j N_j}{\sum_{j=1}^{J} N_j}$, where J is the number of unique frequencies, N_j is the number of data points corresponding to the j^{th} frequency f_j, and $X_{\sigma}^{\mathrm{CIF}}$ represents the shadow fading deviation describing large-scale signal fluctuations. The CIF model becomes the CI model when a single frequency is available to adjust the model ($f_o = f$ or $b = 0$).

Inclusion of the number of traversed walls in the ABG and CIF models
Various indoor PL models have been proposed in the literature to address attenuation loss due to obstructing walls. In [137], the proposed model incorporates a wall loss factor derived by multiplying an attenuation loss associated with each type of wall by the number of such walls. The accumulated wall losses are then calculated for the PL prediction at 864 MHz and 1728 MHz. The model in [138] considers the average attenuation losses for the frequency at 2.4 GHz, considering the type of wall material. In our study, since the constructive material in the different offices, classrooms, and laboratories is the same (plaster panels), we mainly evaluate the effect of the number of walls in the empirical models. The adjusted PL models ABGnw and CIFnw are proposed,

given by

$$\mathrm{PL}^{\mathrm{ABGnw}}[\mathrm{dB}] = 10\alpha \log_{10}(d) + \beta + 10\gamma \log_{10}(f) + \delta n_w + X_\sigma^{\mathrm{ABG}}, \text{ and} \quad (3\text{-}4)$$

$$\mathrm{PL}^{\mathrm{CIFnw}}[\mathrm{dB}] = \mathrm{FSPL}[\mathrm{dB}] + 10n \log_{10}(d) \left(1 + b\left(\frac{f - f_o}{f_o}\right)\right) + \delta n_w + X_\sigma^{\mathrm{CIF}}. \quad (3\text{-}5)$$

They include the number of traversed walls n_w multiplied by the parameter δ representing the wave's average attenuation when traversing the wall. Other scenarios considering the wall material types could also be addressed to evaluate its effect on indoor mmWave PL prediction.

3.4 Methodology for Model Selection

Model selection aims to tune the model's hyperparameters to improve learning [93, 102]. Next, we present the hyperparameters that define each model. The hyperparameters for the ANN design include the number of hidden layers, the number of neurons in each layer, activation function, learning rate, and weight decay. Some hyperparameters of the ANN define its structure and how many parameters (the weights and biases) must be learned; meanwhile, others manage how the parameters are learned.

The SVR machine depends on the number of support vectors, which depend on the regularization parameter C and the width of the tube ε, controlling the model's complexity and prediction accuracy [105]. The hyperparameters of a tree are its depth, the number of predictors considered selecting the best split, and the leaf size, which refers to how many samples are required for a leaf node to exist [44, 139]. The tree parameters define the regions (predictors and split values for the nodes) and the correspondent responses. The RF and the GTB hyperparameters include the number of trees, their maximum depth, minimum leaf size, and the number of predictors at each split node. In addition, the learning rate must be defined as well for the GTB that controls the effect of each new tree on the final response [93].

For the presented results, we first obtain the hyperparameters that minimize the indicator function (maximize performance and, by extension, learning). More specifically, we apply cross-validation (CV) for model selection [93, 102]. We use the K-fold CV, dividing the training data into K subsets $\mathcal{D}_k$, $k = 1 \ldots K$, of approximately equal sizes $|\mathcal{D}_k|$. Thus, we train the model using $K - 1$, and the remaining subset is used for validation.

To evaluate performance, we apply the indicator root mean squared error (RMSE), given by

$$\mathrm{RMSE}_{k,\lambda} = \sqrt{\frac{1}{|\mathcal{D}_k|} \sum_{x^{(i)} \in \mathcal{D}_k} \left[y^{(i)} - \hat{y}_\lambda^{(i)} \right]^2}, \tag{3-6}$$

where $\hat{y}_\lambda^{(i)}$ denotes the model response to $x^{(i)}$ with the hyperparameters λ. The average CV RMSE is

$$\overline{\mathrm{RMSE}}_{\mathrm{CV}} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{RMSE}_{k,\lambda}. \tag{3-7}$$

We use the entire dataset, but part of it is reserved as the test set to evaluate the final learned model. Upon the tuning of the model hyperparameters, we use them when training the final model. Therefore, to refer to the performance evaluation when tuning the hyperparameters (or making a model selection), we will use the term "validation"; meanwhile, the term "test" refers to the performance of final models.

Besides the RMSE, in this thesis, three other performance indicators are employed to evaluate the final models: the mean absolute percentage error (MAPE), the standard deviation (σ), and the coefficient of determination (R^2). MAPE is the average of the absolute percentage errors of the predictions given by

$$\mathrm{MAPE} = \frac{100\%}{I} \sum_{i=1}^{I} \left| \frac{y^{(i)} - \hat{y}^{(i)}}{y^{(i)}} \right|, \tag{3-8}$$

where I is the total number of test samples. In addition, the σ of the error is given by

$$\sigma = \sqrt{\frac{1}{I} \sum_{i=1}^{I} \left(|y^{(i)} - \hat{y}^{(i)}| - \mu \right)^2}, \tag{3-9}$$

where μ is the mean prediction error.

The indicator R^2 describes the percentage of variability of the regressed values that can be explained by the model [45]. In general terms, R^2 measures how well the inputs predict the output [140] and it is given by

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{I} \left(y^{(i)} - \hat{y}^{(i)} \right)^2}{\sum_{i=1}^{I} \left(y^{(i)} - \overline{y} \right)^2}. \tag{3-10}$$

R^2 depends on the sum of squared errors (SSE) of the outputs, and on the sum squared total (SST), the standard deviation of the output.

3.5 Design of the Empirical Models: Results and Analyses

This section presents the results obtained for PL models using the techniques described in the previous section. We use 80% of the dataset to adjust the coefficients of the empirical models and tune the hyperparameters for the ML models. We evaluate the results using the K-fold CV approach

with $K = 5$, leading to a reliable comparison between models. We obtain the coefficients of the empirical models, Section 3.3, using four randomly drawn subsets from the training dataset and evaluate the prediction accuracy using the remaining instances. Table 3.1 and Table 3.2 present the coefficient values for the ABG and CIF models, respectively, for each fold. Finally, the average CV RMSE for the four empirical models are presented in Table 3.3.

Table 3.1: Coefficients of the ABG and ABGnw models obtained from the CV subsets.

CV Subset	ABG			ABGnw			
	α	β	γ	α	β	γ	δ
1	4.83	-7.36	5.47	1.39	1.22	5.34	11.02
2	4.60	-4.53	5.39	1.35	8.12	4.95	10.53
3	4.90	-9.46	5.56	1.56	-2.62	5.56	10.34
4	5.06	6.45	4.39	1.68	2.57	5.12	10.55
5	4.78	-11.66	5.77	1.35	-2.21	5.57	11.03

Table 3.2: Coefficients of the CIF and CIFnw models calculated from the CV subsets.

CV Subset	f_o	CIF		CIFnw		
		n	b	n	b	δ
1	31.81	3.69	-0.19	1.63	-0.41	8.70
2	31.88	3.94	-0.11	1.98	-0.22	8.00
3	31.86	3.98	-0.11	1.87	-0.30	7.80
4	31.81	3.63	-0.17	0.52	-0.72	4.58
5	31.88	3.77	-0.16	1.67	-0.31	9.00

Table 3.3: Performance evaluation of the empirical models with CV, values are in dB.

CV Subset	ABG	CIF	ABGnw	CIFnw
1	11.20	11.47	10.07	10.84
2	11.63	11.51	9.36	9.61
3	11.36	11.79	8.87	9.86
4	11.90	12.58	9.30	10.71
5	10.68	10.91	9.13	9.78
Average	**11.35**	**11.65**	**9.35**	**10.16**

The ABGnw presents the lowest CV RMSE (9.35 dB), which is 2 dB smaller than the one for the ABG (11.35 dB). At the same time, the CIF presents the highest one, 11.65 dB, which is larger than the one for the CIFnw, 10.16 dB. It is clear that including the number of walls between the transmitter and receiver improves the model accuracy.

3.6 Design of the ML Models

3.6.1 Hyperparameters Tuning for the ML Models

For the ML models, the predictors $x_p,\ p = 1 \dots P$ with numerical values and the response y (measured path loss) are independently scaled using the mean and the standard deviation. This allows the data to be normalized and ensures that the predictors and output are on a similar scale. We encode the number of obstructing walls, including cases of one, two, or three walls, to 0.25, 0.5, and 0.75, respectively. The selection process employs the 5-fold CV procedure. We employ a grid search to tune the hyperparameters of the models. All grids contain 180 points to evaluate the performance of the models by considering a wide range of hyperparameter values.

The ANN has one hidden layer where we tested different activation functions, such as ReLU (Rectified Linear Unit), Logistic, and Tanh; the ReLU led to the best results as can be seen in Appendix A. In addition, three hyperparameters are varied; the number of neurons in the hidden layer, the learning rate (η), and the weight decay (α_{wd}) ranging from 10 to 86, 0.001 to 0.01, and 0.001 to 0.01, respectively. The L-BFGS solver [139] is employed for weight optimization using at most 6,000 iterations and early stopping to avoid over-fitting.

For the SVR, one considers an RBF kernel evaluated with σ_{RBF} between 0.1 and 0.3. The other two hyperparameters considered are C and ε ranging between 200 to 2100, and 0.005 to 0.1, respectively. We consider the number of trees ranging from 8 to 246 for the RF hyperparameters. Each tree has a maximum depth ranging from 3 to 6, and each leaf's minimum number of samples varies between 1 and 3. The entire set of predictors can be employed in every node of the trees. For the design of the GTB regression, we employ from 8 to 246 trees. Again, as for the RF, the rules at every node may employ all the predictors, each tree has a maximum depth between 3 and 6, and the minimum number of samples in each leaf varies between 1 and 3. Meanwhile, the learning rate was set to 0.1.

The ANN model attains its lowest 5-fold average RMSE (5.05 dB). As observed from Figure A.1.(a), when the number of neurons increases, the CV RMSE value on the validation set decreases, indicating better performance with increasing the number of neurons. Thus, the optimal hyperparameters are for 74 neurons in the hidden layer, a learning rate set to 0.1, and a weight decay equal to 0.1. The SVR model attains its lowest 5-fold average RMSE

(5.83 dB) for σ_{RBF}, C, and ϵ equal to 0.2, 2,100 and 0.1, respectively. The RF model attains its lowest 5-fold average RMSE (5.18 dB) using 178 trees, maximum depth equal to 6, and minimum samples in each leaf equal to 1, considering the entire set of predictors. For the GTB model, we also observe an improvement in performance (lower CV RMSE value) when the number of trees increases as seen in Figure A.1.(d). Therefore, the GTB model attains its lowest 5-fold average RMSE (4.37 dB) using 246 trees, maximum depth equal to 3, and minimum samples in each leaf equal to 3, considering the entire set of predictors. As one sees, the GTB attains the best performance among the considered models. The visualization plot for the average CV RMSE grid points can be seen in Appendix A, Figure A.2.

3.6.2 Final Models Comparison

To obtain all final models, we employ the same training set. They are obtained using the 80% of the dataset employed for the empirical models' consistency analysis in Section 3.5 and the ML models hyperparameters tuning in Section 3.6.1. The remaining 20% of the dataset instances test the models, and we present the results now. Meanwhile, the adjusted coefficients of the empirical models are presented in Table 3.4 and the hyperparameters of the ML algorithms in Section 3.6.1.

Table 3.4: Adjusted coefficients of the empirical models.

	α	β	γ	δ
ABG	4.65	-0.59	5.03	-
ABGnw	1.47	7.30	4.94	10.38
	n	b	f_o	δ
CIF	4.28	0.311	33.17	-
CIFnw	2.12	0.58	33.17	9.04

Figure 3.2 illustrates the degree of correlation between the measured and predicted PL values for the different models. For the ML models, in the right-hand graph in Figure 3.2, the value of R^2 is always greater than 0.87, indicating a good fit between the measured and predicted values. For the empirical models, in the left-hand graph in Figure 3.2, the R^2 is less than 0.62 presenting greater dispersion and a worse fit.

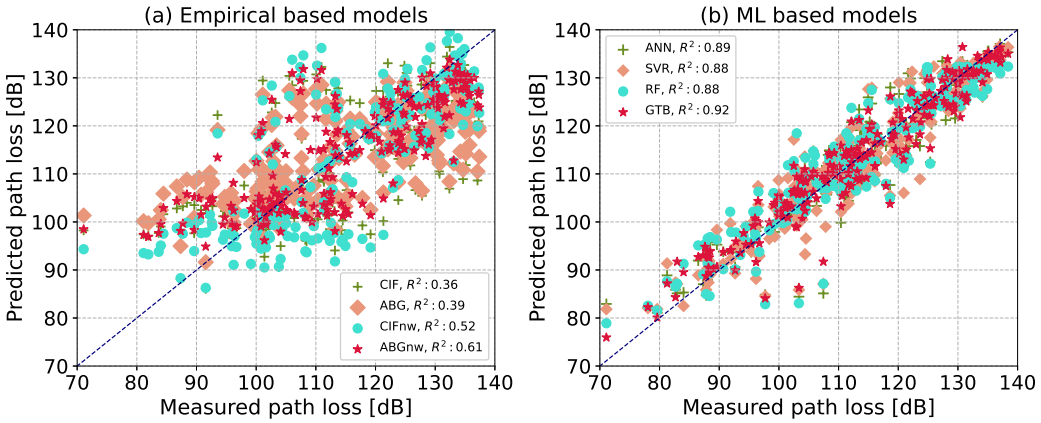

Figure 3.2: Comparison of measured and predicted path losses on the testing set for the empirical and ML models.

3.6.3 Computational Complexity in the ML Models

The training time complexity of an ML algorithm depends on the number of examples used for training, the data dimension, and the number of parameters learned. In the following, we employ simple approximate estimates for the number of operations in naive implementations of the algorithms, for comparison purposes. Consequently, the training time complexity of SVR can be assumed to be $\mathcal{O}(I^2P)$, where I is the dataset cardinality and P the dimension of the input vectors [103], although, the use of a kernel function and computational tricks may reduce it. Similarly, for a decision tree, the training complexity is $\mathcal{O}(I\log(I)P)$ [141]; consequently, the RF and the GTB present a training complexity of $\mathcal{O}(I\log(I)PQ)$, where Q is the number of decision trees employed.

For the ANN, since the learning algorithm is iterative, besides the ANN structure, we have to account for the number of epochs (passes through the dataset), and the training complexity is roughly $\mathcal{O}(EIPN)$, where E is the number of epochs and N the number of neurons in the network [93]. Meanwhile, the run-time complexity corresponds to the number of computations required to obtain the output from the input, thus, it depends on the number of parameters composing the machine and how they are used to produce the output. Again, we consider estimates for the run-time complexity for the purpose of simple comparison. The run-time complexity of the SVR $\mathcal{O}(V)$ depends on the number of support vectors V. For the decision tree, the run-time complexity $\mathcal{O}(J)$ depends on the maximum depth of the tree J; thus, for the RF and the GTB, the run-time complexity is in the order of $\mathcal{O}(JQ)$. The ANN run-time complexity $\mathcal{O}(W)$ depends on the number of weights W.

Table 3.5 lists complexity indicators of the final ML-based regression models: the required time to train the final models (training time), and the space complexity – the memory resources required to run the algorithm also presenting the number of parameters stored. We note that they agree with the discussion in the previous paragraph. The RF and the GTB require storing more information relative to their parameters but are more rapidly learned. On the other hand, the ANN and SVR require less storage but at the expense of higher training time. The experiments were carried out in Python with the Sklearn library [139] on a workstation with an Intel Core i7 8th Gen processor and memory RAM of 16 GB.

Table 3.5: Computational complexity in the ML models.

Model	Training time [s]	Memory storage [KB]	No. of parameters
ANN	13.54	8	444 (weights), 75 (biases)
SVR	13.97	36	681 Sup. Vectors, with 5 features seen during the fit.
RF	0.11	1,399	178 trees (max. deep: 6)
GTB	0.21	289	246 trees (max. deep:3)

In Section 3.8, we further analyze the complexity of the models regarding their input-output mappings as prescribed by interpretability and explainable ML viewpoints. Nevertheless, before that, we assess the performances of the different models we have already discussed.

3.7 Performance of the Empirical and the ML-based PL Models

Table 3.6 presents the performance indicators for the final models for both the training and testing sets. The indicators resulting during training are presented to assess possible over-fitting. The RMSE values for the empirical models for both the training and testing set in Table 3.6 closely match the ones returned by the CV method in Section 3.5. Meanwhile, for the ML models, there is a small gain in the final models compared to those obtained during model selection. Nevertheless, the accordance between the RMSE values for the final models for both the training set (slightly smaller) and the testing set (slightly greater) indicates good generalization performance and no over-fitting.

The ABGnw achieves the best performance among the empirical models, with RMSE, MAPE, σ, and R^2 equal to 9.11 dB, 6.52%, 5.70 dB, and 0.61,

Table 3.6: Performance evaluation of the empirical and ML models on both the training and testing sets.

Model	RMSE [dB]		MAPE [%]		σ [dB]		R^2	
	Train	Test	Train	Test	Train	Test	Train	Test
ABG	11.30	11.40	8.35	8.41	6.54	6.52	0.38	0.39
CIF	11.73	11.67	8.51	8.42	6.88	6.86	0.34	0.36
ABGnw	**9.34**	**9.11**	**6.58**	**6.52**	**5.71**	**5.70**	**0.58**	**0.61**
CIFnw	10.28	10.13	7.18	6.98	6.51	6.58	0.49	0.52
ANN	3.82	4.85	2.65	3.27	2.42	3.25	0.93	0.89
SVR	4.90	4.96	3.21	3.43	3.32	3.22	0.88	0.88
RF	4.53	5.21	3.13	3.62	2.85	3.38	0.90	0.88
GTB	**3.45**	**4.28**	**2.34**	**3.00**	**2.25**	**2.76**	**0.94**	**0.92**

respectively. The CIFnw follows it, then follows the ABG, and at last, the CIF. The inclusion of the number of wall produces accuracy gains for the mmWave propagation indoor scenario, as previously indicated. Regarding the ML regression for PL, the GTB presents the best performance – the lowest RMSE (4.28 dB), the smallest MAPE (3.00%), the lowest σ (2.76 dB), and the greatest R^2 (0.92). Meanwhile, the ANN reaches RMSE, MAPE, σ, and R^2 equal to 4.85 dB, 3.27%, 3.25 dB and 0.89, respectively. They are followed by the SVR (RMSE, MAPE, σ, and R^2 equal to 4.96 dB, 3.43%, 3.22 dB and 0.88, respectively).

3.8 Results: Interpretable Machine Learning Techniques used for Predictors Selection

The previous section has shown that ML models may provide accurate PL prediction for multi-frequency mmWave indoor environments. However, it is still unclear why the ML models improve over empirical models. Considering that, in this section, we investigate how the different predictors affect the responses of the ML-based PL models. Comprehending that, one may devise the best-suited coalition of predictors for the different models.

3.8.1 Mutual Information

We compute MI for predictors pairs and between each predictor and the outcome in the training set. To asses the MI, we use the histogram method [142]. To avoid bias, we use 20-bins histograms for all the predictors [143] presenting continuous values and for the outcome. However, we em-

ploy only three bins for the categorical numbers-of-walls predictor. Figure 3.3 shows the results and the scatter plots for the visualization of the relations between predictors and the path loss and between predictors.

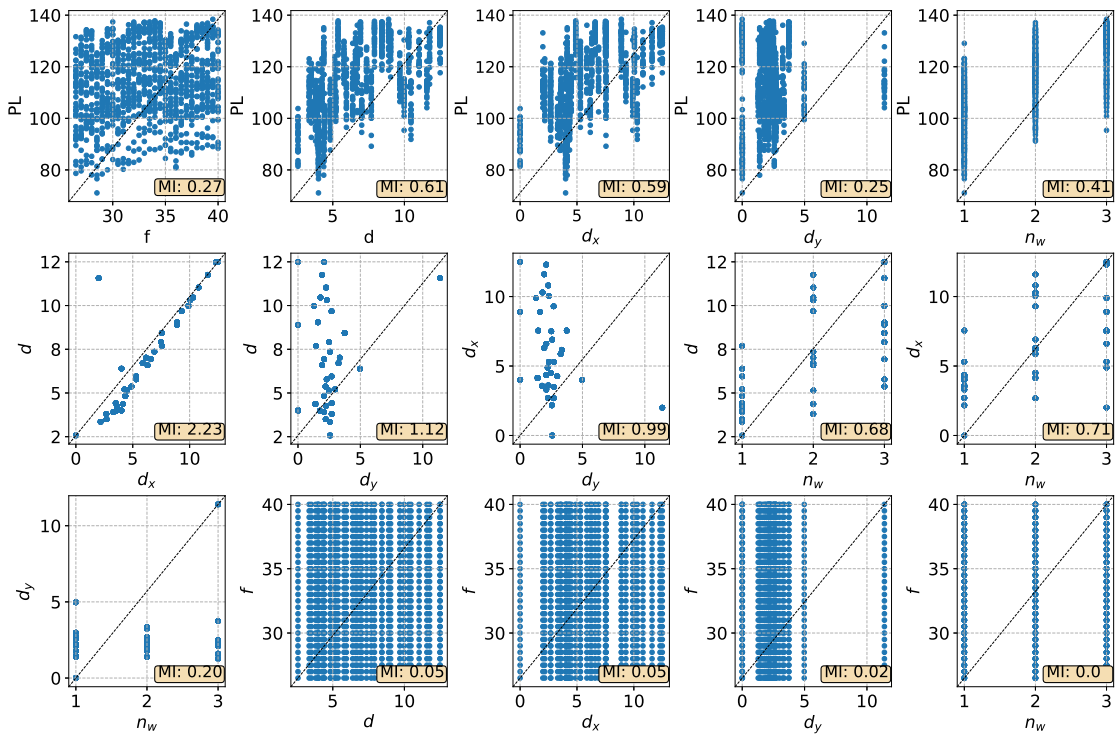

Figure 3.3: Scatter plots between predictors and the path loss and between predictors pairs. Each graph presents the MI value for the pair of variables evaluated.

The predictor d attains the largest MI with the path loss followed by d_x (both are very similar). Figure 3.3 shows a low MI between path loss and f (0.27); this fact may be explained by the presence of strong multipath effects diminishing the path loss dependence with frequency. However, we observe a higher dependency between the path loss and the predictor n_w (0.41). As expected, the higher the number of walls, the higher the path loss. Furthermore, the predictor d_y presents the lowest MI with the path loss. Between predictors pairs, the pairs d–d_x, d–d_y, and d_x–d_y provide the highest MI values; this is expected given their mathematical relations. The MI reduces for the pairs d–n_w and d_x–n_w (and both are similar). When the pair contains f, low MI consistently occurs.

3.9 Model Interpretation Methodology and Results

Our predictor selection methodology examines whether fewer predictors grant more interpretability without compromising prediction performance. We apply the IML tools to the training set for the optimized hyperparameter models described in Subsection 3.6.1.

3.9.1 Methodology

We design machines that employ subsets of predictors (coalitions) S of increased cardinality (but not every possible subset of predictors) for each model. The order in which a new predictor is incorporated employs a forward strategy. The predictors are included in the subset according to their performances for the one-predictor ML regression. We rank the predictors according to the machines' performances – measured using the RMSE, MAPE, σ, and R^2. The rank is then used to train models using coalitions containing more predictors until encompassing all predictors. Each subsequent coalition is the joining of a new predictor with the previous coalition.

3.9.2 Indicator Values

The resulting ranks of predictors for the four models are presented in the second column of Table 3.7. The rows listing the coalitions are obtained using the ranking described in Subsection 3.9.1. Table 3.7 also presents the resulting machines' performance (RMSE, MAPE, σ, and R^2) for the different coalitions. In addition, the pre interpretation of the model uses the JMI and CMI. The post-hoc interpretation uses IAS and $\overline{\text{MEC}}$.

Table 3.7 also presents the JMI and the CMI for the coalitions and the response. They are computed using the k-nearest neighbors algorithm [144]. Since each coalition may present a different distribution for the data, we find the most suitable value of k for computing JMI and CMI for each subset. In [123], the authors compute the MI from the mean of the results for a range of $k = 6, \ldots, 20$. In this study, we compute JMI and CMI for $k = 3, \ldots, 71$. One returns the value of k for which the changes are smaller than 5% for the following 25 values. This methodology allows us to find the value of k in a region where the change in JMI and CMI is small, meaning that they are reasonably stable.

Table 3.7 also presents the IML indicators IAS and $\overline{\text{MEC}}$. Nevertheless, IAS and $\overline{\text{MEC}}$ can be better examined and interpreted if one considers the

Table 3.7: Performance, joint and conditioned mutual information and IML indicators for the machines constructed for the four ML models using different subsets evaluated on the training set. Model performance is measured using RMSE, MAPE, σ, and R^2.

Model	Predictors coalitions	RMSE [dB]	MAPE [%]	σ [dB]	R^2	JMI	CMI	IAS	$\overline{\text{MEC}}$
ANN	d_x	9.18	6.51	5.72	0.59	-	-	-	6.00
	d_x,d_y	6.90	4.75	4.52	0.77	0.31 (k=29)	0.03 (k=44)	1.53	2.58
	d_x,d_y,f	5.47	3.40	3.98	0.85	0.24 (k=30)	0.02 (k=47)	1.18	2.75
	d_x,d_y,f,n_w	3.97	2.76	2.52	0.92	0.23 (k=31)	0.02 (k=10)	1.10	3.61
	d_x,d_y,f,n_w,d	3.82	2.65	2.42	0.89	0.23 (k=31)	0.01 (k=60)	1.17	2.20
SVR	n_w	9.70	7.01	6.08	0.54	-	-	-	1
	n_w,d_y	8.01	5.65	5.07	0.69	0.25 (k=33)	0.04 (k=45)	1.49	2.37
	n_w,d_y,d_x	6.34	4.51	3.91	0.80	0.34 (k=29)	0.20 (k=36)	2.13	2.95
	n_w,d_y,d_x,f	5.13	3.36	3.47	0.87	0.23 (k=31)	0.02 (k=47)	2.00	2.97
	n_w,d_y,d_x,f,d	4.90	3.21	3.32	0.88	0.23 (k=31)	0.01 (k=60)	2.22	1.99
RF	d	7.10	4.88	4.65	0.75	-	-	-	10
	d,n_w	6.22	4.45	3.78	0.81	0.33 (k=30)	0.02 (k=46)	0.44	3.90
	d,n_w,f	5.11	3.46	3.37	0.87	0.19 (k=33)	0.01 (k=50)	0.38	3.56
	d,n_w,f,d_y	4.59	3.18	2.88	0.90	0.20 (k=34)	0.01 (k=55)	0.92	3.42
	d,n_w,f,d_y,d_x	4.53	3.13	2.85	0.90	0.23 (k=31)	0.01 (k=43)	0.48	3.03
GTB	d	6.89	4.74	4.51	0.77	-	-	-	10
	d,f	5.21	3.15	3.90	0.87	0.21 (k=28)	0.02 (k=47)	0.23	9.90
	d,f,n_w	3.58	2.44	2.33	0.94	0.19 (k=33)	0.03 (k=10)	0.21	4.69
	d,f,n_w,d_y	3.49	2.34	2.30	0.94	0.20 (k=34)	0.01 (k=55)	0.51	3.84
	d,f,n_w,d_y,d_x	3.45	2.34	2.25	0.94	0.23 (k=31)	0.01 (k=43)	0.26	3.40

PFI, main effects, and interaction effects jointly. Table 3.8 shows the PFI rank for every model and coalition. We order the predictors vertically for increasing PFI for each model and coalition. The PFI is computed using Equation (2-22) for $D = 10$.

Besides, to visualize the main effect of each predictor on the path loss, we observe the ALE curves for the different coalitions using each model in Figure 3.4. In the figure, the ALE curves in each column correspond to a model, and each row indexes a different predictor. The first row presents the graphs for the best-ranked variable in each of the five predictor subsets. The second row considers the second-ranked predictor, and so on. The abscissa presents the predictor range in each graph, and the ordinate presents the centered ALE (the horizontal dashed line at zero reflects the average main effect). The different curves in each graph present the ALE curves for the correspondent predictor and model when it is included in the coalition following the order of predictors

Table 3.8: PFI rank of the selected predictor shown in crescent order, with RMSE loss for different subsets of predictors.

Number of predictors	Predictors ranked by the PFI (loss: RMSE [dB]) for the different models and coalitions.							
	ANN		SVR		RF		GTB	
2	d_y	9.68	d_y	3.21	d	5.56	f	2.92
	d_x	12.12	n_w	10.94	n_w	10.24	d	13.10
3	f	2.52	d_y	21.93	f	2.18	f	3.57
	d_y	10.25	n_w	27.82	d	6.10	d	8.21
	d_x	14.49	d_x	31.01	n_w	10.67	n_w	10.09
4	f	3.31	f	2.18	f	2.25	f	3.66
	n_w	15.21	d_y	22.51	d	2.27	d	4.20
	d_y	16.59	n_w	29.69	d_y	5.00	d_y	5.05
	d_x	19.78	d_x	34.31	n_w	12.62	n_w	10.87
5	f	3.26	f	2.36	d_x	1.02	d	1.96
	n_w	16.83	d_y	35.43	d	1.32	d_x	2.49
	d_y	21.81	n_w	36.05	f	2.35	f	3.65
	d_x	33.16	d	108.97	d_y	5.10	d_y	4.40
	d	34.39	d_x	110.90	n_w	12.70	n_w	10.86

in the legends. We use at most 30 intervals to divide the predictor range for computing the ALE effects.

A positive ALE in a given predictor interval means an increased effect on the response relative to the average main effect. Negative values denote a decreasing effect relative to the average main effect. The higher the ALE is in a given range, the more informative the predictor is in that range. Thus, if in a given range (x-axis) for a predictor the ALE curve deviates markedly from the horizontal line, then within this range, the predictor significantly influences the model response. One also sees that the predictor's ALE curves vary upon the predictors in the coalition, reflecting the interaction between predictors.

3.9.3 Analysis of the Results

Since the GTB-based PL regression shows the best performance, as readily seen in Table 3.7, we discuss its results first. Complementary to the PFI and ALE effects results for every subset, we also examine the interaction strength between two predictors (2D-ALE) for the GTB-based PL model shown in Table 3.9. The predictor pairs are arranged in increasing order of interaction for every subset of predictors.

In Table 3.7, we note that the more predictors employed, the better the performance irrespective of the considered indicator. Although, for more than three predictors, the performance improvement is meager (we are analyzing

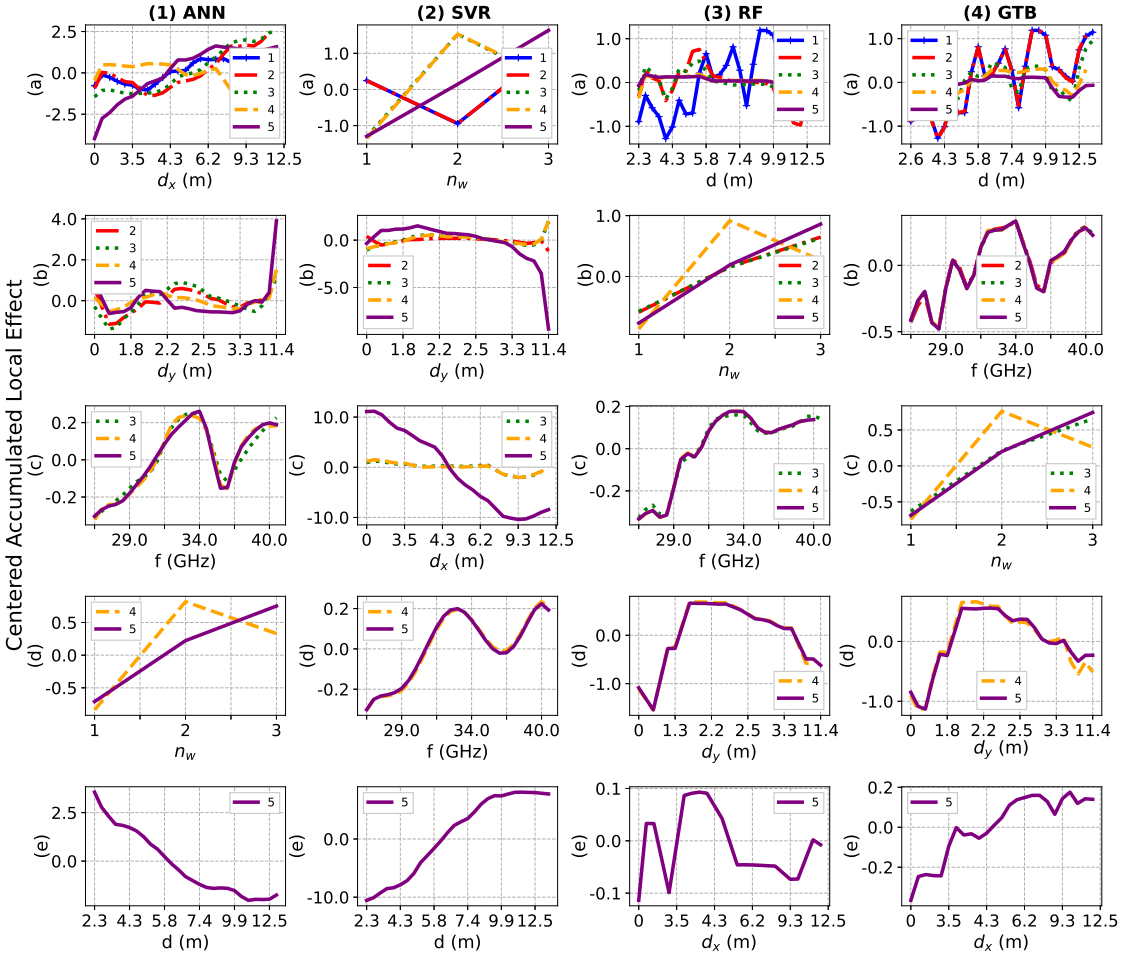

Figure 3.4: ALE graphs for the four ML models. Each line presents the graphs for a different model: (1) ANN, (2) SVR, (3) RF, and (4) GTB. The predictor order in each column corresponds to their inclusions in the model input from upper to lower. Each plot presents the ALE graphs for the predictor, considering the different number of predictors used as input.

Table 3.9: Interaction strength between two predictors (2D-ALE) for the GTB-based PL model, for each number of predictors used for machine design, the first row presents the pair having the lowest interaction, and the pairs follow in increasing interaction order.

Interaction between two predictors (2D-ALE)							
Subset of predictors							
2		3		4		5	
d,f	0.110	f,n_w	0.009	d,d_y	0.004	f,n_w	0.011
		d,f	0.114	f,n_w	0.009	n_w,d_x	0.026
		d,n_w	0.181	f,d_y	0.040	d,d_x	0.030
				d,n_w	0.044	d,n_w	0.032
				d,f	0.080	f,d_y	0.039
				n_w,d_y	0.256	f,d_x	0.040
						d_y,d_x	0.047
						d,f	0.062
						d,d_y	0.166
						n_w,d_y	0.260

the GTB models' performance). The pre-interpretability indicators reported in Table 3.7 indicate that, for the three predictor subset, when the model performance significantly improves, the JMI value reduces. This reflects that larger JMI values do not fully correspond to obtaining machines of improved prediction performance.

Meanwhile, since we compute the CMI of adding a predictor to the previous coalition of predictors, the values in the CMI column reflect the information gained by adding a new predictor to a coalition. Considering the GTB regression, Table 3.7 reports the maximum CMI when incorporating n_w to the coalition containing d and f. Further, the CMI of d_y conditioned on n_w, f, and d is 0.01, lower than the previous two. Finally, when including d_x, the information gain is as irrelevant as the previous one. Although being model agnostic, one notes that the CMI relates to model performance since smaller CMI values correspond to smaller performance improvements as indicated by the RMSE, MAPE, σ, and R^2. This connection is observed for all models.

To quantify the contribution of the predictors we use post-hoc interpretation tools. We compute the marginal contributions to the RMSE, IAS, and the $\overline{\text{MEC}}$ when the coalition expands to include the predictor p,

$$\Delta\text{RMSE}_p = \text{RMSE}(S) - \text{RMSE}(S \cup \{x_p\}), \tag{3-11}$$

$$\Delta\text{IAS}_p = \text{IAS}(S) - \text{IAS}(S \cup \{x_p\}), \tag{3-12}$$

$$\Delta\overline{\text{MEC}}_p = \overline{\text{MEC}}(S) - \overline{\text{MEC}}(S \cup \{x_p\}). \tag{3-13}$$

The ΔRMSE_p is the difference in the RMSE for the coalition of the predictor p and the previous coalition (subset) of predictors. It evaluates the contribution of x_p to the model performance. If ΔRMSE_p is negative, then there is a loss when including x_p in the coalition (the RMSE reduces). In this case, the smaller the modulo of ΔRMSE_p, the smaller the performance improvement. A similar analysis applies to $\Delta\overline{\text{MEC}}_p$.

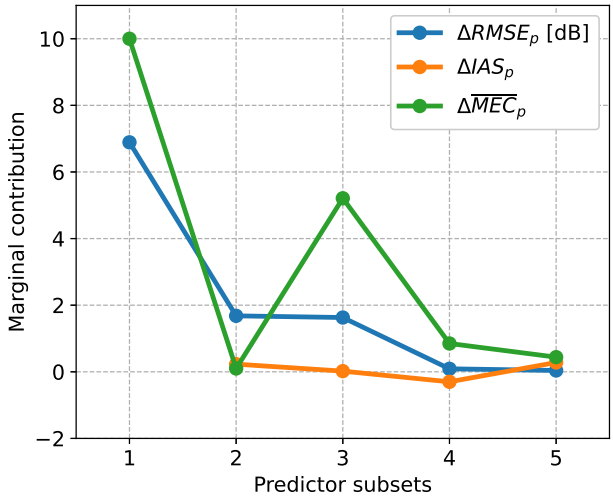

Figure 3.5: Marginal contributions in performance and interpretability for the GTB-based PL model (for the predictors coalitions see Table 3.7).

Figure 3.5 shows these marginal contributions for the GTB-based PL model. The largest marginal contribution in the RMSE and $\overline{\text{MEC}}$ occur when passing from the two to the three predictors (d, f, n_w) coalition; when adding d_y and d_x to the coalition, the marginal contributions are significantly smaller. Thus, the 3-predictors coalition offers a reasonable trade-off between performance and model complexity/interpretability. For this coalition, the values of IAS and $\overline{\text{MEC}}$ are 0.21 (the lowest among the coalitions tested for the GTB) and 4.69, respectively. Thus, when moving from the two for the three predictors coalition, the performance improvement derives from interactions between d and n_w as seen in Table 3.9, and also on the main effects of n_w and f on the response as seen in Figure 3.4.(4)(c). Notwithstanding, the PFI for f does not reflect that, although the PFI for n_w does, as seen in Table 3.8. For this coalition, the ALE curve for d presents less spread in all the range of d, suggesting that the main effect of d is smaller than in the previous coalitions.

A similar analysis for the 4-predictors coalition shows a complexity increase that is not reflected in the performance. The PFI and the ALE for f and n_w remain very similar (Table 3.8), as well as the ALE value of d, as seen in Figure 3.4.(4). The decrease in the marginal contribution of IAS as shown in

Figure 3.5 suggests that the marginal performance improvement is produced mainly by the interaction effects between n_w and d_y as seen in Table 3.9. The marginal contribution of $\overline{\text{MEC}}$ in Figure 3.5 slightly decreases (indicating that the approximation of the ALE curve for the new predictor requires as many linear segments as before) despite employing a new predictor. A similar analysis can be presented for the full-predictors coalition, though there is no significant improvement over the previous ones. Similar analyses apply to the other three ML models, which are presented in Appendix B.

3.9.4 Final ML Models

The results indicate that including d_x in the coalition does not significantly improve performance for the ANN, RF, and GTB models. Besides, the IML indicators are very similar for the four (f, d, d_y, n_w) and five-predictor coalitions, irrespective of the model. For the subset of three predictors, the performances of the ANN, SVR, and RF are significantly degraded with respect to the four predictors coalitions, but not for GTB regression, which already presents a good performance for the three-predictors coalition. Consequently, for the GTB regression, the three-predictors coalition (f, d, n_w) is selected. Meanwhile, for the ANN and the RF, we select four-predictors coalitions (f, d, d_y, n_w). At last, the five-predictor coalition (f, d, d_x,d_y, n_w) is chosen for the SVR. Using these coalitions, we reapply the grid search using a 5-fold CV to find the optimal hyperparameters for each model and, subsequently, train them. Table 3.10 shows the final models, their interpretability indicators (computed on the training set), and their performances.

Table 3.10: Performances and complexities of the four final regression having optimized predictors subsets and hyperparameters.

Model				
ML alg.	ANN	SVR	RF	GTB
Coalition	d_x,d_y,f,n_w	n_w,d_y,d_x,f,d	d,n_w,f,d_y	d,f,n_w
Hyperparameters	neurons=70, η=0.1, α_{wd}=0.1	C=2100, ε=0.1,σ_{RBF}=0.2	trees=178, max.depth=6, min. samples leaf=1	η=0.1, trees=246, max.depth=3, min. samples leaf=3
IAS	2.15	2.22	0.92	**0.21**
$\overline{\text{MEC}}$	2.76	1.99	3.42	**4.69**
Performance (Training set | Testing set)				
RMSE [dB]	3.77 | 4.82	4.90 | 4.96	4.59 | 5.24	**3.58** | **4.29**
MAPE [%]	2.62 | 3.23	3.21 | 3.43	3.18 | 3.66	**2.44** | **3.03**
σ [dB]	2.38 | 3.26	3.32 | 3.22	2.88 | 3.38	**2.33** | **2.75**
R^2	0.93 | 0.89	0.89 | 0.89	0.90 | 0.87	**0.94** | **0.92**

As seen in Table 3.10, the GTB presents the best performance – the lowest RMSE (4.29 dB), the smallest MAPE (3.03%), the lowest σ (2.75 dB), and the greatest R^2 (0.92), together with the lowest interaction among predictors (IAS equal to 0.21) and the highest complexity ($\overline{\text{MEC}}$ equal to 4.69). In terms of performance, the GTB is followed by the ANN, the SVR, and, at last, comes the RF regression. The IML indicators are ranked almost inversely, although the SVR presents the greatest IAS and the smallest complexity. To better assess the predictors' influence on the path loss, Figure 3.6 displays the absolute error between the measured and predicted path loss versus distance and the number of traversed walls for the ABGnw and GTB models, evaluated on the testing set for the single frequencies of 27 GHz, 33 GHz, and 40 GHz. The ML model shows its ability to follow the data distribution with respect to distance, number of traversed walls, and frequency.

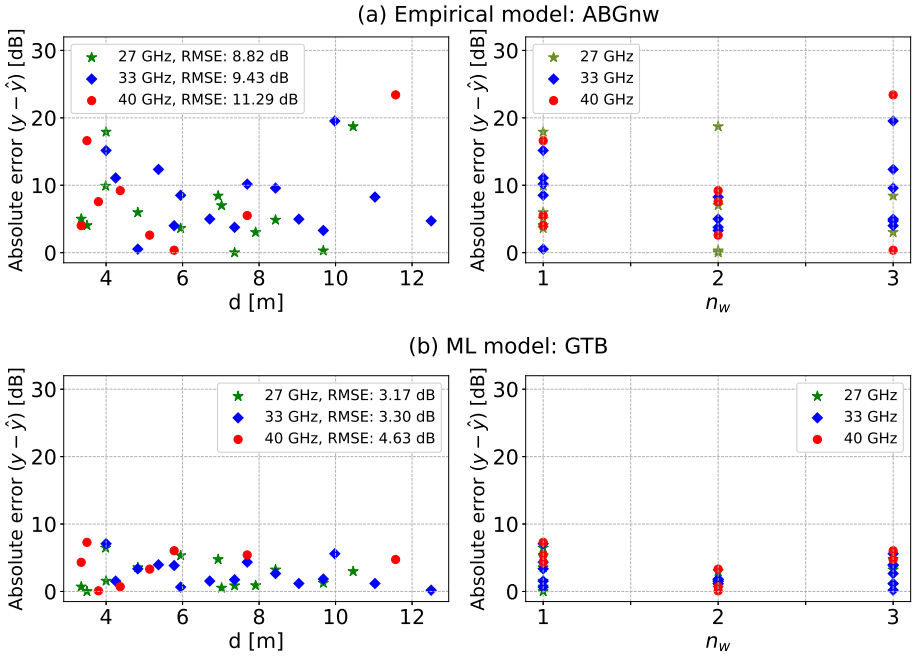

Figure 3.6: Absolute error between the measured and predicted PL on the testing against distance and number of obstructing walls for the frequencies at 27, 33, and 40 GHz.

From the obtained models, we observed for the ANN and SVR that the horizontal and vertical components of the distance (d_x and d_y) contributed to improving the model's performance than simply using the predictor d. This can be attributed to the geometric characteristics of the indoor environment which presents an elongated rectangle implying an asymmetry.

3.10 Generalization Capacity Analysis

We further evaluate the generalization capability of the ML-based PL models. Since the GTB model obtains the best performance with the coalition of three predictors, as seen in Table 3.10, one examines its generalization capability considering that the transmitter point is unknown. We train the GTB model without any instance of the transmitter and, consequently, the correspondent reception points. This best fits a real-world scenario where the transmitter and receiver locations are unknown and the objective is to use the ML-based path loss model to predict the coverage area of potential transmitter locations for which no or limited measured data are available [46].

To this end, first, we partition the database to train it using PL instances from five transmitter points and test it with the instances from the segregated transmitter point. The results are presented in Table 3.11. The first row and column indicate the point of the instances in the testing Tx point (we train the model using the instances associated only with the other five Tx points). The second and third columns show the number of instances in the train and test set resulting from the above-discussed partition of the dataset. The following two columns present the RMSE during training and testing. We note that for each test, we optimize hyperparameters: learning rate (η), trees, maximum depth (md), and minimum samples leaf (msl), which optimal values are listed in the last column in Table 3.11. The results present performance decreases (RMSE increases) in comparison with Table 3.10. However, the performance remains good. Besides, the accordance between the RMSE values for the training set and the testing set (the last is slightly greater) indicates no over-fitting. The highest RMSE (5.62 dB) occurs when Tx 5 is unknown, and the lowest (4.77 dB) when Tx 2 is.

Table 3.11: Performance of the proposed GTB-based PL model when the models are training considering a set of transmitters and the test employs a different one.

Tx-Test	No. samples		RMSE [dB]		Hyperparameters
	Train	Test	Train	Test	
1	812	308	4.63	5.46	η=0.01,trees=161,md=6,msl=2
2	952	168	4.35	4.77	η=0.04,trees=246,md=3,msl=2
3	896	224	4.90	5.20	η=0.07,trees=246,md=2,msl=2
4	924	196	4.13	4.86	η=0.04,trees=246,md=3,msl=2
5	952	168	4.55	5.62	η=0.10,trees=246,md=2,msl=2
6	1064	56	4.60	5.01	η=0.05,trees=246,md=3,msl=3

We move to a harder scenario when fewer training samples are available. We use the measurements collected using only two transmitters for training and the remaining four transmitters for testing. The results are presented in Table 3.12. The results show that the GTB-based PL model still performs well with fewer training samples, for example, using transmitters 2 and 6 (224 samples) for training, the testing RMSE is still reasonably low, 6.15 dB, since it is not far from the previous case with one unknown transmitter location.

Table 3.12: Generalization capability analysis for the GTB-based PL model. The models are trained using two transmitters and evaluated (tested) on the four remaining ones.

Tx's (No. samples)		RMSE [dB]		Hyperparameters
Train	Test	Train	Test	
1,3 (532)	2,4,5,6 (588)	4.19	6.74	η=0.1,trees=246,md=2,msl=3
1,4 (504)	2,3,5,6 (616)	3.64	5.84	η=0.1,trees=140,md=2,msl=10
2,4 (364)	1,3,5,6 (756)	4.23	5.94	η=0.1,trees=40,md=3,msl=10
2,6 (224)	1,3,4,5 (896)	4.14	6.15	η=0.1,trees=30,md=3,msl=10

In order to further assess the generalization capabilities of the proposed GTB PL model, we apply the design methodology to the indoor database furnished by the Yonsei University, Korea [145, 146]. This database is constructed using mixed measurements and Ray-tracing models at 28 GHz, a bandwidth of 100 MHz, omnidirectional antennas, transmission power of 20 dBm, and a receiver threshold equal to -100 dBm [146]. It reports the PL in dB for various floors inside Yonsei University's building. (http://www.cbchae.org/). We extract the data corresponding to floors 2, 3, and 4. On each floor, there are 90 instances accounting for ten transmission locations and nine reception points with distances ranging between 20 to 80.46 meters. There are up to nine traversed walls on floors 2 and 4 and up to ten on floor 3. The PL spans between 76.85 dB to 158.70 dB on the second floor, 76.73 dB to 135.89 dB on the third floor, and 77.15 dB to 144.38 dB on the fourth floor.

Due to the greater Tx-Rx distances and spanned number of traversed walls for the Yonsei database, it is not possible to apply directly the trained models in Table 3.10. However, we can verify if the proposed methodology performs well by evaluating if the GTB-based PL model using the selected predictors (f, d, n_w) provides accurate responses for this different scenario. The small number of training samples in the Yonsei database may hamper learning good models. Therefore, we optimize and train (with the learning rate set to 0.1) the models using one floor and the remaining two floors are

used for testing, as presented in Table 3.13. For the GTB-based PL model, similar RMSE results are obtained for training and testing when any floor is used to build the model with optimized hyperparameters. Furthermore, the RMSE for the testing sets present similar values to those presented in Table 3.12 when only two transmitters are used to train the models at the PUC-Rio, CETUC environment. The RMSE values for the training and testing for the ABGnw and CIFnw are higher than the ones returned by the GTB-based PL model.

Table 3.13: GTB and empirical PL models performance using the database from Yonsei University, Korea.

Floor	Set	Samples	RMSE [dB]			Hyperparameters
			ABGnw	CIFnw	GTB	
2	**Train**	90	9.98	9.97	**5.96**	
3	Test	90	11.42	11.36	**6.43**	trees=150, md=2,msl=3
4	Test	90	10.49	10.43	**6.91**	
3	**Train**	90	7.80	7.29	**5.12**	
2	Test	90	13.04	13.22	**6.55**	trees=200,md=2,msl=4
4	Test	90	10.34	9.96	**5.52**	
4	**Train**	90	9.54	9.12	**5.75**	
2	Test	90	10.91	11.17	**6.64**	trees=200,md=4,msl=2
3	Test	90	8.76	8.32	**6.01**	

In conclusion, since the proposed GTB-based path loss model provides accurate responses from few measurements, it can be used for practical PL prediction for mm-wave links in medium-size buildings.

3.11 Discussion

This chapter has investigated mmWave indoor PL modeling. We obtained different machine learning- and empirical-based models for the PL of mmWaves in an indoor environment. They are based on a measurement campaign considering a wide frequency band spanning from 26.5 GHz to 40 GHz. We proposed extensions of two empirical models, ABGnw and CIFnw, that incorporate the number of traversed walls and improve the accuracy of the PL prediction over the original ABG and CIF. The machine learning regression PL models significantly improved the performance over empirical models. Notwithstanding, we observed that the four ML regression models are much more accurate than the empirical ones. The GTB led to the best PL prediction among the ML models.

We proposed a methodology to approach the selection of the coalition of predictors for ML regression, aiming at model interpretability. The results show that the presented methodology can select a good coalition of predictors with interpretability gains without compromising the predictive performance. In addition, by analyzing the predictors belonging to the coalitions, we observed that the pre-interpretation tool known as conditional mutual information might be helpful in detecting relevant predictors before training when the data dimension is low.

The GTB regression attains the best performance, followed by the ANN, the SVR, and the RF. A possible explanation for the best performance of the GTB comes from the post-hoc interpretation tools. They indicate that this regressing model relies most heavily on the main effects of each predictor in the coalition. We also saw that the resulting GTB-based PL model can accurately predict the PL at unforeseen transmission and reception locations. Therefore, the model is of practical use for mmWave PL prediction.

Using the methodology described in this chapter, in the following chapter, PL prediction in mmWave outdoor links is tackled using a measurement campaign performed at the university campus of PUC-Rio. Besides, two extended empirical models are also proposed considering system- and environment-parameters predictors and compared with the ML models, as presented for the path loss for mmWave indoors in this chapter.

4
Path Loss Prediction for mmWave Outdoor Communications using Machine Learning Techniques

The outdoor propagation of mmWaves in non-line-of-sight (NLOS) conditions presents very high losses due to blockage by buildings, vegetation, and absorption by atmospheric gases. In addition, the multipath propagation in NLOS conditions is less effective due to lower diffraction effects. Thus, line-of-sight (LOS) is required in short-range links [23, 34, 147]. Propagation through vegetation suffers losses that increase proportionally with the logarithm of the frequency [23]. This severely impacts the mmWave link with attenuations around 6 dB/m at 28 GHz and 11 dB/m at 60 GHz [23].

This chapter addresses PL prediction for short-range links in an outdoor environment at frequencies between 27 GHz and 40 GHz. We design several ML models including ANN, SVR, RF, and GTB. The model's inputs regard the influence of the frequency, distance, height difference between the transmitter and receiver, and the link vegetation depth. The predictor coalition selection technique described in the previous chapter is applied searching to the ML model with the best performance; thus, using the post-hoc interpretation tools, we examine the interpretability for each predictor subset.

4.1
Related Work

The multi-frequency ABG and CIF models are also commonly employed for outdoor environments, while the AB and CI models are used for single-frequency scenarios [148, 149]. The work in [148] studied the effect of the rain attenuation in path loss prediction for short-range millimeter wave link at 38 GHz on a path length of 300 m using an extended version of the CI model. The proposed model obtained an adjusted R^2 value of 0.85. In [149], the authors employed the ABG model using measurement campaign conducted in a parking environment at 28 GHz and 38 GHz, and compared their results using a extension of the log-distance model that incorporates the effects of presence of cars around the transmitter and receiver by considering link distances from 14 to 30 m. In the experiments, the ABG model achieved an average σ value of 2.4 dB and the proposed model achieved an average σ value of 3.25 dB.

Concerning machine learning techniques, in [32], an ML-based PL model is proposed at 28 GHz for a measurements campaign in Manhattan City. The predictors are extracted from the LiDAR point cloud, and a deep learning model is proposed for PL prediction. The authors in [25] propose a deep learning algorithm to predict the path loss exponent at 28 GHz in an outdoor environment. Terrain data and the shape and height of the buildings were used as input images, and ray-tracing simulations were employed to generate the data. In [26], a deep learning method is proposed for large-scale fading prediction for mmWave using the environment topographical predictors as inputs.

We propose two extended empirical models based on the ABG and CIF using the vegetation profile and heigth difference between the transmitter and receiver as predictors for comparison with the ML models. We also compare the results of the obtained models with the study proposed in [150] based on the Fuzzy technique derived from the measurement campaign employed in this chapter. Finally, we analyze the generalization capacity of the model to assess the performance of the ML and empirical models. Given that the measurement campaign used only one transmitter location with multiple receiver positions, we evaluate the generalization capability of the models considering unknown receiver points.

4.2 The Measurement Campaign: Dataset Description

We consider measurements performed at frequencies between 27 GHz to 40 GHz in steps of 1 GHz at the university campus of PUC-Rio. The environment comprises two of the tallest buildings on the university campus, surrounded by abundant vegetation and other shorter buildings, as seen in Figure 4.1. The measurements were carried out in October 2018 and were conducted by colleagues from the radio propagation laboratory in CETUC [151]. The transmitter was located at the rooftop of a university building at a height of 50 m from the ground to simulate a typical micro-cell system, with the base station represented by the yellow star in Figure 4.1.

In the measurement campaign, 23 receiver positions were utilized, varying heights and distances to the transmitter. Tx-Rx distances range from 50 to 280 m, corresponding to points around the university campus where the transmitted wave was detectable. To capture the influence of height differences between the transmitter and receiver, different heights for the receiver antenna relative to the transmitter were considered. The height differences range from 15 to 53 m. The measurement campaign used a directional antenna at the

transmitter and the receiver, which are aligned with each other to maximize the measured power. The alignment process was facilitated by Bosch GRL 825 laser pointers.

Figure 4.1: Distribution of the transmitter and receivers map of the outdoor measurement scenario, PUC-Rio [150].

For this mmWave environment, the same system setup described in Chapter 3, Subsection 3.2.1 was employed. At each receiver position and frequency, measurements were taken three times and the results were averaged. The PL in dB is obtained from the measured received power and system parameters utilizing Equation (3-1).

Due to abundant vegetation in the university campus, we evaluate the influence of vegetation depth (v_{depth}) on PL prediction for mmWave short-range links. This predictor is determined based on the obstruction of the link caused by vegetation. To obtain the vegetation depth values, a satellite image was utilized to derive the vegetation profile between the transmitter and receiver positions using an algorithm to process the maximum height vegetation as proposed in [33]. Some receivers partially obstructed by foliage, such as RX1, RX2, RX4, RX10, RX15, and RX16 were identified. The profile vegetation of those receivers is shown in Figure 4.2. The data from two receiver positions (RX3 and RX13) were not considered due to their high path loss values arising from foliage fully obstructing the link to the transmitter. Thus, the dataset provides 294 PL instances. The points of vegetation intersection along the direct path are used to calculate the v_{depth} value.

Thus, each PL sample considered is associated with four numerical attributes: carrier frequency (f) in GHz, the Tx-Rx distance (d) in meters,

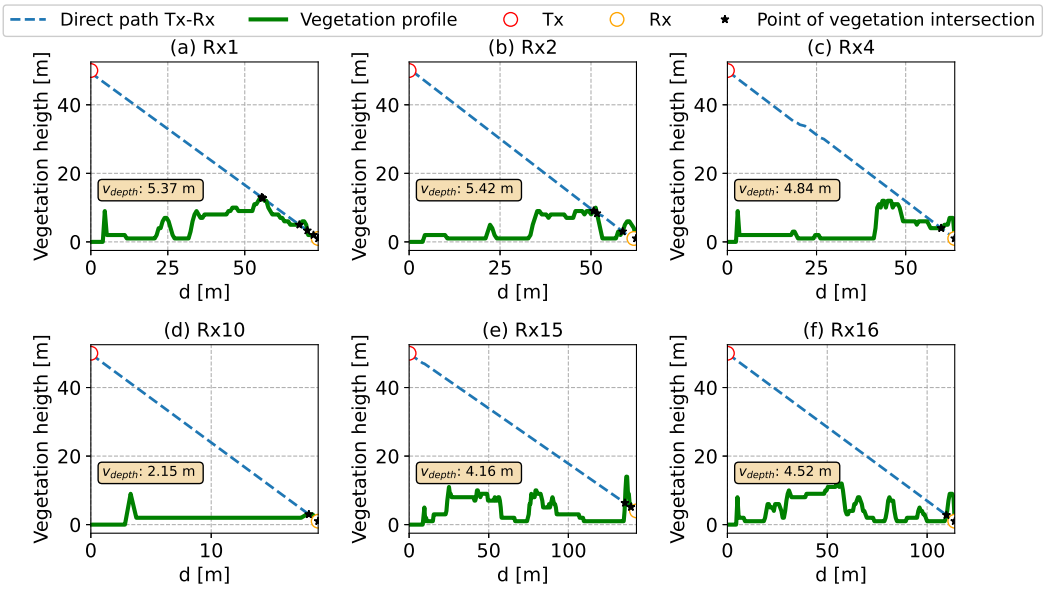

Figure 4.2: Vegetation profile for the receivers partially obstructed by foliage: (a) RX1, (b) RX2, (c) RX4, (d) RX10, (e) RX15, and (f) RX16. Each graph presents the v_{depth} value for the receiver evaluated.

the height difference between the transmitter and the receiver Δ_h in meters, and the vegetation depth (v_{depth}) in meters. Figure 4.3 shows the dependency between those predictors and the path loss and the correspondent MI value. The predictor d has the largest MI with the path loss (0.75), followed by Δ_h and f. The lowest MI is between v_{depth} and path loss (0.30). From these scatter plots and MI values, one sees that path loss tends to increase with frequency. In addition, as the distance between the transmitter and receiver increases, the signal strength decreases and, subsequently, the higher the path loss [34]. Also, the height difference between the transmitter and receiver can significantly impact the path loss; when there is a larger height difference, the signal encounters additional obstacles, such as buildings and other terrain features, which can obstruct or scatter the signal [20]. Finally, the vegetation depth can have an impact on path loss in the mmWave spectrum, since dense vegetation absorbs and scatters the signal, causing higher path loss [71].

4.3 Proposed Empirical Path Loss Model

To evaluate the influence of the predictors Δ_h and v_{depth} in the ABG and CIF models, two adjusted path loss models are also proposed. Their proposal aims to incorporate the effects of those predictors into the path loss and they

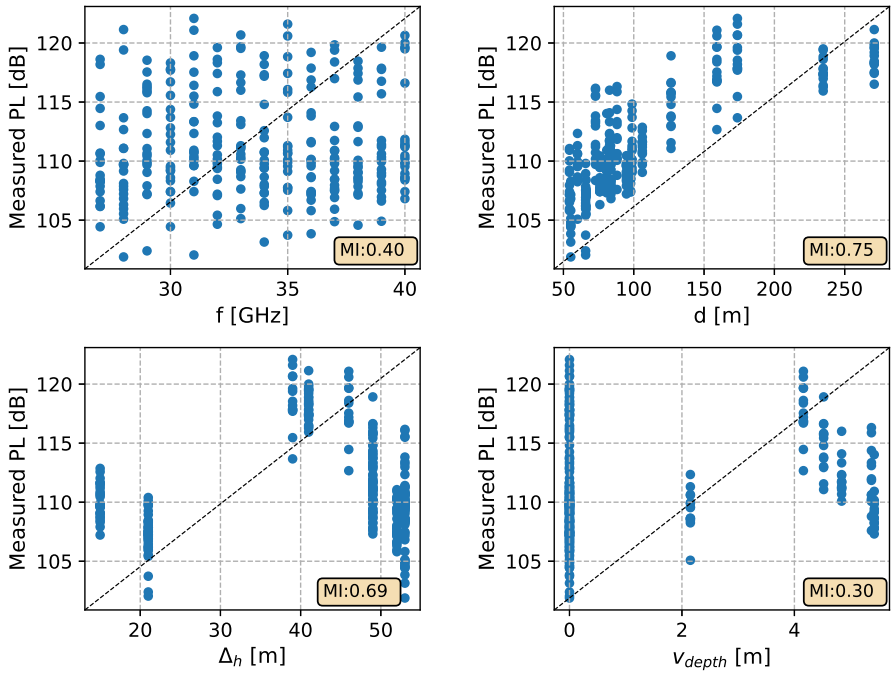

Figure 4.3: Scatter plots between predictors and the path loss for the mmWave outdoor environment.

are given by

$$\mathrm{PL}^{\mathrm{ABG}\Delta_h,\,v_{\mathrm{depth}}}[\mathrm{dB}] = 10\alpha\log_{10}(d) + \beta + 10\gamma\log_{10}(f) + \zeta\log_{10}\left(\frac{\Delta_h}{d}\right) + \varphi(v_{\mathrm{depth}}) + X_\sigma^{\mathrm{ABG}}, \tag{4-1}$$

$$\mathrm{PL}^{\mathrm{CIF}\Delta_h,\,v_{\mathrm{depth}}}[\mathrm{dB}] = \mathrm{FSPL}[\mathrm{dB}] + 10n\log_{10}(d)\left(1 + b\left(\frac{f - f_o}{f_o}\right)\right) + \zeta\log_{10}\left(\frac{\Delta_h}{d}\right) + \varphi(v_{\mathrm{depth}}) + X_\sigma^{\mathrm{CIF}}. \tag{4-2}$$

where ζ represents the PL coefficient characterizing the dependence on the relative height between the transmitter and receiver, and φ represents the mean vegetation attenuation.

4.3.1 Model Selection and PL Model Design

As in Chapter 3, 80% of the dataset is used for adjusting the coefficients of the empirical models and tuning the hyperparameters of the ML models. The 5-fold CV technique is applied to the training set to obtain the best hyperparameters in each ML model. The empirical model coefficients are also adjusted on each training set and evaluated on the remaining validation

set. The coefficient values for the ABG and ABG$\Delta_h, v_{\text{depth}}$, and CIF and CIF$\Delta_h, v_{\text{depth}}$ are presented in Table 4.1 and Table 4.2, respectively.

Table 4.1: Coefficients of the ABG and ABG$\Delta_h, v_{\text{depth}}$ models obtained from the CV subsets.

CV Subset	ABG			ABG$\Delta_h, v_{\text{depth}}$				
	α	β	γ	α	β	γ	ζ	φ
1	1.92	68.82	0.28	2.32	63.38	0.21	4.11	0.16
2	1.87	72.05	0.14	2.26	65.50	0.16	4.01	0.16
3	1.84	71.97	0.19	2.23	64.41	0.27	3.87	0.21
4	1.95	68.49	0.27	2.29	62.47	0.29	3.42	0.20
5	1.93	70.15	0.21	2.38	63.19	0.18	4.37	0.12

Table 4.2: Coefficients of the CIF and CIF$\Delta_h, v_{\text{depth}}$ models calculated from the CV subsets.

CV Subset	f_o	CIF		CIF$\Delta_h, v_{\text{depth}}$			
		n	b	n	b	ζ	φ
1	33.64	0.43	-0.79	0.53	-0.73	5.08	0.13
2	33.44	0.44	-0.91	0.54	-0.74	5.31	0.12
3	33.43	0.45	-0.98	0.54	-0.71	5.34	0.17
4	33.50	0.44	-0.80	0.52	-0.72	4.58	0.17
5	33.55	0.45	-0.88	0.55	-0.74	5.18	0.10

The performance evaluation of the empirical models, conducted on each subset, is summarized in Table 4.3. The ABG$\Delta_h, v_{\text{depth}}$ model attains the lowest average CV RMSE (2.18 dB). This model presents an improvement of 0.18 dB compared to the ABG model. On the other hand, the CIF model achieves the highest average CV RMSE (2.57 dB). However, when Δ_h and v_{depth} are included as predictors in the CIF model, its performance improves to 2.20 dB. Overall, the evaluation results indicate that the inclusion of Δ_h and v_{depth} leads to performance improvements for both the ABG and CIF empirical models.

Table 4.3: Performance evaluation of the empirical models with CV, values are in dB.

CV Subset	ABG	CIF	ABG$\Delta_h, v_{\text{depth}}$	CIF$\Delta_h, v_{\text{depth}}$
1	2.39	2.69	2.25	2.29
2	2.35	2.51	2.16	2.14
3	2.51	2.50	2.42	2.37
4	2.49	2.79	2.21	2.25
5	2.05	2.36	1.87	1.93
Average	**2.36**	**2.57**	**2.18**	**2.20**

4.4 ML-Based Models

We consider the four ML models described in Chapter 3, Subsection 3.6.1. The ML models' predictors and measured PL values are independently scaled using the mean and standard deviation.

4.4.1 Hyperparameter Tuning for the ML Models

The 5-fold CV technique is employed on the training set for each ML model to find the optimal hyperparameters. The ranges of the values of the hyperparameters are those described in Subsection 3.6.1. The ANN model achieves its lowest average CV RMSE (1.96 dB) by using 78 neurons in the hidden layer, with a learning rate and weight decay equal to 0.1 and 0.1, respectively, and the ReLU activation function. The SVR attains its lowest average CV RMSE (1.98 dB) for σ_{RBF}, C, and ϵ equal to 0.1, 200, and 0.05, respectively. The RF model attains its lowest average CV RMSE (1.77 dB) by utilizing 110 trees. The maximum depth of each tree is set to 6, and the minimum number of samples required to be at each leaf node is set to 3. Lastly, the GTB model attains its lowest average CV RMSE (1.61 dB) by employing 161 trees, maximum depth equal to 3, minimum samples in each leaf equal to 2, and learning rate of 0.1, considering the entire set of predictors.

4.5 Final Models Comparison

The final models are designed using 80% of the dataset to adjust and train the empirical and ML models, respectively. The remaining 20% of the dataset is reserved for testing the models. The adjusted coefficients of the empirical models are shown in Table 4.4. The ML models are trained using the hyperparameters reported in Subsection 4.4.1. Table 4.5 shows the final model comparison for both the training and testing sets. In the last row in Table 4.5, we include the results for the Fuzzy clustering model reported in [150] for comparison.

The results in Table 4.5 show that the performance of the ABG and CIF improve when the relative height between the Tx-Rx and the vegetation depth are included as seen in the average CV RMSE values in Subsection 4.3.1. Among the empirical models, the ABG$\Delta_h, v_{\text{depth}}$ achieves the best performance with RMSE, MAPE, σ, and R^2 equal to 2.52 dB, 1.63%, 1.64 dB, and 0.66, respectively. The CIF$\Delta_h, v_{\text{depth}}$ achieves a performance very close to the

Table 4.4: Adjusted coefficients of the empirical models.

	α	β	γ	ζ	φ
ABG	1.93	64.93	0.53	-	-
ABG$\Delta_h, v_{\text{depth}}$	2.27	58.94	0.55	3.43	0.27
	n	b	f_o	ζ	φ
CIF	0.44	-0.68	33.37	-	-
CIF$\Delta_h, v_{\text{depth}}$	0.52	-0.61	33.37	4.65	0.24

Table 4.5: Performance evaluation of the ML models on both the training and testing sets.

Model	RMSE [dB]		MAPE [%]		σ [dB]		R^2	
	Train	Test	Train	Test	Train	Test	Train	Test
ABG	2.29	2.69	1.64	1.87	1.38	1.67	0.71	0.61
CIF	2.49	2.91	1.75	2.01	1.54	1.81	0.66	0.54
ABG$\Delta_h, v_{\text{depth}}$	**2.08**	**2.52**	**1.35**	**1.63**	**1.34**	**1.64**	**0.76**	**0.66**
CIF$\Delta_h, v_{\text{depth}}$	2.11	2.54	1.48	1.71	1.31	1.64	0.76	0.65
ANN	1.23	2.00	0.83	1.34	0.82	1.31	0.92	0.78
SVR	1.52	1.98	0.93	1.29	1.12	1.35	0.87	0.79
RF	1.38	1.68	0.95	1.13	0.89	1.11	0.90	0.85
GTB	**0.94**	**1.39**	**0.64**	**0.94**	**0.62**	**0.90**	**0.95**	**0.90**
Fuzzy [150]	-	2.20	-		-		-	

ABG$\Delta_h, v_{\text{depth}}$ with RMSE, MAPE, σ, and R^2 equal to 2.54 dB, 1.71%, 1.64 dB, and 0.65, respectively.

The GTB presents the best performance among the proposed ML models, with RMSE, MAPE, σ, and R^2 equal to 1.39 dB, 0.94%, 0.90 dB, and 0.90, respectively. It is closely followed by the RF model, which attains RMSE, MAPE, σ, and R^2 equal to 1.68 dB, 1.13%, 1.11 dB, and 0.85, respectively. The ANN and SVR models present lower performances compared to the GTB and RF models. Figure 4.4 displays the measured and predicted path loss values for the empirical and ML models for comparison.

The work in [150] presents a Fuzzy technique based on the subtracting clustering algorithm to predict the PL. The approach uses data from the same measurement campaign described in this chapter. In that study, a number of samples from the dataset were dedicated to specifying the fuzzy logic, while the remaining samples were reserved for testing. However, the authors did not provide detailed information about the train-test split technique employed and the allocation of samples to each set. The Fuzzy model in [150] reported

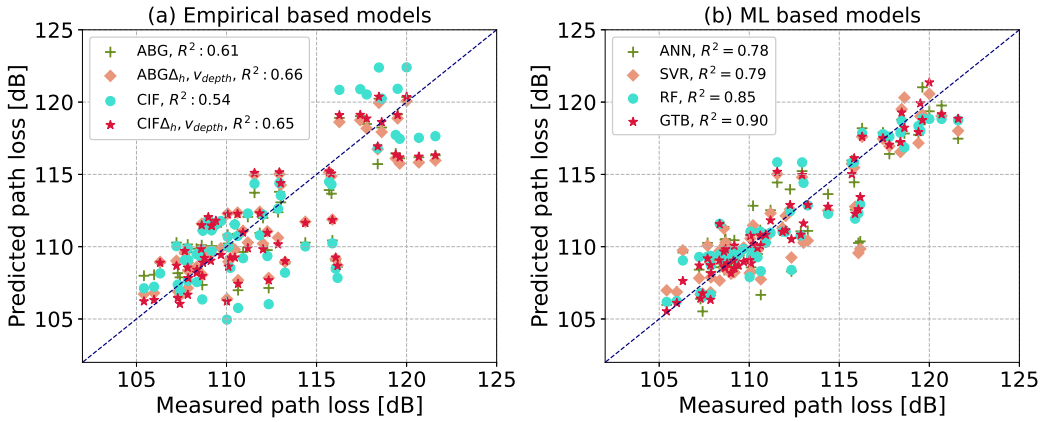

Figure 4.4: Comparison of measured and predicted path losses on the testing set for the ML and empirical models.

an RMSE of 2.20 dB for the PL prediction on the testing set for the multi-frequency case. The RMSE value is very similar compared to the proposed ABG$\Delta_h, v_{\text{depth}}$ and CIF$\Delta_h, v_{\text{depth}}$ models, with values of 2.52 dB and 2.54 dB, respectively. However, its performance is worst than that of all the proposed ML models, more significantly for the GTB model (1.39 dB).

We now employ the predictor selection technique described in Section 3.9 to investigate the relations between predictors and PL, specifically in the GTB model, which presented the best performance among the designed ML models. The performance evaluation results of each predictor coalition and their interpretability are shown in Table 4.6, while Figure 4.5 displays the ALE curves. From the results in Table 4.6, one sees that the prediction performance is significantly dependent on the predictors d and f; when adding more predictors to the coalition, the improvement is very small. In addition, the IAS and $\overline{\text{MEC}}$ values are similar for all the coalitions, this can be also seen in the ALE curves; when adding a new predictor, the centered accumulated local effects (y-axis) remain similar. Besides, the predictors d and f show the highest ALE values indicating a larger effect on the path loss prediction.

Table 4.6: Performance and IML indicators for the GTB model using different coalitions measured on the training set.

Predictors coalitions	RMSE [dB]	MAPE [%]	σ [dB]	R^2	IAS	$\overline{\text{MEC}}$
d	1.62	1.09	1.07	0.86	-	3.00
d,f	1.04	0.68	0.72	0.94	0.09	3.02
d,f, v_{depth}	0.99	0.66	0.68	0.95	0.11	3.02
d,f, v_{depth},Δ_h	0.94	0.64	0.62	0.95	0.11	3.02

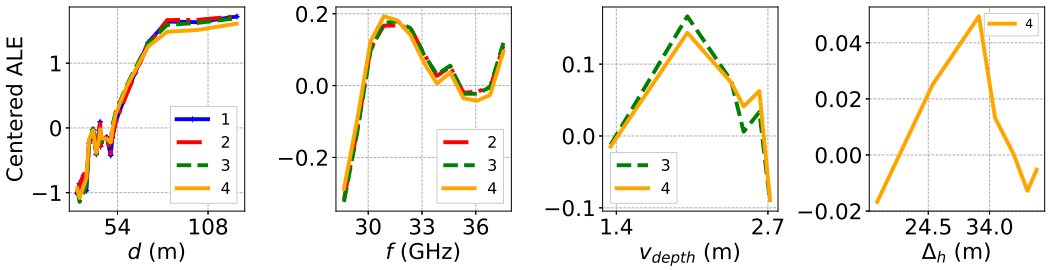

Figure 4.5: ALE plots for the GTB model evaluated from the subset of one predictor (d) until the four predictor subset (d, f, v_{depth}, Δ_h) in the mmWave outdoor environment.

For the subset of four predictors, the model relies mainly on the interaction between d and f as seen in Table 4.7. The low IAS values for all the subsets suggest that the model depends more heavily on the individual effects of each predictor than on interactions between pairs of them. In addition, the PFI technique is applied to the subset of four predictors obtaining the results shown in Table 4.8. From the results, the higher dependency on the predictor d is observed, which is followed by f, v_{depth}, and Δ_h.

Table 4.7: 2D-ALE for the GTB model for the subset of four predictors.

Predictor	2D-ALE value
v_{depth},Δ_h	0.00
f,Δ_h	0.02
f, v_{depth}	0.03
d,Δ_h	0.06
d, v_{depth}	0.08
d,f	0.40

Table 4.8: PFI rank value of the four predictors subset shown in crescent order.

Predictor	PFI value (loss: RMSE [dB])
Δ_h	0.25
v_{depth}	0.53
f	0.98
d	4.51

Therefore, the predictor Δ_h and v_{depth} present the lower influence to improve PL prediction for the GTB model; however, given the observed improvement on the ABG and CIF models upon including the predictors Δ_h and v_{depth}, we maintain them in the following subsection for comparison.

4.6 Generalization Capacity Analysis

To address the generalization capacity of the models for short-range PL at mmWave, an additional evaluation is performed by considering unknown receivers. In this experiment, the GTB model using the subset of four predictors is trained using seventeen randomly selected receivers (approximately 80% of the total samples, resulting in 238 PL instances), and the remaining four receivers are used for testing (56 PL instances). This process is repeated three times, resulting in three different sets for training and testing, as shown in Figure 4.6 and Table 4.9. For each set, the optimal hyperparameters are determined using a 5-fold CV, and the selected hyperparameters are listed in the last column in Table 4.9. The coefficients of the ABG and $\text{ABG}_{\Delta_h, v_{\text{depth}}}$ are fit for each training set, and their values are presented in Table 4.10.

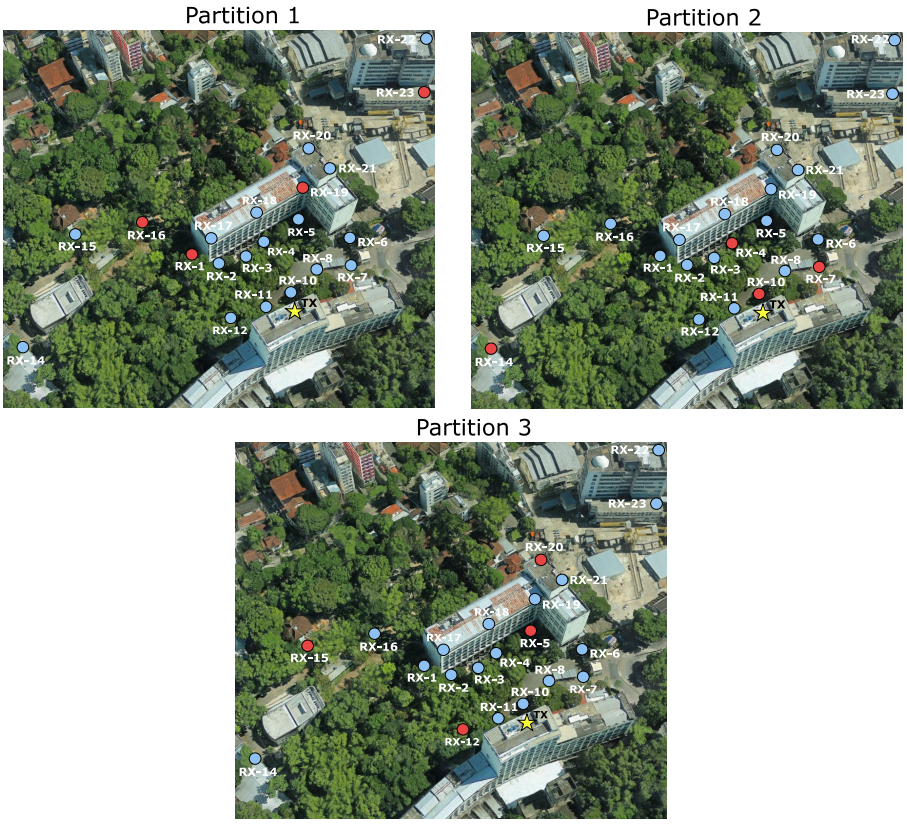

Figure 4.6: Generalization test capacity for unknown receivers for the mmWave outdoor links.

Table 4.11 compares the final models using the different training and testing sets partitions. The results show that the GTB-based PL prediction reaches the best generalization performance, although the $\text{ABG}_{\Delta_h, v_{\text{depth}}}$ attain good results, and the ABG presents the lowest performance.

Table 4.9: Generalization capacity analysis for the PL models. The models are adjusted/trained using seventeen receivers and evaluated (tested) on the four remaining ones.

	Rx's employed		
Partition	Train	Test	hyperparameters
1	RX_2,RX_4,RX_5,RX_6, RX_7,RX_8,RX_9,RX_{10}, RX_{11},RX_{12},RX_{14},RX_{15}, RX_{17},RX_{18},RX_{20},RX_{21},RX_{22}	RX_1,RX_{16},RX_{19},RX_{23}	η=0.1,trees=260, md=2,msl=2, absolute error loss
2	RX_1,RX_2,RX_5,RX_6, RX_8,RX_9,RX_{11},RX_{12}, RX_{15},RX_{16},RX_{17},RX_{18}, RX_{19},RX_{20},RX_{21},RX_{22},RX_{23}	RX_4,RX_7,RX_{10},RX_{14}	η=0.1,trees=30, md=2,msl=1, squared error loss
3	RX_1,RX_2,RX_4,RX_6, RX_7,RX_8,RX_9,RX_{10}, RX_{11},RX_{14},RX_{16},RX_{17}, RX_{18},RX_{19},RX_{21},RX_{22},RX_{23}	RX_5,RX_{12},RX_{15},RX_{20}	η=0.1,trees=60, md=1,msl=3, absolute error loss

Table 4.10: Adjusted coefficients of the ABG and ABG$\Delta_h, v_{\text{depth}}$ models for the generabizability analysis.

		α	β	γ	ζ	φ
	Partition 1	1.96	66.34	0.41	-	-
ABG	Partition 2	1.90	68.76	0.29	-	-
	Partition 3	2.01	67.35	0.27	-	-
	Partition 1	2.35	59.87	0.41	3.69	0.22
$\text{ABG}_{\Delta_h,v_{\text{depth}}}$	Partition 2	2.17	64.47	0.29	3.18	0.23
	Partition 3	2.40	60.84	0.27	4.05	0.19

4.7 Discussion

This chapter has investigated PL modeling for mmWave in an outdoor environment. The study proposed two adjusted empirical models, the $\text{ABG}_{\Delta_h,v_{\text{depth}}}$ and $\text{CIF}_{\Delta_h,v_{\text{depth}}}$. By considering the additional predictors $(\Delta_h, v_{\text{depth}})$, the derived models can capture the influence of the difference in antenna height between the transmitter and the receiver and vegetation depth along the path profile. Also, we delved into the design of optimized ML-based PL models for mmWave links outdoors. In this category, the best performance was attained for the GTB-based PL model. The GTB-based PL model has shown the potential to accurately predict path loss in the considered environment. However, additional measurement data from different scenarios still need to be addressed to further evaluate the proposed ML-based PL model.

Table 4.11: Performance evaluation of the empirical and GTB models considering a set of receivers for training and the test employs a different set.

		Training set				Testing set			
	Partition	RMSE [dB]	MAPE [%]	σ [dB]	R^2	RMSE [dB]	MAPE [%]	σ [dB]	R^2
ABG	1	2.42	1.74	1.44	0.68	2.21	1.56	1.33	0.67
	2	2.33	1.60	1.50	0.71	2.70	1.90	1.60	0.54
	3	2.42	1.73	1.46	0.63	2.24	1.55	1.36	0.78
$\text{ABG}_{\Delta_h, v_{\text{depth}}}$	1	2.23	1.53	1.45	0.73	1.91	1.38	1.09	0.75
	2	2.14	1.47	1.37	0.75	2.40	1.51	1.65	0.63
	3	2.22	1.54	1.42	0.68	2.00	1.36	1.25	0.83
GTB	1	1.41	0.83	1.09	0.89	1.89	1.37	1.56	0.75
	2	1.84	1.21	1.38	0.82	2.29	1.62	1.43	0.66
	3	1.95	0.76	1.37	0.80	1.98	1.36	1.24	0.83

Given that the measurement campaign for the outdoor environment presents mostly LOS conditions in comparison with the indoor environment, the connections between predictors and path loss are best learned in this environment, as demonstrated by the better performance prediction for the ML and empirical models for outdoor propagation models than for indoors.

The following chapter studies the applicability of the ML models design methodology for links provided by an urban macrocell in lower frequencies (sub-6 GHz), in which the propagation environment is defined by larger Tx-Rx distance and the presence of buildings and dense vegetation in the surroundings.

5
Path Loss Prediction for Macrocell Coverage at sub 6-GHz using Machine Learning Techniques

The next generation of mobile communication systems relies on mmWave technology to fulfill high data rates and low latency requirements. However, 5G technology involves the integration of overlapping cells consisting of macrocells designed to provide extensive coverage for the so-called heterogeneous network [4, 33]. Typically, the macrocells will operate using sub-6 GHz bands, such as 700 MHz, 2.5 GHz, and 3.5 GHz [33].

This chapter tackles modeling PL prediction for macrocell coverage in the sub-6 GHz bands. The models are derived using an extensive measurement campaign that encompasses two Routes in Rio de Janeiro, Brazil. For the design of the ML models, one extracts predictors (features) from the path profile between the transmitter and the receiver.

5.1
Related Work

With the deployment of macrocell mobile networks, many models have appeared to help predict the coverage considering different system-and environment-dependent predictors. The Egli model [37] includes predictors such as distance, frequency, and antenna heights; Okumura's model proposes different models according to the terrain type [38] considering frequency, distance, and antenna heights; and Lee model [39] uses predictors such as frequency, distance, antenna heights, building heights and street orientation.

The site-specific model proposed in [33] combines three models to a single model: the discrete mixed Fourier transform split-step parabolic-equation (DMFT-SSPE) [152], the building-transmission model (BTM) [153], and the International Telecommunication Union of Radio-communications (ITU-R) model for attenuation in vegetation [33]. The study employs profile information from the ground, the buildings, and the vegetation. In addition, the diffracted components arising from the building rooftops and terrain peaks present in the radio path are also considered. The model's performance was assessed using data from a measurement campaign in Rio de Janeiro conducted in two Routes. Some sub-regions from the two Routes refereed as calibration scenarios were

used to adjust the coefficient parameters of the proposed propagation model represented by the red line in Figure 5.1. The results were also compared with the log-distance and Okumura-Hata models; the authors claim a better performance for their model. We employed the same dataset for PL modeling for macrocell coverage at the sub-6 GHz spectrum and compare the obtained results against the ones in [33].

Figure 5.1: Distribution of samples for the model design of the work proposed in [33]. The regions represented by the red line were used to adjust the coefficient parameters in the models.

ML-based PL modeling for macrocell coverage has already been proposed in [63, 64]. In [63], an ANN model is developed to predict PL and shadowing in a suburban environment considering frequencies 450 MHz, 1500 MHz, and 3600 MHz. The dataset is split for training (80%) and testing (20%) using random sampling. The principal component analysis (PCA) algorithm is applied to extract relevant predictors. The ANN model was evaluated using different predictor subsets, including a subset with only one predictor (distance) and a subset with four predictors (transmitting antenna height, receiving antenna height, transmitting/receiving antenna height ratio, and distance). When using only the distance as the predictor, the ANN model achieved an RMSE, MAPE, and R^2 of 8.61 dB, 4.94%, and 0.66, respectively, on the testing set. With the inclusion of four predictors, the model performance shows almost no improvement, attaining an RMSE, MAPE, and R^2 equal to 8.40 dB, 4.86%, and 0.67, respectively.

The work in [64] focuses on the development of a combined PL model for an urban environment in the 3.5 GHz frequency band using a ray-tracing

technique to generate the dataset. It proposes an ABG model for PL prediction when LOS conditions exist, while for NLOS cases, a model optimized by least-square regression is employed. The study includes the PCA algorithm for predictor selection, and nine predictors are employed to characterize the building profile. The authors report the performance prediction by the mean error and σ, with values of 0.00 dB and 12.28 dB, respectively, using a transmitter height of 30 m. Most of the previously cited works employ predictors such as distance, antenna height, and building information. In contrast, this thesis considers a more extensive framework; we identify fourteen predictors from the building, vegetation, and ground profiles and employ several ML techniques to characterize the entire profile.

The ML-based PL models are first built using the entire set of predictors. Then, the predictor coalitions methodology (see Section 3.9) is applied to identify the most significant predictors for the ML-based PL model with the best performance for macrocell coverage within the sub-6 GHz frequency band. The performance of the final ML model is compared with the empirical models log-distance and Okumura-Hata, and the model proposed in [33]. Furthermore, to assess the generalization capacity of the ML techniques, we compare the performances obtained using conventional training-testing splitting and a training-testing approach that segregates the data from different streets.

5.2 The Measurement Campaign: Dataset Description

The measurements campaign was carried out in 2020 and was conducted by colleagues from the radio propagation laboratory in CETUC and the Polytechnic Institute of Leiria, Portugal. The measurement campaign includes two Routes, as depicted in Figure 5.2. The blue line corresponds to the first Route, where the transmitter is positioned on the rooftop of a building with an antenna height of 53 meters from the ground, represented by the blue star. The second Route, represented by a magenta line, covers the Leblon area. The magenta star marks the transmitter's position for this scenario, with an antenna height of 71 meters. The measurements were carried out at different frequencies, including 750 MHz, 2.5 GHz, and 3.5 GHz [33].

The campaign used vector signal generators to generate continuous-wave (CW) signals at 10 dBm. The transmitter is an Aaronia HyperLOG 60100 antenna with a 5 dBi gain. The receiver is an RFS I-ATO1-380/6000 H-plane omnidirectional antenna. The measurement software also controlled a GPS device to assign the position coordinates to each measurement sample acquired [33]. The PL in dB is obtained from the measured received power (P_{RX}) and

system parameters using Equation (3-1). The samples collected from both Routes and the three frequencies are shown in Table 5.1. There were 15,824 samples collected during the measurement campaign.

Figure 5.2: Map of the measurement campaign in the macrocell coverage at the sub-6 GHz frequency band.

Table 5.1: Samples collected in each Route and frequency.

Route	Frequency	Samples
#1	750 MHz	3,887
	2.5 GHz	3,886
	3.5 GHz	3,747
#2	750 MHz	1,459
	2.5 GHz	1,459
	3.5 GHz	1,386

5.2.1 Identification of the Predictors from the Profile Path between the Transmitter and the Receiver

The path profile for each measured PL is available at [33]. They include buildings, vegetation, and ground profiles, as well as the identification of the diffraction heights in the path. The profile path between the Tx and the Rx provides extensive information to identify valuable predictors for the PL prediction. An example of the profile path between the transmitter and the

receiver is shown in Figure 5.3. From the profile environment and system parameters, fourteen predictors are identified. They include frequency, in MHz (f), the height difference between the transmitter and receiver (Δ_h), in meters, and the Tx-Rx distance (d), in meters. Given the urban environment characteristics, where buildings can obstruct radio paths, leading to additional propagation loss [64], we identified four predictors related to building height variability along the link. They are quantified by the mean and standard deviation of the building height ($\overline{b}_h$ and σ_{b}), in meters, as well as the obstruction caused by buildings, measured by the building depth (b_{depth}), in meters, and the number of intersected buildings (n_b).

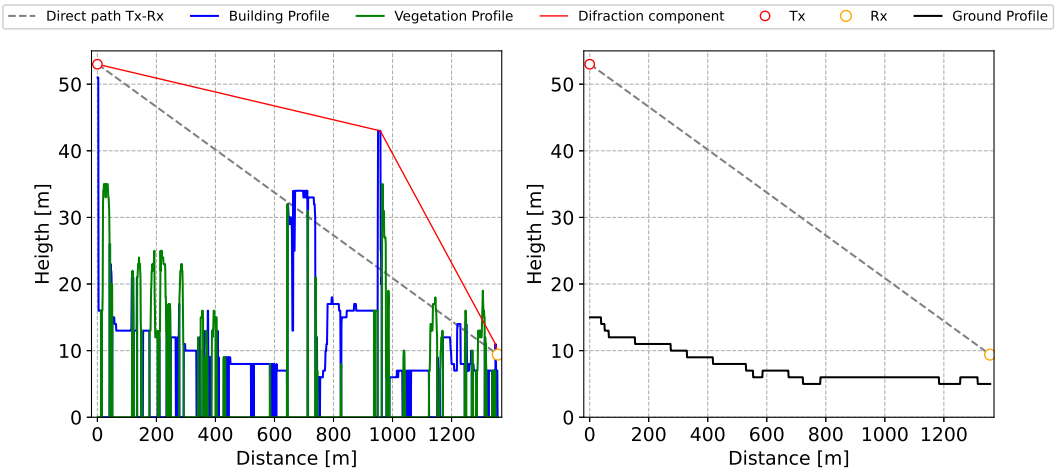

Figure 5.3: Example of the profile environment between the transmitter and receiver. The left-hand graph shows the building, vegetation and diffraction profile along distance, and the right-hand graph shows the ground profile.

In addition, the presence of vegetation can significantly impact signal propagation, leading to additional path loss and signal attenuation [33], we extract four predictors from the vegetation profile that include the mean and standard deviation of the vegetation height along the radio path ($\overline{v}_h$ and σ_{v}), in meters, and the obstruction of vegetation measured by the vegetation depth (v_{depth}), in meters, and the number of intersected vegetation/trees (n_v). Furthermore, we also consider the variability of the ground height along the direct path by measuring its mean and standard deviation ($\overline{g}_h$ and σ_{g}), in meters. Lastly, from the diffraction profile, we obtain the mean of the diffraction height ($\overline{\mathrm{dif}}_h$), in meters. The ranges for the predictors in each Route are shown in Table 5.2.

Table 5.2: Range values of the predictors for the Route #1 and Route #2.

Predictor	Route #1	Route #2
f	750 MHz - 3.5 GHz	750 MHz - 3.5 GHz
Δ_h	31.58 - 49.50 m	56.07 - 63.53 m
d	171.28 - 3,542.11 m	121.93 - 1,881.94 m
$\overline{b}_h$	2.91 - 27.34 m	6.32 - 21.84 m
$\overline{v}_h$	2.14 - 28.05 m	0.94 - 14.51 m
$\overline{g}_h$	5.76 - 26.09 m	6.31 - 18.50 m
σ_{b}	4.49 - 20.48 m	6.29 - 20.73 m
σ_{v}	5.24 - 27.46 m	6.79 - 20.37 m
σ_{g}	0.64 - 16.90 m	4.50 - 13.55 m
b_{depth}	0 - 241.71 m	0 - 353.30 m
v_{depth}	0 - 307.04 m	0 - 153.13 m
n_b	0 - 26	0 - 37
n_v	0 - 27	0 - 29
$\overline{\mathrm{dif}}_h$	20.50 - 81.79 m	27.45 - 88.44 m

5.2.2 Distribution of the Measured PL

Figure 5.4 provides a visual representation of the distribution of the measured PL samples using the cumulative distribution function (CDF). The plots show that the CDFs for the measured PL are similar across the three frequencies in each Route. Furthermore, Figure 5.5 displays a scatter plot showing the dependence of path loss on distance, which shows that there is a great dispersion in the PL when ordered by the Tx-Rx distance.

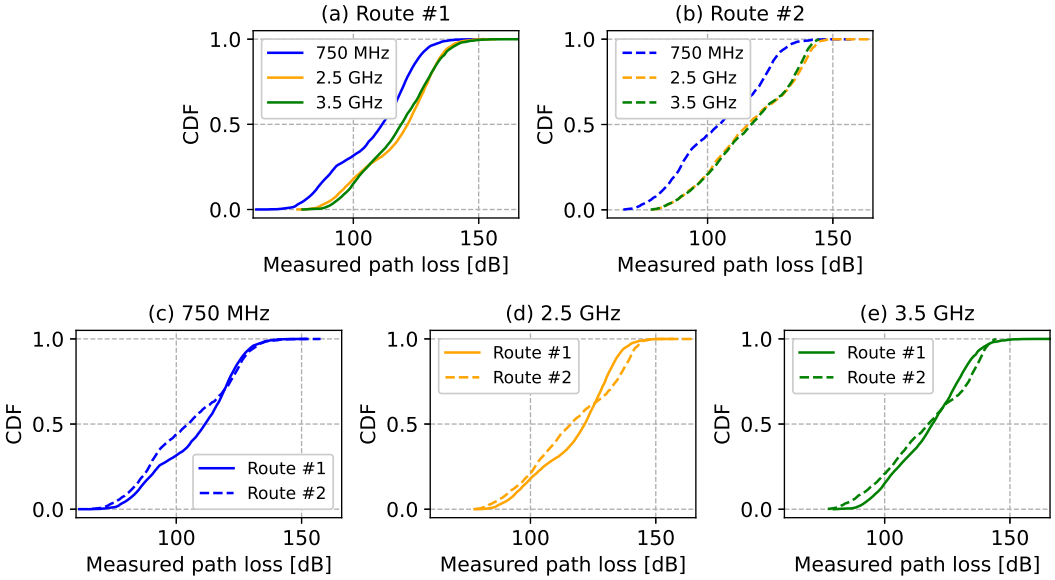

Figure 5.4: CDF of the measured PL in Route #1 and Route #2.

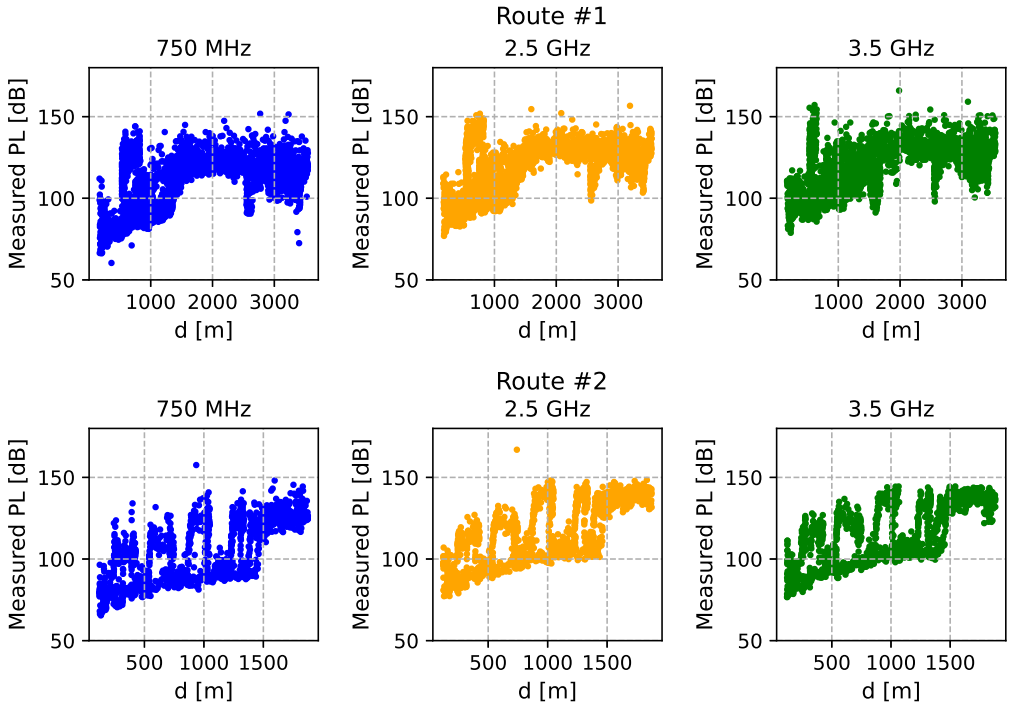

Figure 5.5: Scatter plots of the measured path loss versus distance for each Route and frequency.

In addition, we calculate the variance (σ^2) and mean ($\overline{y}$) of the measured PL to examine the similarity between the statistical distribution in both Routes and frequencies, which are presented in Table 5.3. The results show that the mean of the measured PL in both Routes is similar. Route #2 presents higher variances, indicating a wider dispersion of the samples in Route #2 than in Route #1.

Table 5.3: Parameters of the distributions for σ^2 and mean of the measured PL.

	σ^2 of measured PL [dB^2]			mean of measured PL [dB]		
	750 MHz	2.5 GHz	3.5 GHz	735 MHz	2.5 GHz	3.5 GHz
Route #1	263.88	231.42	213.84	107.94	117.88	117.64
Route #2	331.42	331.26	308.51	104.69	116.63	116.58

5.3 Path Loss Empirical Model

This section describes two common empirical models used for macrocell coverage. They are the log-distance and Okumura-Hata models. The log-distance PL model is an extension of Friss free-space model [27], given by

$$\mathrm{PL}[dB] = \mathrm{L}_0 + 10n\log_{10}(d), \tag{5-1}$$

where L_0 is the reference PL, n is the path loss exponent, and d is the Tx-Rx distance in meters, indicating the rate at which the path loss increases with the distance. The coefficients L_0 and n are estimated using the least squares regression technique [154]. Its values can vary depending on the specific environment and frequency.

The Okumura-Hata model is valid for microwave frequencies from 150 to 1500 MHz and covers receiver antenna heights between 1 and 10 m, transmitter antenna heights ranging from 30 to 200 m, and link distances from 1 to 10 km. For an urban environment [52], the Okumura-Hata is given by

$$\begin{aligned} \mathrm{PL}[dB] =& 69.55 + 26.16\log_{10}(f) - 13.82\log_{10}(h_b) - \\ & C_H + (44.9 - 6.55\log_{10}(h_b))\log_{10}(d), \end{aligned} \tag{5-2}$$

where f is the frequency, in MHz, h_b is the heigth of the transmitter, in meters, h_m is the height of the receiver, in meters, C_H is the antenna height correction factor $(3.2(\log_{10}(11.75h_m))^2 - 4.97)$, in dB, and d is the distance, in kilometers.

5.4 ML-based PL Models

This section compares the conventional methodology for train-test splitting and a proposed approach for splitting the dataset for designing the macrocell PL prediction model. A comparison of the obtained ML model with the empirical models and with the study in [33] are presented in Section 5.5 and Section 5.6, respectively.

5.4.1 Training-Testing Methodology to Force "Unknown Streets"

As we have already discussed in the previous chapter, for the development of the ML model, it is necessary to use a percentage of the data for learning the behavior of the predictors and separating the remaining data to test the model obtained. Furthermore, the training set must represent the testing set [40]. The conventional approach is to randomly shuffle the dataset and allocate a certain percentage, such as 70% or 80% for training the model [32]. By using this split technique in the context of PL macrocell coverage prediction, we can assess the model's capacity to interpolate, that is, to measure the generalization prediction for instances (and cases) that are close to those presented for the model learning as discussed in [32, 40]. In that case, the samples (predictors and output) on the training and testing set would have a very similar statistical distribution; and, the model is expected to generalize well using the conventional train-test split and relevant predictors since model

prediction is sensitive to the training data domain. However, it is important to measure how well the model performs on completely new and unseen data from unknown streets, that is, when extrapolating [32].

Therefore, we compare different train-test splits as illustrated in Figure 5.6. In the first, one randomly shuffles the dataset instance (80% for training and 20% for testing) from both Routes. In the second, the train-test splitting considers the streets (samples' positions) such that the testing set has instances collected in streets that are unknown during training to further analyze the generalization performance of the models.

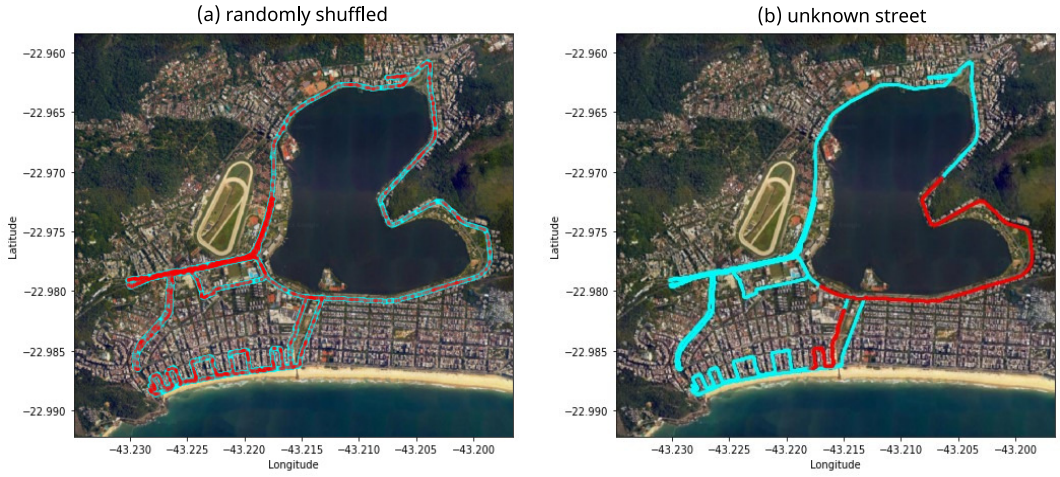

Figure 5.6: Example of the randomly shuffled (graph on the left) and unknown streets (on the right), for the training (in blue) and testing (red) sets division methodology.

Table 5.4 presents the variance and mean of the PL instances in the training and testing sets using the two dataset splits above discussed. As expected, the σ^2 and mean of the measured PL on the training and testing sets divided using random split have very close values compared with the training and testing for "unknown streets", where the mean of the measured PL is similar, however, σ^2 is higher for the training set.

Table 5.4: Comparison of the train-test randomly shuffled and unknown streets strategies in terms of σ^2 and mean for the PL values they contain.

		σ^2 of measured PL [dB^2]			mean of measured PL [dB]		
Splitting technique	Set	750 MHz	2.5 GHz	3.5 GHz	735 MHz	2.5 GHz	3.5 GHz
Randomly shuffled	Train	285.87	259.00	240.72	107.02	117.73	117.30
	Test	278.32	258.18	235.30	107.21	116.77	117.56
Unknown street	Train	298.14	272.81	245.54	104.91	115.34	115.44
	Test	114.90	83.54	87.99	116.48	127.20	127.42

The GTB model is employed considering all the predictors available in the dataset to compare the impacts of using the train-test splits discussed.

For the random shuffling approach, the dataset is randomly divided into 80% of the instances for training (12,659) and 20% for testing (3,165). The 5-fold CV technique using a grid search explores the range of the hyperparameters for the GTB PL model. For this scenario, we observe for all the machines that using the same hyperparameters values as those used in the mmWave scenarios tend to produce over-fitting; therefore, for the macrocell coverage, we employ a narrower range of hyperparameters values to avoid that issue. The hyperparameters considered include the number of trees (ranging from 2 to 70), the maximum depth (ranging from 1 to 3), the minimum number of samples in a leaf (ranging from 1 to 3), and the learning rate (ranging from 0.001 to 0.1). During the grid search, two different loss functions are tested, squared error and absolute error. The lowest CV RMSE (6.41 dB) is achieved with 70 trees, maximum depth equal to 3, minimum samples leaf equal to 1, and a learning rate of 0.1, using the squared error loss function.

For the train-test split considering unknown streets, the train-test split follows the division presented in Figure 5.6.(b) – employing samples from both Routes to highlight the effects of training using different environments. The training set contains 13,024 samples extracted from both Routes (9,575 from Route #1 and 3,449 from Route #2). The remaining samples (2,800) are assigned to the testing set; among these samples, 1,945 are from Route #1, while 855 are from Route #2. The GTB model reaches the lowest CV RMSE (6.37 dB) using 70 trees, a maximum depth of 3, and the minimum samples per leaf equal to 1 with the squared error loss function. Wrapping up, if the two strategies of train-test splits use very different numbers of instances (and from different numbers of scenarios), the comparison is hindered.

The performance evaluation of these GTB models for both train-test splits is shown in Table 5.5. As expected, the model using the random train-test split has the best generalization PL prediction. This can be attributed to the testing set containing samples that are very close (in the predictor and output spaces) to those on which the model was trained, which achieves an RMSE, MAPE, σ and R^2 equal to 6.41 dB, 4.33%, 4.22 dB and 0.85, respectively. However, for PL prediction on similar streets, the proposed train-test split emphasizes the applicability of the trained ML-based PL model. For the proposed train-test split, the GTB model achieves an RMSE, MAPE, σ, and R^2 on the testing equal to 7.36 dB, 4.61%, 4.86 dB, and 0.56, respectively. Although the performance prediction is lower than the train-test random split, this approach considers the actual application of the trained ML-based model.

Table 5.5: Performance evaluation of the GTB model measured on the training and testing sets.

	Training set				Testing set			
Train-test	RMSE [dB]	MAPE [%]	σ [dB]	R^2	RMSE [dB]	MAPE [%]	σ [dB]	R^2
random split	6.26	4.22	4.10	0.86	6.41	4.33	4.22	0.85
unknown street	6.17	4.21	4.06	0.87	7.36	4.61	4.86	0.56

5.4.2 ML-based PL Models

In this subsection, the "unknown streets" train-test split is applied to design four ML models: ANN, SVR, RF, and GTB for macrocell at sub 6-GHz links. The training and testing sets are depicted in Figure 5.6.(b). The training set is used to optimize the hyperparameters of each model using a 5-fold CV. The predictors for the models are shown in Table 5.2.

During the grid search, the ANN model considers a number of neurons varying from 1 to 20, learning rate and weight decay from 0.001 to 0.1 for both hyperparameters, activation functions ReLU, Logistic, and Tanh, with the solver L-BGS and early-stopping. The lowest average CV RMSE (5.91 dB) is achieved using 19 neurons, a learning rate equal to 0.001, and a weight decay of 0.1 with the activation function ReLU. The grid search for the SVR model considers C ranging from 1 to 100, ϵ ranging from 0.0001 to 0.01, and σ_{RBF} ranging from 0.0002 to 0.02. The lowest average CV RMSE (6.63 dB) is achieved for C equal to 100, ϵ to 0.0001, and σ_{RBF} to 0.002.

The grid search for the RF model includes the number of trees varying from 2 to 70, the maximum depth varying from 1 to 3, and the minimum samples per leaf varying from 1 to 3 with the squared error or the absolute error as the loss function. The lowest average CV RMSE (8.52 dB) is attained using 44 trees, with a maximum depth of 3, minimum samples leaf of 3, with the squared error loss function. For the GTB, the employed hyperparameters are those found in Subsection 5.4.1: 70 trees, maximum depth of 3, minimum samples leaf of 1, learning rate of 0.1, and squared error loss function.

The results for the final ML models are displayed in Table 5.6. The GTB model achieves the best generalization performance with an RMSE, MAPE, σ, and R^2 equal to 7.36 dB, 4.61%, 4.86 dB, and 0.56, respectively. The lowest performance prediction is provided by the SVR model with an RMSE, MAPE, σ, and R^2 equal to 10.80 dB, 6.19%, 7.79 dB, and 0.05, respectively. Figure 5.7 shows the measured and predicted path loss values for the ML models.

Table 5.6: Performance evaluation of the ML-based PL models for the sub-6 GHz macrocell environment measured on the training and testing set.

	Training set				Testing set			
Model	RMSE [dB]	MAPE [%]	σ [dB]	R^2	RMSE [dB]	MAPE [%]	σ [dB]	R^2
ANN	5.71	3.91	3.73	0.89	10.51	6.98	6.30	0.10
SVR	6.57	4.40	4.42	0.85	10.80	6.19	7.79	0.05
RF	8.55	5.96	5.46	0.75	9.18	5.80	5.80	0.31
GTB	6.17	4.21	4.06	0.87	7.36	4.61	4.86	0.56

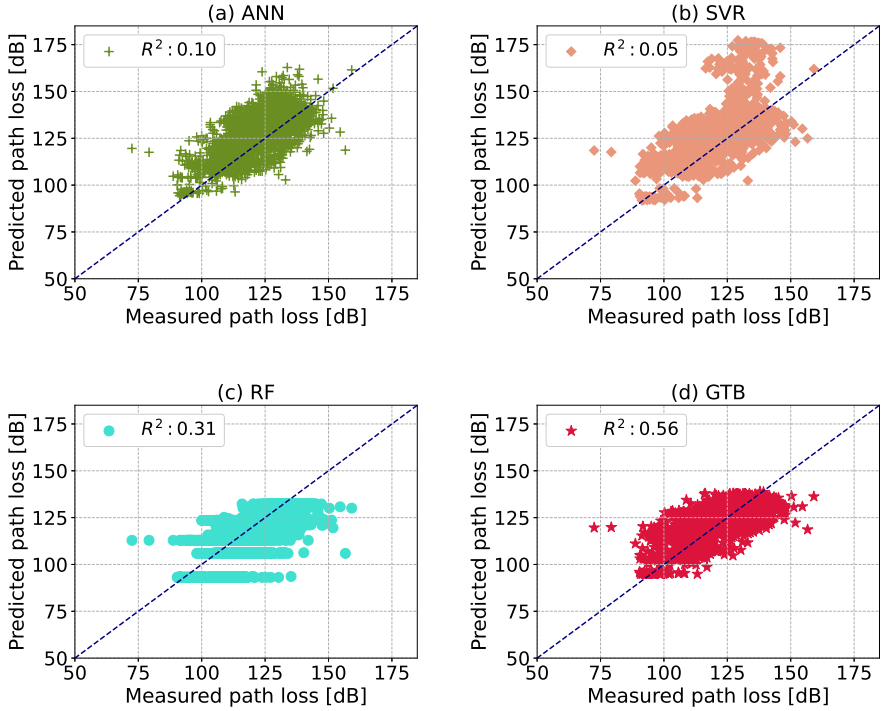

Figure 5.7: Comparison of measured and predicted path losses on the testing set for the ML models in the macrocell environment.

5.4.3 Predictor Coalition Selection

Since the GTB model achieves the best performance, as can be seen in Table 5.6, our predictor selection methodology described in Section 3.9 is employed to identify the most relevant predictors for this model. Table 5.7 presents the results of performance evaluation and interpretability on each subset. One notes that the inclusion of a new predictor in the coalition leads to an improvement in performance, more significantly for the first three predictors b_{depth}, d, and f. After the sixth predictor (v_{depth}), the influence

of a new predictor in the performance reduces. Furthermore, the IAS and $\overline{\text{MEC}}$ indicators consistently maintain similar values in all subsets, implying that including a new predictor does not introduce strong interactions with the previously selected predictors.

Table 5.7: Performance and IML indicators for the GTB model for macrocell in the sub-6 GHz using different coalitions measured on the training set. Model performance is measured using RMSE, MAPE, σ, and R^2; interpretability is assessed by IAS and $\overline{\text{MEC}}$.

Predictors coalitions	RMSE [dB]	MAPE [%]	σ [dB]	R^2	IAS	$\overline{\text{MEC}}$
b_{depth}	10.25	7.61	6.19	0.65	-	2.00
b_{depth},d	8.63	6.34	5.21	0.75	0.11	2.68
b_{depth},d,f	6.94	4.79	4.46	0.84	0.10	2.52
b_{depth},d,f,$\overline{g}_h$	6.51	4.49	4.24	0.86	0.11	2.04
b_{depth},d,f,$\overline{g}_h$,Δ_h	6.43	4.42	4.20	0.86	0.11	2.11
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth}	6.32	4.35	4.14	0.87	0.11	2.01
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$	6.31	4.33	4.13	0.87	0.10	1.99
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$	6.30	4.32	4.13	0.87	0.10	2.01
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$,σ_b	6.29	4.30	4.13	0.87	0.11	1.96
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$,σ_b,σ_g	6.15	4.21	4.04	0.87	0.10	2.03
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$,σ_b,σ_g,σ_v	6.13	4.19	4.03	0.87	0.10	2.01
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$,σ_b,σ_g,σ_v,n_v	6.13	4.19	4.02	0.87	0.11	2.06
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{v}_h$,$\overline{b}_h$,σ_b,σ_g,σ_v,n_v,$\overline{\text{dif}}_{b_h}$	6.12	4.19	4.01	0.87	0.11	2.04
b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth},$\overline{b}_h$,$\overline{v}_h$,σ_b,σ_g,σ_v,n_v,$\overline{\text{dif}}_{b_h}$,$n_b$	6.17	4.21	4.06	0.87	0.09	2.07

In addition, Figure 5.8 and Table 5.8 present the ALE curves and PFI values for the selected subset (coalitions) of predictors, respectively. The PFI values indicate that the predictor b_{depth} has the highest influence in improving PL prediction, followed by the predictors d, f, and $\overline{g}_h$. The lower PFI values are for the predictors v_{depth} and Δ_h. From the ALE curves (Figure 5.8.(c)), it can be seen that the predictor f has a quite linear effect on the predicted PL, contrary to the GTB models in mmWave when frequency had a non-linear effect as discussed in Chapter 3, Subsection 3.9.3.

The previous results indicate that including the predictors such as $\overline{v}_h$,$\overline{b}_h$,σ_b,σ_g,σ_v,n_v,$\overline{\text{dif}}_{b_h}$, and n_b do not significantly enhance the performance of the model. As a result, the subset of six predictors (b_{depth},d,f,$\overline{g}_h$,Δ_h,v_{depth}) is selected to build the final GTB model.

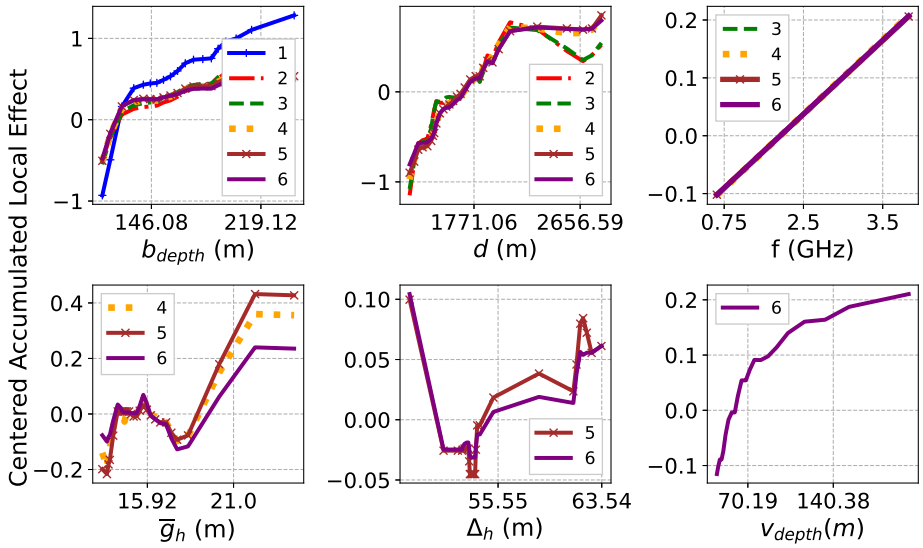

Figure 5.8: ALE plots for the GTB model evaluated from the subset of one predictor (b_{depth}) until the selected subset (b_{depth},d,f,$\overline{g}_h$,Δ_h, v_{depth}). The y-axis shows the centered accumulated local effect values.

Table 5.8: PFI rank value of the selected subset of predictors shown in crescent order.

Predictor	PFI value (loss: RMSE [dB])
Δ_h	0.51
v_{depth}	0.89
$\overline{g}_h$	1.36
f	3.38
d	4.94
b_{depth}	8.23

5.5 Final Models Comparison between the GTB and Empirical Models

In this section, we compare the performance at the different frequencies of the GTB-based PL model against single-frequency empirical models, such as the log-distance and Okumura-Hata models obtained in Section 5.3. For the GTB, a grid search with a 5-fold CV is applied to identify the optimal hyper-parameters for the selected subset using the same range of values described in Subsection 5.4.2. The GTB model achieves the lowest CV RMSE (6.47 dB) using 70 trees, with a maximum depth equal to 3, a minimum samples leaf of 3, a learning rate of 0.1, and the squared error loss function. The performance evaluation results on the training and testing, for each set of samples obtained for Route #1 and Route #2 are presented in Table 5.9.

Figure 5.9 and Figure 5.10 compare the measured and predicted PL values on the GTB model on the testing set for each Route. Furthermore,

Table 5.9: Performance evaluation of the GTB model measured on the training and testing sets using the optimized subset of predictors for the sub-6 GHz macrocell environment.

Training set				Testing set				
RMSE [dB]	MAPE [%]	σ [dB]	R^2	Route	RMSE [dB]	MAPE [%]	σ [dB]	R^2
6.33	4.35	4.14	0.86	#1	7.88	4.92	5.29	0.44
				#2	6.40	4.03	3.98	0.72

we compare the performance of the GTB model and the empirical models log-distance and Okumura-Hata shown in Table 5.11 and Table 5.12. The coefficients for the log-distance model (L_o and n) are estimated using Equation (5-1) and employed on the training set. They are reported in Table 5.10. For the Okumura-Hata model, Equation (5-2) is employed on the testing set for PL prediction.

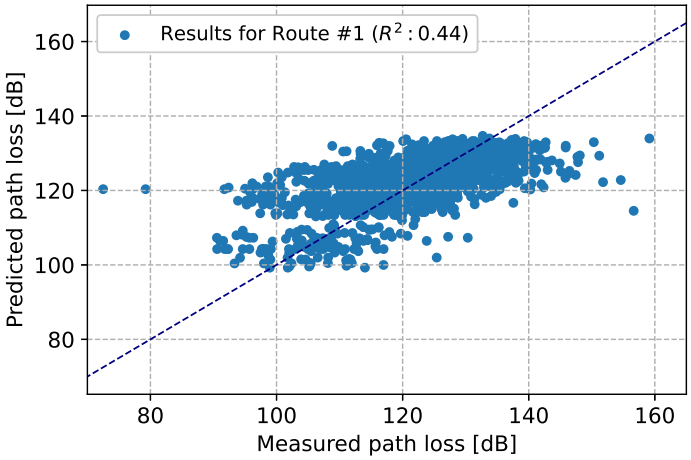

Figure 5.9: Comparison of measured and predicted path losses on the testing set for the GTB-based PL model measured on the Route #1.

Table 5.10: Values of the coefficients for the log-distance model in each frequency, estimated on the training set.

Coefficients	750 MHz	2.54 GHz	3.5 GHz
L_o	13.82	27.28	34.55
n	3.05	2.95	2.70

Table 5.11 and Table 5.12, present the performance indicators including RMSE, MAPE, σ, R^2 and R^2_{oos}, which measures the out-of-sample R^2 [155]

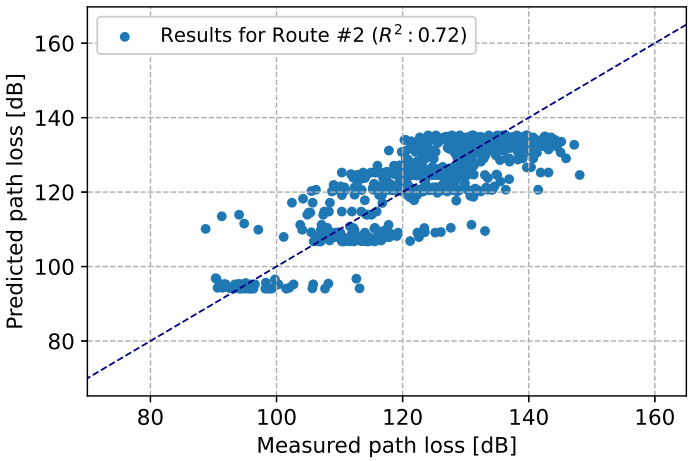

Figure 5.10: Comparison of measured and predicted path losses on the testing set for the GTB-based PL model measured on the Route #2.

and is given by

$$R^2{}_{\text{oos}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{I}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{I}\left(y^{(i)} - \overline{y}_{\text{train}}\right)^2}. \tag{5-3}$$

$R^2{}_{\text{oos}}$ compares the sum of squared errors $(y^{(i)} - \hat{y}^{(i)})^2$ from the trained or adjusted model, and the sum of squared errors $(y^{(i)} - \overline{y}_{\text{train}})^2$ concerning the mean of the measured PL on the training set (the simplest fit), instead from the testing set as calculated for R^2 in Equation (3-10). The evaluation of similarity in those indicators could provide additional insight into how close the statistical distribution of the PL instances on the training and testing set is. Thus, when the mean of the measured PL from both sets is similar, the values of R^2 and $R^2{}_{\text{oos}}$ are also expected to be similar.

The results presented in Table 5.11 and Table 5.12 show a significant improvement by the GTB model compared to the log-distance and Okumura-Hata in both testing subsets. The GTB model achieves the best generalization performance for the samples from Route #2 (as seen in Table 5.12), with the lowest RMSE, MAPE, and σ, and the highest R^2 and $R^2{}_{\text{oos}}$ values; also the smallest difference between R^2 and $R^2{}_{\text{oos}}$ for each frequency is obtained for this testing subset; which means a closer value of $\overline{y}_{\text{test}}$ and $\overline{y}_{\text{train}}$, consequently its generalization is better. The negative value for R^2 on the testing for the empirical models indicates that the mean of the measured PL performs better in predicting the PL than the adjusted model.

Table 5.11: Performance evaluation for each frequency measured on the testing set from samples of Route #1 in the macrocell coverage.

Model	Frequency	RMSE [dB]	MAPE [%]	σ [dB]	R^2	R^2_{oos}
Log-distance	750 MHz	11.70	8.40	7.05	-0.32	0.40
	2.5 GHz	9.97	6.55	6.04	-0.42	0.45
	3.5 GHz	10.01	6.44	5.99	-0.20	0.53
Okumura-Hata	750 MHz	15.61	11.45	9.48	-1.34	-0.07
	2.5 GHz	17.44	12.29	9.24	-3.34	-0.69
	3.5 GHz	19.88	14.31	9.64	-3.74	-0.86
GTB	750 MHz	8.76	5.84	5.87	0.26	0.66
	2.5 GHz	6.62	3.95	4.48	0.37	0.76
	3.5 GHz	8.18	5.01	5.26	0.20	0.68

Table 5.12: Performance evaluation for each frequency measured on the testing set from samples of Route #2 for macrocell coverage.

Model	Frequency	RMSE [dB]	MAPE [%]	σ [dB]	R^2	R^2_{oos}
Log-distance	750 MHz	13.91	10.49	6.24	-0.38	0.35
	2.5 GHz	14.48	9.75	6.28	-1.09	0.37
	3.5 GHz	13.12	8.88	5.94	-0.81	0.36
Okumura-Hata	750 MHz	11.65	8.44	6.59	0.03	0.54
	2.5 GHz	9.83	6.49	5.31	0.04	0.71
	3.5 GHz	9.97	6.71	6.14	-0.05	0.63
GTB	750 MHz	7.05	4.55	4.65	0.64	0.83
	2.5 GHz	6.20	3.89	3.59	0.61	0.88
	3.5 GHz	5.85	3.64	3.57	0.64	0.87

5.6 Generalization Capacity Analysis

When applying the trained model to a new scenario, an additional test is performed to assess the GTB model's generalization capacity to learn the relationship between predictors and PL in macro-condition environments. The testing set is selected to represent a different scenario and enables the evaluation of the model's generalization to unseen data. The test involves training the model on the entire set of samples from Route #1 (11,520 samples) and evaluating its performance on the complete set of samples from Route #2 (4,304 samples). The spatial distribution of the PL instances in the two Routes is provided in Figure 5.11. The 5-fold CV technique is applied over the training set to determine the optimal hyperparameters using the grid search values described in Subsection 5.4.2.

The final GTB model is designed using 70 trees, a maximum depth equal to 3, and a minimum number of samples per leaf equal to 2. On the training set, this model achieves an RMSE, MAPE, σ and R^2 equal to 6.38 dB, 4.23%, 16.07

dB of 0.84, respectively, and an RMSE, MAPE, σ and R^2 on the testing set equal to 9.45 dB, 6.72%, 18.86 dB and 0.75, respectively. The results for each frequency on the testing set are summarized in Table 5.14, which compares the performance prediction achieved by the GTB, the log-distance, and Okumura-Hata models. The estimated coefficients L_o and n for the log-distance model at each frequency are presented in Table 5.13. When comparing the results of the GTB model to the previous ones, we observe a little decrease in performance indicated by higher values of RMSE, MAPE, and σ. This decrease can be attributed to the ML model being tested in a new scenario from its training environment. However, the GTB model still performs better than all the other empirical models.

Table 5.13: Coefficients value for the log-distance model estimated over the training set for the generalization test.

Coefficients	750 MHz	2.54 GHz	3.5 GHz
L_o	21.59	35.75	39.26
n	2.77	2.64	2.52

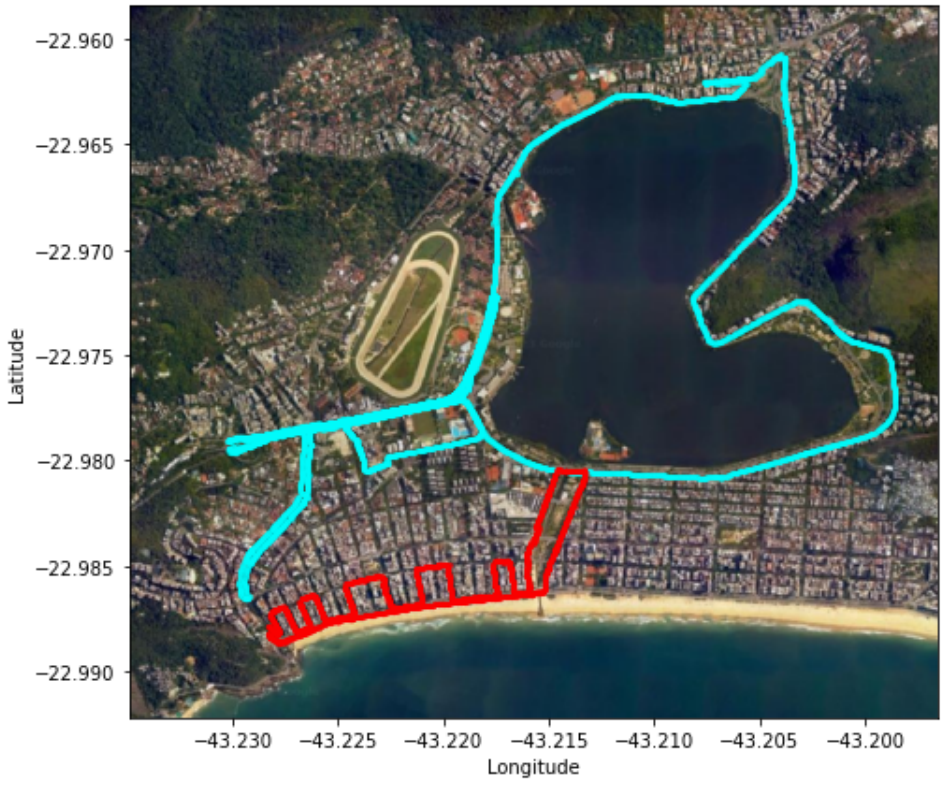

Figure 5.11: Measurements samples for the generalization capacity test for the sub-6 GHz macrocell environment. Samples in blue are used for training and in red for testing.

Table 5.14: Performance evaluation for the empirical models and GTB PL model trained using the instances from Route #1 and tested using the samples from Route #2 in the different frequencies for the sub-6 GHz macrocell environment.

Model	Frequency	RMSE [dB]	MAPE [%]	σ [dB]	R^2	$R^2{}_{\text{oos}}$
Log-distance	750 MHz	14.50	12.54	6.37	0.36	0.39
	2.5 GHz	14.46	11.13	6.20	0.37	0.37
	3.5 GHz	13.98	10.84	5.93	0.37	0.43
Okumura-Hata	750 MHz	14.18	12.40	7.22	0.39	0.41
	2.5 GHz	13.77	10.81	7.28	0.43	0.43
	3.5 GHz	14.78	11.15	9.10	0.29	0.29
GTB	750 MHz	8.42	6.32	5.43	0.79	0.79
	2.5 GHz	9.95	6.81	5.85	0.70	0.70
	3.5 GHz	9.90	7.03	5.71	0.68	0.68

From Table 5.14, it is noted that R^2 and $R^2{}_{\text{oos}}$ values are equal for the three frequencies, indicating very similar values of the mean of the measured PL on the training and testing set as seen in Table 5.3. The scatter plots between the measured and predicted PL for the ML and empirical models are shown in Figure 5.12. Furthermore, when examining the ranges of the values for the selected coalitions in Table 5.2, it is observed that certain predictor values on the testing set lie outside the range of values observed in the training set, as for b_{depth}, Δ_h, and d, although close. Hence, if the range of the predictor values on the unseen testing sets are within or close to the training inputs and output ranges, the model is expected to present good generalization, as discussed in [32].

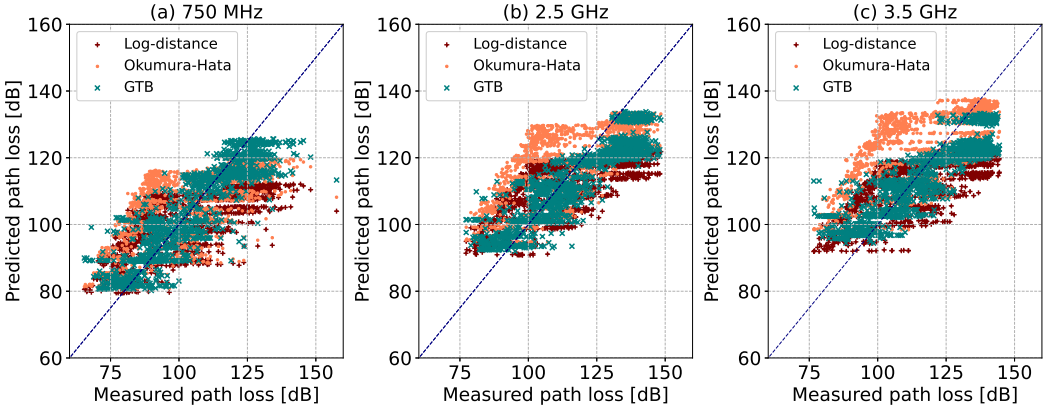

Figure 5.12: Comparison of measured and predicted path loss on the testing set for the ML and empirical models.

The results presented in [33] are displayed in Table 5.15. The adjusted model, derived from the sub-regions shown in Figure 5.1, is used to predict PL using all samples on both Routes at 750 MHz, 2.5 GHz, and 3.5 GHz. This

study performs a thorough comparison between the proposed PL model, the log-distance, and the Okumura-Hata models, as outlined in Table 5.15. While the study in [33] employed a small number of samples from both Routes for model adjustment, the approach we presented used samples from Route #1 for training and Route #2 for testing, i.e., a dataset crossed split methodology. As a result, a closer comparison between the models becomes possible by specifically examining the results for Route #2 in Table 5.15 and comparing them to the GTB model results provided in Table 5.14.

Table 5.15: Reported RMSE values in dB at [33] for the compared PL models for the Route #1 and Route #2, respectively.

Study in [33]	Route	750 MHz	2.5 GHz	3.5 GHz
Log-distance	#1	13.64	12.75	13.08
	#2	14.53	14.17	13.60
Okumura-Hata	#1	14.24	13.92	13.61
	#2	19.21	21.31	21.19
Proposed model	#1	12.69	12.55	13.94
	#2	14.20	13.19	14.20

The GTB model shows an outstanding generalization performance in comparison to the log-distance and Okumura-Hata models, as shown by the results in Table 5.14 and the study in [33] detailed in Table 5.15. This emphasizes the applicability of the ML-based PL model as an effective alternative for coverage planning in macrocell environments.

5.7 Discussion

This chapter has presented an optimized ML-based PL model design based on the profile terrain along the direct path between the transmitter and receiver for a macrocell coverage in sub-6 GHz considering a multi-frequency scenario. The results show that the GTB model can predict PL for multiple frequencies, antenna heights, and the macro conditions of the environment, such as building depth, vegetation depth, and the average ground height in the area. A higher dependency on the total number of intersected buildings is observed. Since the optimized GTB-based PL considers a variety of radio-path parameters such as multi-frequency, distance, antenna heights, and environment conditions, it is suitable for PL modeling in a macrocell coverage.

Examining the overall interpretability complexity of the final optimized GTB models in the mmWave and sub-6 GHz bands, we observe a lower level of complexity for the sub-6 GHz GTB model. This can be seen from the lower IAS and $\overline{\text{MEC}}$ values presented in Table 5.7 than the ones in Table 3.7. The

reduced complexity in the sub-6 GHz model can be attributed to more linear effects of the predictors on the response, more significantly for frequency and distance. Additionally, in the case of the lower frequencies (sub-6 GHz), the GTB model relies more on the individual effects of the predictors rather than on their interactions, as seen from the IAS values. Consequently, one can say that the ML-based PL prediction for higher frequencies, more specifically for mmWave in the indoor environment, is more complex to interpret.

The following chapter addresses dynamic scenarios faced by vehicular communications assessing V2I and V2V links PL prediction. Predictors from building, vegetation, and ground profiles are also used, and their effects on PL prediction – considering a lower transmission antenna height (as defined by the V2I link condition), are assessed for V2I links. Meanwhile, a satellite image extraction process is proposed to feed the input of a deep learning model based on CNN and transfer learning to predict the PL in the V2V scenario.

6
Path Loss Prediction for V2I and V2V using Machine Learning Techniques

In recent years, vehicular communications involving vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) scenarios have gained significant attention since they can contribute to improving vehicle road safety, reduce traffic congestion, and support applications such as autonomous vehicles, emergency services and traffic optimization, among others [10, 18]. However, the PL prediction for vehicular wireless channels presents a more challenging scenario than the traditional cellular channel [10]. In V2I, both the transmitter and receiver have low antenna heights, and additionally, the receiver or transmission can be in motion [14, 17]. In the V2V case, the channel is even more dynamic due to the relative movement between the transmitter and receiver vehicles, rapidly changing the surrounding environment [18, 27].

Therefore, existing objects, such as moving and stationary vehicles, buildings, and vegetation, affect the radio propagation [15, 28]; in particular, the presence of buildings poses a challenge to model PL due to multiple propagation mechanisms such as diffraction and reflections [15]. In conclusion, V2I and V2V channel propagation characteristics are influenced by the type of environment, vehicle speed, and road traffic density [12, 19].

This chapter addresses PL modeling for V2I and V2V wireless channels through a measurements campaign carried out in Rio de Janeiro, Brazil. For the V2I study, the profile environment is employed to design the ML-based PL models; the proposed methodology for coalition selection is applied to determine the optimal predictors for this vehicular channel, and the final model is compared with the log-distance model.

For the V2V environment, we use measured data considering the case when vehicles are driving in the same direction for training ML-based PL models. To design the ANN, SVR, RF, and GTB models, the distance and the speed between vehicles predictors are evaluated, as well as their decomposition into the vertical and horizontal components. In addition, as the environment highly affects path loss in a vehicular channel, we explore the potential of using satellite imagery to improve PL prediction using a convolutional neural network (CNN) for automatic predictor extraction.

6.1 PL Prediction for V2I

6.1.1 Related Work

Most V2I PL modeling studies have focused on empirical models in the 5 GHz frequency band. In [14], the Doppler effect is studied in an urban environment at 5.9 GHz, and the path loss is modeled with the ITU-R P.1411-5, two-ray and street canyon model for LOS conditions with the Tx-Rx distance ranging from 1 to 1,000 m. In [17], different antenna heights are tested in urban and suburban environments, and the two-ray and log-distance model are employed considering distances from 5 to 700 m, in the frequency band of 5.9 GHz. The estimated coefficients for L_o and n in the urban environment are equal to 44.8 and 1.98, respectively, for a transmitter height of 3.5 m, and 45.5 and 1.92, for a transmitter height of 1.5 m.

The work in [28] considers the 2.4 GHz band and studies the effect of trees in a rural area for varying heights of the receiver antenna. The results indicate that the trees considerably affect the received power for V2I communications; the experiments were performed using the ITU-R, FITU-R, and COST235 models. At a receiver antenna height equal to 5 m, the models attain σ values equal to 5.55 dB, 6.30 dB, and 9.32 dB, respectively. In [15], the authors addressed the impact of building obstruction for a V2I communication link at 2.4 GHz for an urban micro-cell. The experiments were compared with free-space loss, log-distance, and log-normal models. The log-normal model achieved the best performance with RMSE equal to 11.85 dB.

Limited studies in the literature concerning ML techniques address PL prediction for the V2I channel. In [13], angle spread and delay spread in a V2I urban environment at 28 GHz are predicted using a deep learning network; the ray-tracing technique is used to generate the simulated data employed, in order to capture the point-cloud information of the 3D scenario directly and map it with the channel characteristics using a deep learning model.

In this study, we investigate if using predictors extracted from the profile environment as the model's input can provide effective PL prediction in a V2I urban environment within the sub-6 GHz frequency band. Two Routes (streets) are considered, with one route allocated for training and the other for testing. The predictor coalition methodology of Subsection 3.9.1 is applied to the ML model leading to the best performance, and the final optimized ML model is compared with the log-distance model.

6.1.2 The Measurement Campaign: Dataset Description

We employ the data from a wideband measurement campaign carried out in Rio de Janeiro during 2020 conducted by colleagues from the radio propagation laboratory in CETUC and the Polytechnic Institute of Leiria, Portugal. The campaign considers using a low-height transmitter antenna to enable V2I communication with the receiver under real driving conditions. The measurements were performed at 735 MHz, 2.54 GHz, and 3.5 GHz frequencies in two Routes as depicted in Figure 6.1. In the first Route, the transmitter was located in Baixo Gávea, as represented by the white start, while in the second Route, it was positioned on João Borges Street. The transmitting antenna height was 3 meters from the ground while the receiver antenna height was set to 3.5 meters in both Routes. The environments are characterized by dense vegetation and building surroundings. The collected samples shown in Figure 6.1 correspond to the points leading to valid OFDM symbols after applying the constant false alarm rate (CFAR) filtering technique [156]; thus, the PL in dB is obtained from the measured received power (P_{RX}) and system parameters using Equation (3-1). This results in 756 samples on Route #1 and 825 samples on Route #2 in NLOS conditions.

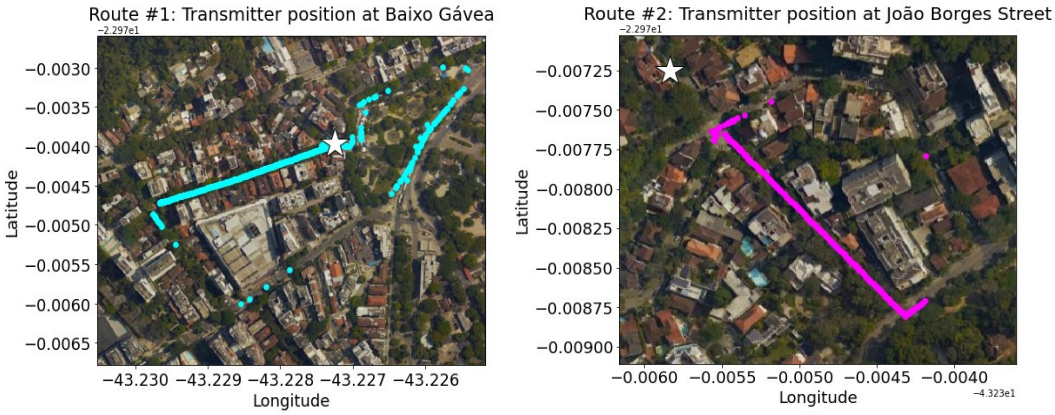

Figure 6.1: Measurement scenarios for V2I.

Route #1 contains 198 samples at the frequency of 735 MHz, 289 samples at 2.54 GHz, and 269 samples at 3.5 GHz. On the other hand, Route #2 consists of 353 samples at 735 MHz, 235 samples at 2.54 GHz, and 237 samples at 3.5 GHz. Predictors are extracted from the profile environment to design the ML-based PL model for the V2I case. In total, 12 predictors are considered, including the frequency (f) in MHz, Tx-Rx distance (d) in meters, mean ($\overline{b}_h$, $\overline{v}_h$, $\overline{g}_h$), and standard deviation (σ_b, σ_v, σ_g) of building, vegetation, and ground height, in meters. Additionally, the building depth (b_{depth}), vegetation depth

(v_{depth}), and the number of intersections of building and vegetation (n_b, n_v) along the path from the transmitter to the receiver are also included.

Table 6.1 provides an overview of the predictor value ranges in both Routes. The predictor value ranges in Route #1 encompass the ones in Route #2, except for the predictor n_v. Additionally, Table 6.2 presents the variance σ^2 and mean values of the measured PL values at each frequency. For the frequency of 735 MHz, the mean values are similar in both Routes. However, for the frequencies of 2.54 GHz and 3.5 GHz, Route #2 shows a higher mean accompanied by lower σ^2. This difference in the statistical distribution pattern can be assessed in the CDF curves presented in Figure 6.2. In addition, for the frequencies 2.54 GHz and 3.5 GHz, Route #2 shows a narrower range of measured PL values, ranging from 60.52 dB to 81.02 dB and from 68.34 dB to 90.65 dB, respectively.

Table 6.1: Ranges of predictor values in Route #1 and Route #2 in the V2I scenario.

Predictor	Route #1	Route #2
d	1.26 - 275.48 m	51.06 - 233.90 m
f	735 MHz - 3.5 GHz	735 MHz - 3.5 GHz
σ_b	2.17 - 18.66 m	7.32 - 18.66 m
σ_v	0 - 11.81 m	5.01 - 11.81 m
σ_g	0.01 - 3.26	0.32 - 0.73 m
$\overline{b}_h$	0 - 19.06 m	5.54 - 19.06 m
$\overline{v}_h$	0 - 19.10 m	1.23 - 19.10 m
$\overline{g}_h$	0 - 153.02 m	16.86 - 17.92 m
b_{depth}	0 - 188.92 m	1.06 - 109.79 m
v_{depth}	0 - 275.48 m	0 - 112.48 m
n_b	0 - 16	1 - 14
n_v	0 - 8	0 - 14

Table 6.2: Variance and mean of the measured PL in Route #1 and Route #2.

	σ^2 of measured PL [dB^2]			mean of measured PL [dB]		
	735 MHz	2.54 GHz	3.5 GHz	735 MHz	2.54 GHz	3.5 GHz
Route #1	128.71	110.14	108.99	39.17	62.12	69.96
Route #2	66.07	21.71	29.87	34.97	76.05	84.09

Figure 6.3 presents the scatter plots between path loss and distance in the different Routes and frequencies. The plots show that the path loss increases for higher frequencies. Additionally, when the distance is small, reflections from the surrounding environment significantly influence the power of the received signal, dispersing the path loss to higher values as reported in [17].

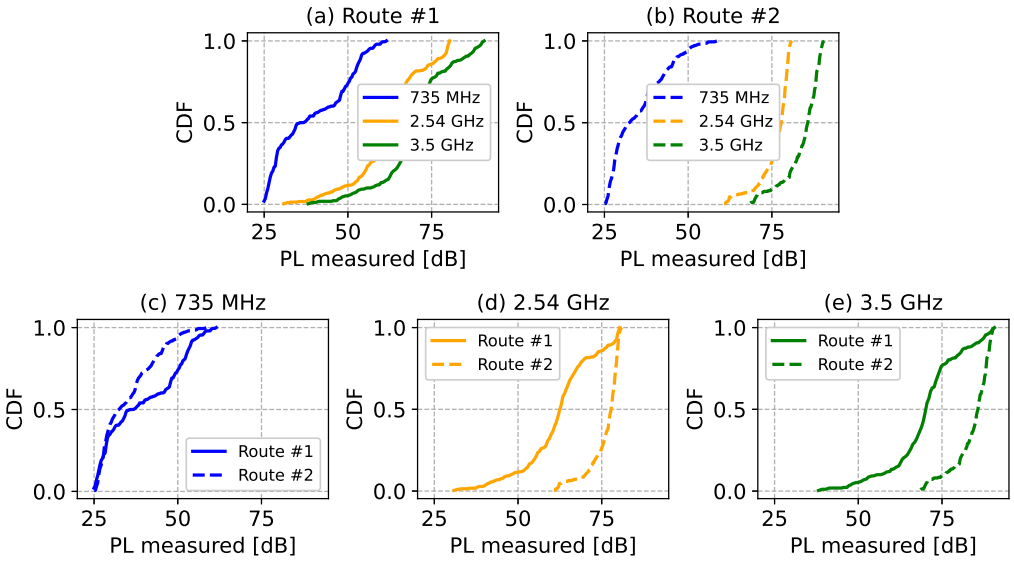

Figure 6.2: CDF of the measured PL in Route #1 and Route #2 in the V2I scenario.

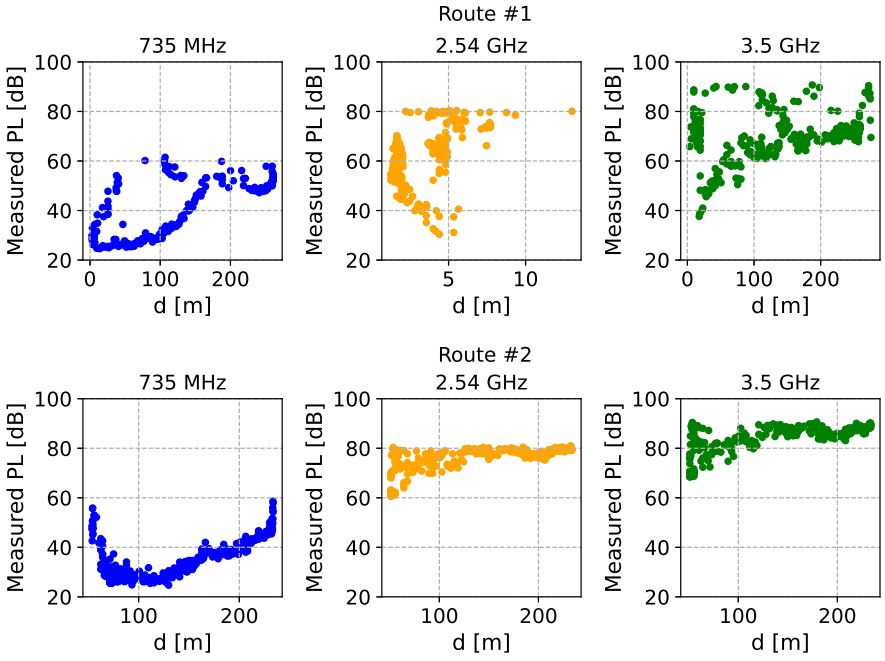

Figure 6.3: Scatter plot of path loss versus distance for the Route #1 and Route #2 at each frequency. For Route #2, the variability of PL with distance is lower, more notably in the frequency of 2.54 GHz and 3.5 GHz in the distance higher than 100 m.

6.1.3 ML-based PL Model Design

In this study, we select the samples from Route #1 for training and the samples from Route #2 for testing, aiming at the practical application of the models. The 5-fold CV technique is employed to identify the optimal hyperparameters while considering all predictors for each ML model on the training set. During the grid search, the ANN model considers the same hyperparameters range values employed in the macrocell environment, Chapter 5, that include a number of neurons varying from 1 to 20, learning rate and weight decay from 0.001 to 0.1, and activation functions ReLU, Logistic, and Tanh. Using the Logistic activation function, the lowest CV RMSE (4.55 dB) is achieved with 11 neurons, a learning rate equal to 0.01, and a weight decay of 0.1.

For the SVR, the hyperparameters C, ϵ, and σ_{RBF} are explored with the same hyperparameters values described in Chapter 5. C is varied from 1 to 100, ϵ ranges from 0.0001 to 0.01, and σ_{RBF} spans from 0.0002 to 0.02. The lowest CV RMSE (5.49 dB) is attained with C, ϵ, and σ_{RBF} equal to 100, 0.01, and 0.02, respectively. For the RF model, the number of trees varies from 2 to 70. For this vehicular scenario, we employ a higher range of values for the hyperparameters maximum depth and minimum number of samples to assess their effect in the model's prediction. Thus, the maximum depth varies from 2 to 12, and the minimum number of samples per leaf varies from 2 to 10 with the squared error and absolute error as loss functions. The lowest CV RMSE (5.99 dB) is attained using 34 trees, with a maximum depth of 2, a minimum number of samples of 10, and the squared error loss function.

The grid search for the GTB considers the same ranges for building the trees as the RF, with a learning rate equal to 0.14 and squared error and absolute error loss functions. The lowest CV RMSE (4.88 dB) is reached using 70 trees, with a maximum depth of 8, minimum samples of 10, and the absolute error loss function. The results for the final ML models are presented in Table 6.3.

One notes that the models cannot generalize well with a notably high over-fitting. One possible reason for that is because the training and testing sets are disjointed (different range values for the predictors and also for the output), as can be seen in Table 6.1 and Figure 6.2, and as also discussed in [40]. Path loss and distance differ in each Route, as reflected in the PL values in Figure 6.3, more markedly for the frequencies at 2.5 GHz and 3.5 GHz. Although the RF model presents the closest performance values between the training and testing sets, the GTB model performs better, as observed

Table 6.3: Performance evaluation of the PL-based ML models measured on the training and testing set for the V2I case.

	Training set				Testing set			
Model	RMSE [dB]	MAPE [%]	σ [dB]	R^2	RMSE [dB]	MAPE [%]	σ [dB]	R^2
ANN	3.48	4.32	2.55	0.95	28.80	37.95	17.36	-0.51
SVR	4.56	5.11	3.66	0.92	35.79	75.23	23.52	-1.33
RF	8.85	11.23	6.49	0.70	15.20	25.55	6.93	0.58
GTB	3.58	3.09	3.18	0.95	12.99	23.82	6.40	0.69

in the evaluation of the testing set. The predictor selection methodology is then employed in the GTB model to address over-fitting and identify the most significant predictors for further generalization testing.

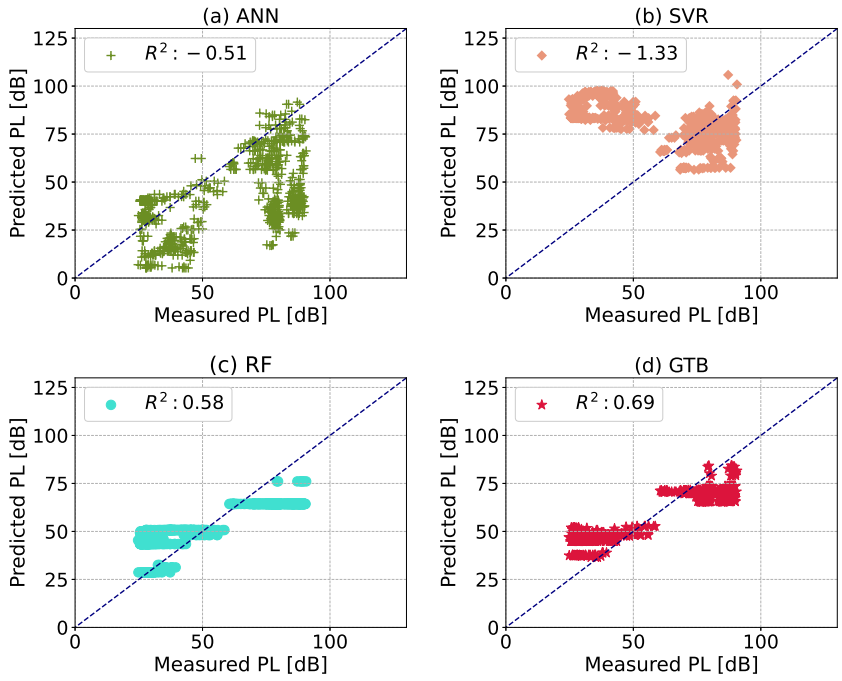

Figure 6.4: Comparison of measured and predicted path losses on the testing set (Route #2) for the ML models in the V2I environment.

6.1.4 Predictor Selection

Following the methodology in Subsection 3.9.1, the resulting coalitions of predictors for the GTB PL model are shown in Table 6.4. Upon examining the predictor's coalition formed, one observes that the performance prediction on the training set improves steadily as new predictors are included. This

improvement continues until reaching the final subset, which incorporates all predictors. Also, from Figure 6.6, it is noted that the higher marginal contribution (ΔRMSE_p) occurs for the subset of seven predictors (f,n_b,σ_b,n_v,σ_v,σ_g,d), beyond this subset, the marginal contribution significantly reduces. There is a risk of producing over-fitting by incorporating new predictors, as reported by the $\overline{\text{MEC}}$ values, which increase for more than seven predictors, indicating a more complex model. The increase of the IAS value to 0.43 for the subset of seven predictors suggests that the model improvements depend on the interaction with the predictor d, also seen when including b_{depth} (0.45). The same behavior is not as strong when including other predictors in the coalition. Furthermore, from the ALE curves shown in Figure 6.5, it can be seen that the coalition with seven predictors presents quite linear effects on the predicted PL.

Table 6.4: Performance indicators for the GTB model for the V2I case using different subsets measured on the training set. Model performance is measured using RMSE, MAPE, σ and R^2 and the interpretability is measured with IAS and $\overline{\text{MEC}}$.

Predictors coalitions	RMSE [dB]	MAPE [%]	σ [dB]	R^2	IAS	$\overline{\text{MEC}}$
f	10.87	16.29	6.97	0.55	-	1.00
f,n_b	9.24	11.69	6.62	0.68	0.14	1.23
f,n_b,σ_b	7.11	6.87	5.96	0.81	0.26	2.09
f,n_b,σ_b,n_v	6.26	5.93	5.26	0.85	0.27	1.78
f,n_b,σ_b,n_v,σ_v	5.89	5.42	5.07	0.87	0.27	1.79
f,n_b,σ_b,n_v,σ_v,σ_g	5.54	4.83	4.85	0.88	0.26	1.74
f,n_b,σ_b,n_v,σ_v,σ_g,d	4.45	3.77	4.00	0.92	0.43	1.87
f,n_b,σ_b,n_v,σ_v,σ_g,d,b_{depth}	4.07	3.46	3.65	0.94	0.45	1.94
f,n_b,σ_b,n_v,σ_v,σ_g,d,b_{depth},$\overline{g}_h$	3.99	3.35	3.59	0.94	0.27	2.25
f,n_b,σ_b,n_v,σ_v,σ_g,d,b_{depth},$\overline{g}_h$,$\overline{b}_h$	4.03	3.41	3.62	0.94	0.26	2.60
f,n_b,σ_b,n_v,σ_v,σ_g,d,b_{depth},$\overline{g}_h$,$\overline{b}_h$,$\overline{v}$	3.70	3.08	3.20	0.95	0.25	2.57
f,n_b,σ_b,n_v,σ_v,σ_g,d,b_{depth},$\overline{g}_h$,$\overline{b}_h$,$\overline{v}_h$,v_{depth}	3.58	3.09	3.18	0.95	0.25	2.77

Using the subset of predictors f, n_b, σ_b, n_v, σ_v, σ_g, and d, the GTB model achieves an RMSE, MAPE, σ, and R^2 on the testing set equal to 8.24 dB, 14.87 %, 4.45 dB and 0.88, respectively. The results indicate a significant reduction in over-fitting by employing a smaller subset of predictors. In addition, the PFI values of the selected subset of predictors are shown in Table 6.5. The results show that the predictors f, d, and σ_v present the highest influence for improving the predicted PL in the V2I environment. Meanwhile, the predictors n_b, σ_g, and σ_b have lower but similar importance values, which emphasize that

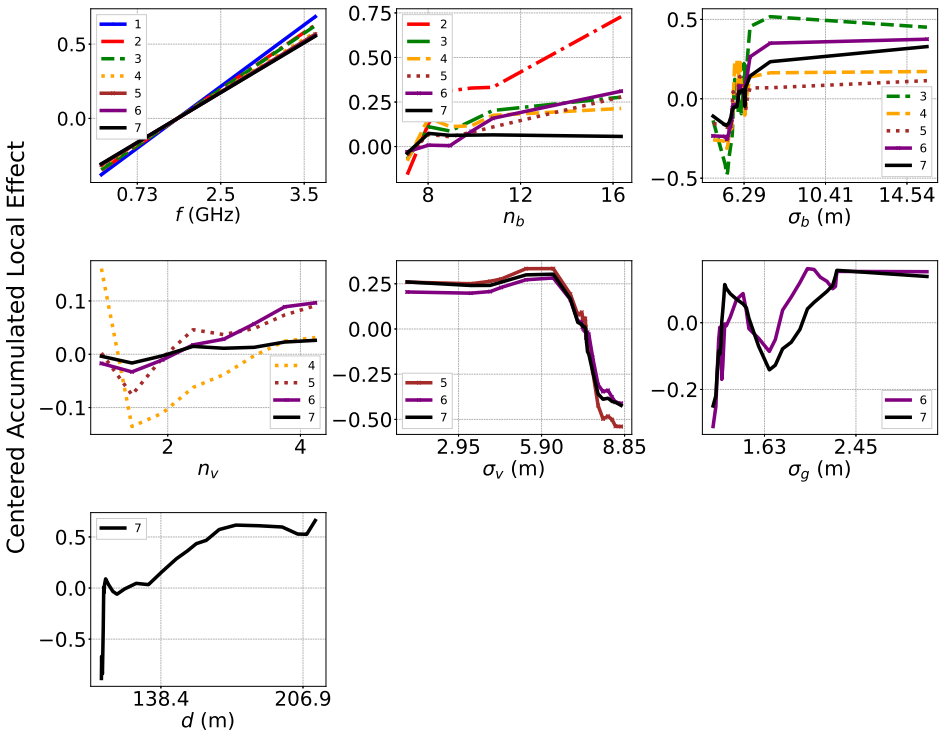

Figure 6.5: ALE plots for the GTB model including the subset with one predictor (f) until the selected subset (f, n_b, σ_b, n_v, σ_v, σ_g, d).

knowing the building and ground profiles also contribute to improving PL prediction; at last, the lowest PFI value is observed for n_v.

Table 6.5: PFI ranking in increasing order of the selected subset of predictors shown.

Predictor	PFI value (loss: RMSE [dB])
n_v	0.13
σ_b	1.24
σ_g	1.86
n_b	1.98
σ_v	3.47
d	7.34
f	14.32

6.1.5 Final Models Comparison

In this subsection, the performance of the GTB regressor is compared with the log-distance model obtained from Equation (5-1). The 5-fold CV technique is applied to find the optimal hyperparameters for the selected subset

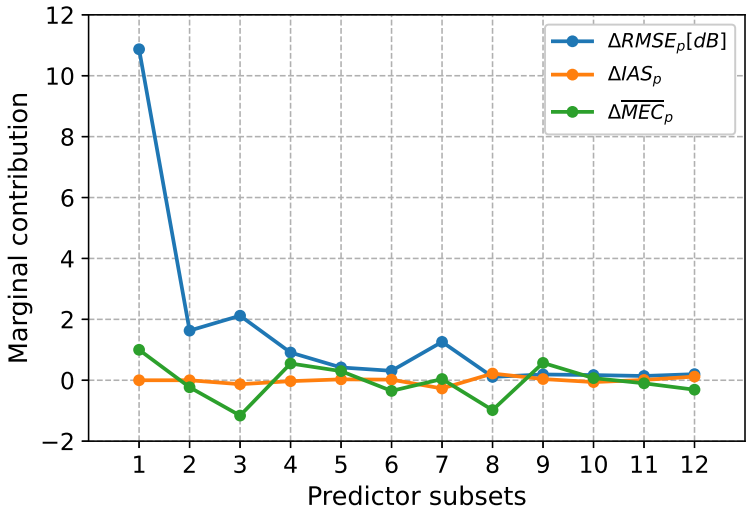

Figure 6.6: Marginal contributions in performance and interpretability for the GTB-based PL model (for the predictors coalitions see Table 6.4).

of predictors. The number of trees in the grid search varies from 2 to 70, the maximum depth varies from 2 to 12, with the squared error and absolute error as loss functions. The minimum number of samples per leaf in this case varies from 10 to 20, to address over-fitting as discussed in [79]. The lowest RMSE value (5.66 dB) is obtained using 70 trees, maximum depth equal to 10 and minimum samples leaf equal to 14, for a learning rate of 0.14 and the absolute error loss function. The final model reaches an RMSE, MAPE, σ and R^2 on the training set of 4.46 dB, 3.80%, 4.00 dB, respectively, and 0.92, and an RMSE, MAPE, σ and R^2 on the testing set equal to 6.33 dB, 10.22%, 3.79 dB and 0.93, respectively.

To compare against the log-distance model, first, its coefficients L_o and n are estimated, and the resulting values are shown in Table 6.6. Values of the path loss exponent n lower than 2 indicate better propagation than free space conditions as a result of constructive interference due to multi-path [19]. For the V2I environment, this is expected due to the existence of building faces on both sides of the street [14].

Table 6.6: Coefficient values for the log-distance model for each frequency, estimated over the training set.

Coefficients	735 MHz	2.54 GHz	3.5 GHz
L_o	14.09	54.72	61.74
n	1.34	1.77	0.43

The performance indicators of the ML-based and log-distance PL models for each frequency on the testing set are shown in Table 6.7. The results show

a significant improvement in the GTB model over the log-distance model at the three frequencies. The GTB model achieves the highest R^2 value of 0.37 at 735 MHz, followed by an R^2 value of 0.14 at 2.54 GHz. However, the GTB model does not generalize well for the 3.5 GHz data, as shown in the R^2 value (-0.97). The log-distance, model also performs better for the subset sample at 735 MHz and smaller performance at higher frequencies.

Table 6.7: Performance evaluation for the testing set at the different frequencies for the GTB and log-distance PL models.

Model	Frequency	RMSE [dB]	MAPE [%]	σ [dB]	R^2	R^2_{oos}
Log-distance	735 MHz	9.88	28.57	4.50	-0.48	-0.17
	2.54 GHz	15.91	20.63	3.67	-10.66	-0.17
	3.5 GHz	14.26	15.63	4.80	-5.80	0.11
GTB	735 MHz	6.46	15.74	3.74	0.37	0.50
	2.54 GHz	4.32	4.01	3.26	0.14	0.91
	3.5 GHz	7.66	8.18	3.13	-0.97	0.74

In addition, Table 6.7 presents a marked difference between the R^2 and $R^2{}_{\text{oos}}$ values for the frequency at 2.54 GHz and 3.5 GHz, suggesting that the distribution of the measured PL samples differ in terms of variance and mean between the training and testing sets – as analyzed in Table 6.2, and further visualized in Figure 6.2 and Figure 6.3. The comparison between the predicted and measured path loss for each frequency are shown in Figure 6.7.

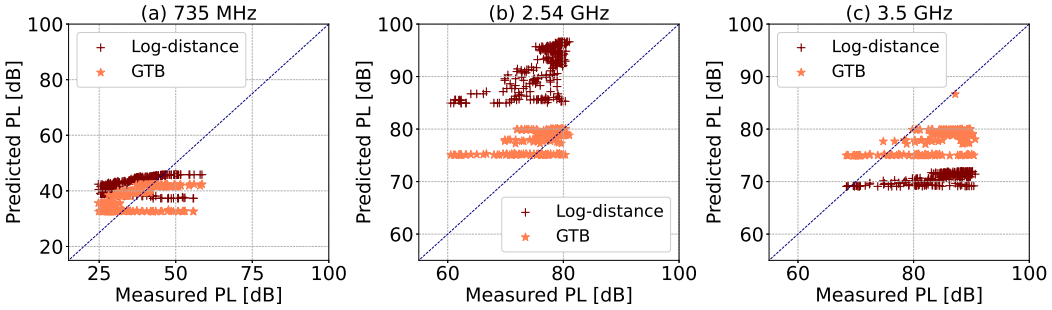

Figure 6.7: Scatter plots between the predicted and measured data for each frequency for the log-distance and GTB model.

6.2 PL Prediction for V2V

6.2.1 Related Work

Traditional approaches for PL prediction in V2V links consider the log-distance model based on measurement campaigns [10, 19]. In [10], an extensive overview of channel characteristics for highway, urban, suburban, and rural V2V communications is presented. Considering a suburban area of the city of Valencia, Spain, the authors of [19] analyze the influence of Tx-Rx distance in the prediction of the PL in V2V links at 5.9 GHz; authors report that the path loss exponent in the log-distance model ranges from 1.83 to 3.27, and σ from 2.57 dB to 4.55 dB.

Most recently, ML techniques have also been explored for channel modeling in V2V scenarios [11, 18, 157]. For example, in [157] the power delay profile (PDP) and PL are the predictors for estimating the angle-of-arrival using an SVR model in V2V channels. In [11], the vehicular channel parameters (PL, shadow fading, RMS-DS, and K-factor) extracted from measurement campaigns in different scenarios are used to train an ANN to classify the vehicular scenario.

In [18], the authors propose using ANN, RF, and CNN for PL and root mean squared delay spread (RMS DS) prediction in a V2V urban environment. Ray tracing technique was used to simulate data, in a total of 24,500 instances, considering LOS and NLOS conditions. The three ML models employed twelve predictors related to Tx and Rx positions: Tx-Rx distance, number of intersected buildings, street widths at the Tx and Rx positions, link type (LOS/NLOS), and the speed of vehicles. The best PL model was the RF model with an RMSE equal to 0.63 dB in LOS and to 0.44 dB in NLOS condition.

A comparison of the performances of the RF and the log-distance models for V2V PL prediction is presented in [70] using data from a measurement campaign held in an urban environment. The measurements were collected at the same location during day and night, and 10,000 samples are used for training and 3,000 for testing. Four predictors were employed including the transmitter and receiver positions, Tx-Rx distance, link type (LOS/NLOS), and different subsets of these predictors were also used. The performance evaluation used the mean absolute error and standard absolute error as indicators. Concerning PL models based on deep learning techniques, some studies propose CNN models primarily focusing on urban environments. These models aim to learn patterns extracted from images such as the distance

between transmitter and receiver and the density of buildings and vegetation surrounding the pair of Tx and Rx [18, 52, 55, 111, 112]. In [52], some layers provided by a geographic information system (GIS) enter the CNN model; they include the land clutter classification, terrain, and building. In addition, data from antenna height, azimuth, tilt, and antenna gain are also employed. To generate the dataset, ray-tracing techniques and measurement campaigns were combined, covering 40 cities worldwide, resulting in around 800,000 simulated samples and 400,000 field measurements.

The work in [55] used a CNN for PL modeling for the frequency of 900 MHz. The dataset contained more than 35,000 instances from 146 urban environments, which were generated using ray-tracing techniques. Building height was used to construct the input image. For each PL instance, a square patch image was assigned, taking into account the region bounded from the transmitter and the receiver between Tx and Rx. Various sizes of input images were tested, ranging from 16x16 to 256x256 pixels. The dataset was randomly split into 80%/20% for training/testing.

In [112], the authors propose using satellite images extracted from the Mapbox API for PL prediction at 811 MHz and 2630 MHz frequencies. Ray-tracing simulations with urban and suburban characteristics were conducted at the Technical University of Denmark. Each PL sample was associated with a square patch image centered in the receiver position. The simulated data consisted of 60,000 data points from which some Routes on the campus were used for training and the remaining for testing. In addition, data augmentation was used to increase the number of images during training. The work in [120] proposes a CNN model for operation at the UHF frequency band. The authors proposed concatenating four satellite images tiled to extract four different positions along the Tx-Rx path.

The previously cited works dealing with CNN models for PL prediction are trained from scratch. They are based on large datasets generated via ray-tracing simulations; the need for large datasets derives from the many parameters that need to be learned when training these models. However, one of the significant challenges in developing a CNN model for radio-propagation prediction is the usually small cardinality of the datasets available.

To tackle this challenge, our study employs a strategy known as transfer learning [106]. This approach entails using a pre-trained model on a large dataset considering a given task and applying it to a new task as explained in Subsubsection 2.3.5. In the context of radio propagation, the work in [114] proposed the use of a pre-trained model for path loss exponent (n) and shadowing factor classification at 900 MHz for an urban scenario. However, to

the best of our knowledge, no works address the use of CNNs for PL prediction using pre-trained models. We employ transfer learning to develop a PL model for V2V using real-world V2V measurements. The proposed methodology for image extraction is explained in Subsubsection 6.2.4. Subsequently, the performance of the proposed CNN-based PL model is compared with the ANN, SVR, GTB, and log-distance models.

6.2.2 The Measurement Campaign: Dataset Description

The measurement campaign used for this study was carried out in neighbor of Jardim Oceânico, Barra da Tijuca, Rio de Janeiro, in November 2017, conducted by colleagues of the radio propagation laboratory of CETUC. The vehicles move on a Route of high traffic density and are surrounded by buildings, trees, vegetation, and other obstructing objects. The Tx and Rx vehicles moved in the same directions with approximation and estrangement conditions, thus mostly in line-of-sight (LOS). The LOS condition considers a scenario without an obstacle vehicle between Tx and Rx.

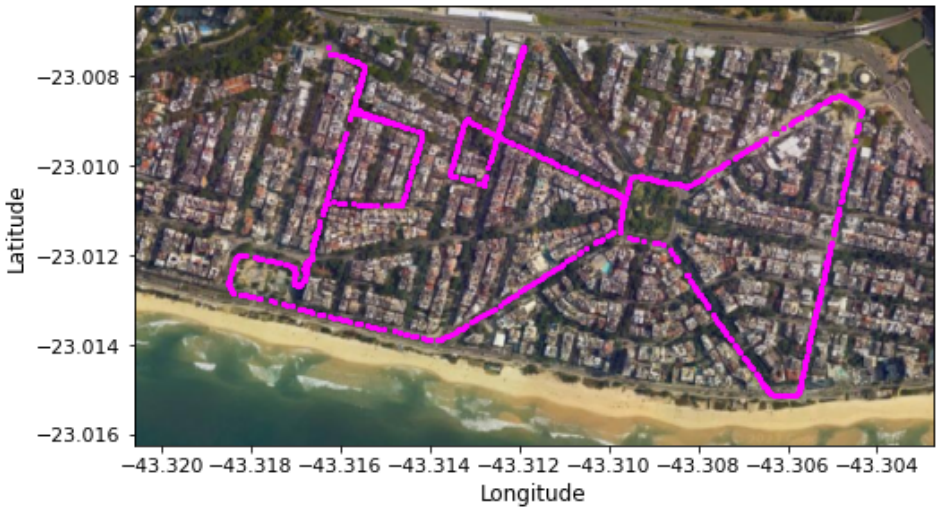

Figure 6.8: Positions of the collected samples in the V2V measurement campaign.

A signal generator from Anritsu model MG3700A was used to transmit orthogonal frequency division multiplexing (OFDM) symbols with a Zadoff-Chu (ZC) sequence at 10 dBm. The receiver is a signal analyzer from Anritsu, model MS2692A. Ten dBi gain omnidirectional antennas were used in the transmitter and the receiver with heights of 1.7 meters from the ground. The number of received multi-paths and their amplitudes were characterized using tools developed in MATLAB for the data post-processing [158]. Finally, the PL in dB is obtained from the measured received power (P_{RX}) and system

parameters using Equation (3-1). The campaign provides a collection of 1,374 PL instances. Each instance has five numerical attributes associated with it: the Tx-Rx distance (d) in meters and its decomposition into horizontal (d_x) and vertical (d_y) components, the relative speed between the vehicles (v) in meters/second and its decomposition into horizontal (v_x) and vertical (v_y) components. The predictor d varies from 3.31 to 176.54 meters, and the predictor v varies from 0 to 10.79 meters/second. The scatter plot between the Tx-Rx distance and path loss is shown in Figure 6.9.

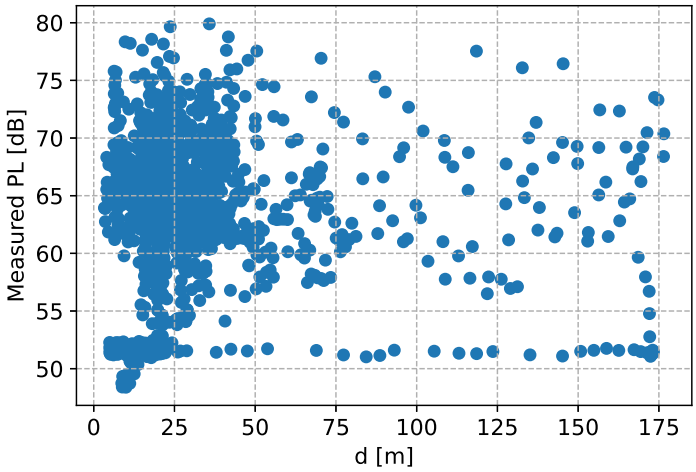

Figure 6.9: Scatter plot between path loss and distance for the V2V scenario.

6.2.3 Proposed CNN-based PL model

To address the limitation of the small dataset when designing a CNN model for PL prediction, we propose using a pre-trained CNN model as an automatic predictor extractor to capture the environment characteristics between the transmitter and receiver positions. In the literature review, we initially identified the pre-trained residual network (ResNet) capability to perform land use classification using satellite images [106, 115, 159]. The ResNet architecture has multiple variations, including ResNet16, ResNet18, ResNet50, ResNet110, and ResNet164 [115]. The number in the architecture name indicates the number of layers that are employed.

Land use datasets often contain images with various elements such as objects, buildings, and vegetation [115]. Meanwhile, the work in [115] uses two pre-trained CNN architectures, ResNet and dense convolutional network (DenseNet), to solve the land use classification considering two datasets: the FMOW (470,086 images and 63 classes) and the NWPU-RESISC45 (31,500

images JPG and 45 classes with a resolution of 256 x 256 pixels). In [159], the EuroSAT dataset was used in the ResNet50 and GoogLeNet for land use classification. The dataset consists of 27,000 satellite images from European cities from 34 countries, obtained from the satellite mission Sentinel-2 (the images present a resolution of 20 meters). The results in [159] show that even though the pre-trained network was trained on images from a totally different domain (ImageNet dataset), the pre-trained model generalized well. Given the good results of adapting pre-trained models for tasks like land use classification, and since satellite imagery has demonstrated its efficacy as input for PL prediction as in [112, 120], we propose using the pre-trained ResNet18 for designing a PL model.

Figure 6.10 shows the architecture of the ResNet18. There are 18 layers in this network (17 convolutional layers and a fully connected layer, additionally a Soft-max layer performs the classification task). The model uses a residual learning method for training, since deeper layers can results in a degradation of the output [113]. Thus, the architecture employs shortcut connections in a residual block framework that facilitates the optimization of the overall network [113]. Lastly, the layers are stacked to learn a residual mapping (function) to solve the issue of accuracy degradation in deeper networks [113]. The ResNet18 has been originally trained on the ImageNet dataset containing 14 million images and 1,000 classes.

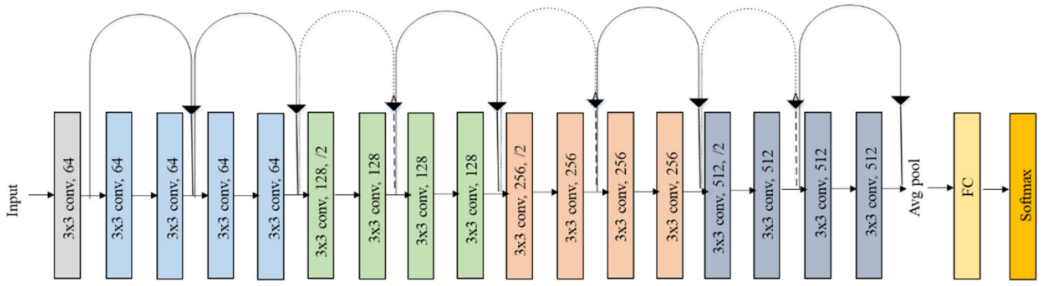

Figure 6.10: The ResNet18 architecture using 18 layers based on a residual learning framework [113].

6.2.3.1 The ResNet18

The main hyperparameters defining the CNN consist of the kernel size, the number of kernels, and the activation functions. In the pooling layer, the important hyperparameters are the pooling method, and the filter size; and, in the fully connected layer, they are the number of hidden layers and number of neurons in each layer, and activation function. Other hyperparameters

are the optimizer, the learning rate, the loss function, batch size, and regularization [117, 118]. The evaluation of the pre-trained CNN model to solve the land use classification is presented in Appendix C. As seen in the results, the predictor extractor of the pre-trained ResNet18 can extract patterns from the satellite images. We adapt the pre-trained ResNet18 model for PL prediction in the V2V environment. The feature extractor is performed by the convolutional layers while prediction is done by the regression layers.

6.2.4 Proposed Methodology for Image Extraction

To enable detailed observation of buildings, vegetation, and other structures that significantly impact the magnitude and variability of the receiver power (e.g., large-scale fading) high-resolution images are necessary [112]. Thus, in our image extraction process, we use satellite images from Google Earth (GE) and obtain their georeferencing using the QGIS software. The methodology for clipping each patch image is visualized in Figure 6.11. The main image (see Figure 6.11.(a)) was extracted from GE using a zoom with an altitude eye of 1.61 Km. That zoom allowed to georeference the image with 0.48 meters per pixel using the QGIS Georeferencer tool [160]. In that tool, different ground control points are used to find common points between the extracted image and a defined GIS layer (GE map) to assist in the georeferencing process using a re-sampling method for extrapolation.

To extract the patch image having the transmitter and receiver positions and given the shorter Tx-Rx distance in this V2V environment, at most 175 m – as can be seen in Figure 6.9, we calculate the average latitude and longitude coordinates between the Tx and Rx. The height and width of the cropped image are determined based on the Tx-Rx distance as shown in Figure 6.11.(c), this allows capturing the propagation environment, i.e., the path between transmitter and receiver sites. The patch image is then automatically cropped from the extracted satellite image (see Figure 6.11.(a)) using the Python libraries Pyproj and GDAL [161]. Finally, each image is resized to a square of 224×224 pixels that feeds the ResNet18 model. Some samples of clipped patch images for our V2V dataset are shown in Figure 6.12.

Figure 6.13 shows the proposed CNN-based PL model and methodologies used for the experiments, which comprises the pre-trained ResNet18 as the predictors (features) extractor and a fully connected layer that receives an input vector of dimensions 512×1. The architecture employs a single hidden layer with 256 neurons before the output neuron computing the PL for V2V. To train and test the proposed architecture, we employ a training set containing

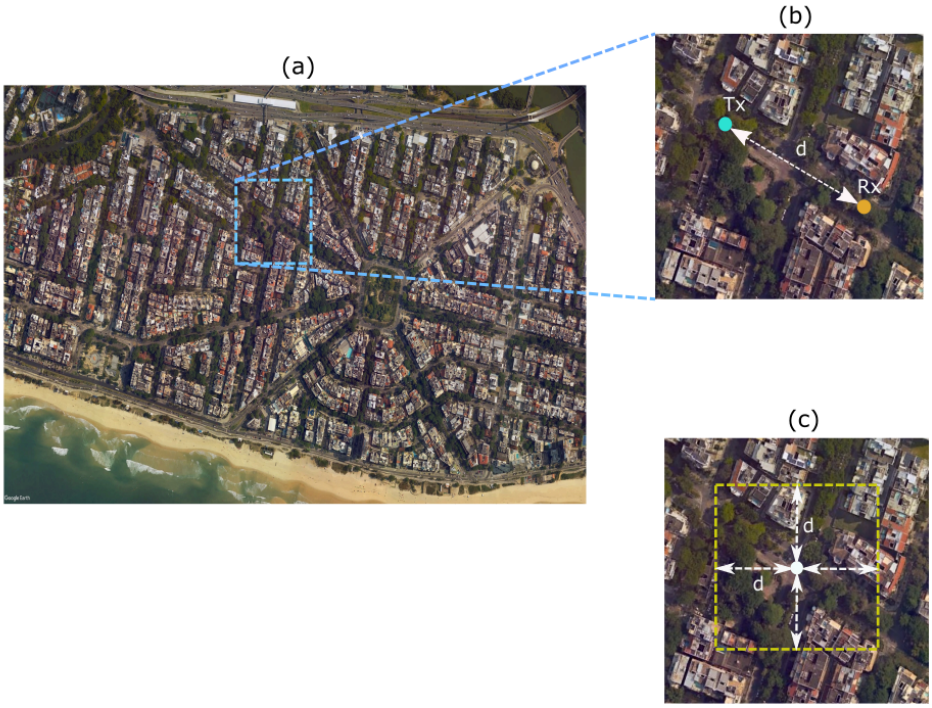

Figure 6.11: Obtaining the images containing the Tx and Rx sites and the link between them: (a) georeferenced image, (b) example to extract the Tx-Rx patch image, and (c) average latitude and longitude coordinates between the Tx and Rx. The yellow dashed line identifies the final square cropped from the image.

Figure 6.12: Some examples of clipped patch images considering different Tx-Rx distances. As seen in the images, the proposed clipping technique captures the effect of the distance between the transmitter and receiver on the clipped patch. The Tx-Rx distance samples range from 12.03 to 103.50 m.

1,000 samples and the remaining instances (374) for the testing set from the measurement campaign described in Subsection 6.2.2. The sites of the vehicles in the two subsets are displayed in Figure 6.14.

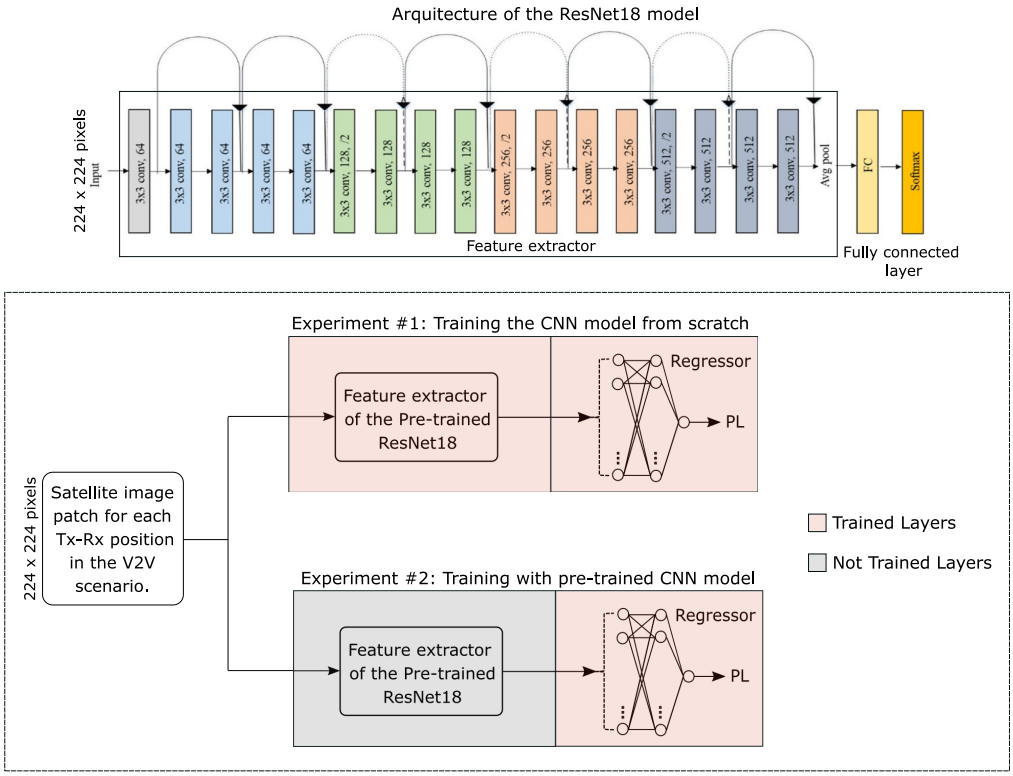

Figure 6.13: Experiments for PL prediction using the ResNet18. Two training approaches are tested: the firts where the model is trained from scratch (experiment #1), and the second where only the regression layers are trained (experiment #2).

Figure 6.14: The points in blue are used for training and the samples in red are used for testing.

The experiments were carried out in Python with the Pytorch framework on a workstation with an Intel Core i7 8th Gen processor, 64 GB of RAM, and the GPU GeForce RTX 2060. For the hyperparameters setting, the 5-fold CV technique is applied on the training set, varying the hyperparameters as shown in Table 6.8. The best performance is provided by the learning rate equal to 0.001, with a drop-out coefficient of 0.2, batch size of 15, Adam optimizer, ReLU activation function, mean squared error loss function, with early-stopping. The measured PL is normalized using the MinMaxScaler to scale the output in the range of 0 to 1.

Table 6.8: Hyperparameters range for the proposed CNN for path loss modeling.

Category	Value
Learning rate	0.001 - 0.005
Dropout rate	0.1 - 0.5
Batch size	10 - 30
Optimizer	Adam/AdamW/SGD
Activation function	ReLU/Tanh

We also train the model from scratch to assess if there is a gain in using a pre-trained CNN model and how large it is. Figure 6.15 compares the RMSE of the predicted PL using the pre-trained ResNet18 and training it from scratch. The training curves demonstrate that using the pre-trained CNN model presents consistently better performance than training the model from scratch. Besides, when using the pre-trained CNN, one reaches more rapidly the lower and stable region of the RMSE curve.

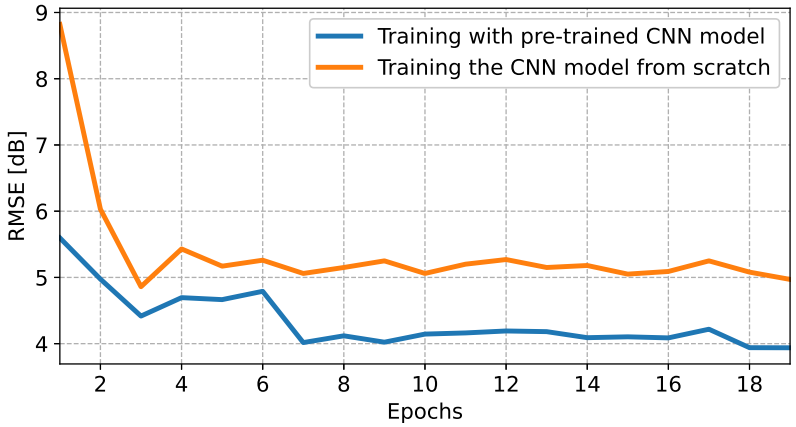

Figure 6.15: RMSE curves for the PL during training when using the pre-trained ResNet18 model and training the ResNet18 from scratch.

6.2.5 Final Model Comparison

In this subsection, we compare the ML models for V2V PL prediction. The ML models trained using tabular data – including ANN, SVR, RF, and GTB, consider two predictor coalitions. The first subset includes the predictors d and v, while the second subset uses its components d_x, d_y, v_x, and v_y. The 5-fold CV technique is employed to find the optimal hyperparameters for each ML model. The ranges of the hyperparameters in the four ML models are those described in Subsection 6.1.3.

The ANN using the subset of two predictors attains the lowest CV RMSE (4.93 dB) using 11 neurons, a learning rate equal to 0.1, a weight decay of 0.001 with the ReLU activation function. When using four predictors, the optimal hyperparameters (providing RMSE equal to 4.79 dB) are 16 neurons, with a learning rate equal to 0.1, and a weight decay of 0.1 with the ReLU activation function. Using two predictors, the SVR reaches the lowest CV RMSE (5.45 dB) for C equal to 75, ϵ to 0.0001 and σ_{RBF} to 0.02; and, using four predictors, C is equal to 65, ϵ to 0.0001 and σ_{RBF} to 0.02 leading to (5.37 dB).

The two-predictors RF presents the lowest CV RMSE (4.85 dB) using 14 trees, with a maximum depth of 3, minimum samples of 10, and the squared error loss function; meanwhile, when four predictors are employed, the best performance (4.74 dB) is obtained using 12 trees, the maximum depth of 4, minimum samples of 10 and the squared error loss function. When the GTB uses two predictors, the lowest CV RMSE (4.88 dB) is obtained using 14 trees, with a maximum depth of 3, a minimum number of samples equal to 10, and the absolute error loss function; and, using four predictors, the lowest CV RMSE (4.73 dB) is reached using 58 trees, with the maximum depth of 11, at least 10 samples per node and the absolute error loss function. The ML-based and log-distance model performances are compared in Table 6.9 together with the CNN-based PL models.

The log-distance model achieves an RMSE, MAPE, σ, and R^2 on the testing set equal to 9.74 dB, 12.80%, 5.75 dB, and -0.77, respectively. Table 6.9 shows that the ML models using tabular data are vulnerable to over-fitting, mainly for the ANN and SVR (two predictors), which present the highest difference between the RMSEs for the training and the testing sets. A possible reason for this behavior is the small size of the training data hampering learning the patterns from the dataset. When using four predictors, the over-fitting increases for the ANN model, leading to the worst generalization performance on the testing set. The performance prediction of the SVR on the testing improves significantly when using the subset of four predictors.

Table 6.9: PL performance evaluation and interpretability for the ML and log-distance models. The estimated coefficients L_o and n for the log-distance model for the V2V training set is 46.36 and 1.23, respectively.

	Training set				Testing set					
Model	RMSE [dB]	MAPE [%]	σ [dB]	R^2	RMSE [dB]	MAPE [%]	σ [dB]	R^2	IAS	$\overline{\text{MEC}}$
Log-distance	6.16	8.58	3.36	0.18	9.74	12.80	5.75	-0.77	-	-
ANN (two predictors)	4.84	5.85	3.23	0.49	16.66	18.81	11.54	-4.17	0.21	2.43
ANN (four predictors)	4.35	5.24	2.92	0.59	19.12	21.08	13.77	-5.80	0.37	2.59
SVR (two predictors)	5.43	6.73	3.51	0.36	15.52	18.02	10.32	-3.48	0.11	2.00
SVR (four predictors)	5.27	10.52	3.58	0.40	8.50	10.50	5.17	-0.35	0.19	1.95
RF (two predictors)	4.73	5.62	3.24	0.51	7.59	9.55	4.57	-0.07	0.27	2.49
RF (four predictors)	4.43	5.18	3.07	0.57	7.53	9.50	4.47	-0.06	0.25	2.62
GTB (two predictors)	4.78	5.53	3.37	0.50	8.55	10.80	5.03	-0.36	0.33	2.90
GTB (four predictors)	3.28	2.90	2.75	0.77	7.78	9.89	4.60	-0.13	0.46	3.78
CNN trained from scratch	4.97	6.11	3.11	0.36	8.81	10.93	5.43	-0.44	-	-
pre-trained CNN	**3.94**	**5.15**	**2.36**	**0.60**	**6.97**	**9.18**	**3.85**	**0.10**	-	-

The performance in the RF on the training and testing sets slightly improves using four predictors. The GTB model achieves a better performance using four predictors. Finally, the best performance among the tabular ML models is provided by the RF model with four predictors, which achieves an RMSE, MAPE, σ and R^2 of 7.53 dB, 9.50%, 4.47 dB, and -0.06, respectively.

In addition, the interpretability indicators in Table 6.9 can bring some further conclusions. In the ANN model, when adding more predictors in the subset, both IAS and $\overline{\text{MEC}}$ increase, with values of 0.37 and 2.59, respectively, indicating a higher interaction between predictors. In the case of the SVR using four predictors, the IAS increases to 0.19, with similar $\overline{\text{MEC}}$ values between the two subsets. The IAS and $\overline{\text{MEC}}$ values on the RF models remains similar for the two subset. The GTB report an increase in the IAS value when using four predictors and a more significant increase of the $\overline{\text{MEC}}$ value for the GTB (3.78).

Table 6.9 also brings the results for the pre-trained and trained from scratch CNN models. As seen in Figure 6.15, the best performance prediction during training occurs for the pre-trained model, performing the best generalization results on the testing set. Therefore, the best performance among the ML and empirical models is obtained with the pre-trained CNN with an RMSE, MAPE, σ and R^2 on the testing set equal to 6.97 dB, 9.18%, 3.85 dB, and 0.10, respectively. The comparison between the measured and predicted path loss values for the log-distance, RF and pre-trained CNN model

are shown in Figure 6.16. For this vehicular environment, more measurement samples could be collected to explore the improvement in the PL prediction in this scenario.

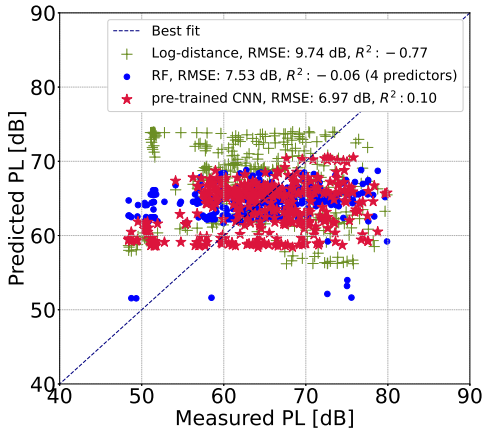

Figure 6.16: Comparison of the measured and predicted PL for the log-distance, RF and pre-trained CNN model on the testing set.

There are some interpretability methods to explain deep learning models that can be employed for the CNN models. Among them, one finds the gradient explanation technique based on the gradient attribution method [67]. This tool quantifies how much a change in each input would change the predictions in a small neighborhood around the input; and the guided back-propagation for visualizing the predictors learned by CNNs using a deconvolution technique. Deconvolution networks are employed to provide insight into the function of intermediate feature layers of an already trained CNN, mapping the feature activity in intermediate layers back to the input feature space.

6.2.6 Discussion

This chapter started with models for path loss between Tx-Rx pairs for a V2I channel using the profile along the direct path. The V2I results indicate that the 735 MHz channel was easier to model since its models attained the best performance indicators. This is most probably due to the closer statistical distributions between the training and testing sets in the different Routes considered. The receiver's velocity, traffic street density data, and the width of the street impact on the model accuracy still need to be further addressed for the V2I scenario.

In continuation, the chapter studied PL models for the V2V environment. A pre-trained model extracts the feature map from a satellite image. The

results show a performance improvement when using the transfer learning approach compared to training from scratch and better performance compared to log-distance and ML models using the tabular dataset. For both scenarios, we believe that using more extensive datasets may improve the performances of the V2I and V2V models. The greater the dataset, the larger the chance for the models to learn a broader range of patterns and variations of the propagation environment, leading to improved generalization capacity.

7
Conclusions and Future Work

The design of empirical path loss models aims at specific frequency bands and environments, given the particular propagation characteristics in each channel. This thesis has employed the same approach for machine learning-based path loss models. Starting with an extensive literature review, we identified key aspects from the most relevant works on ML techniques and methodologies for PL modeling. Those aspects include the machine learning technique, operating frequency, data generation for training, and train-test split methods for generalization testing. We considered them for designing, comparing, and analyzing the PL models in this thesis.

Many previous studies propose using ML techniques (ANN, SVR, and RF) with varying predictors, mainly for links operating at a single frequency in the UHF band. While some are based on measurements, others employ data simulated using the ray-tracing technique for training. However, in the last case, the representative radio propagation characteristics for model training depend on the accuracy of the solver generating the synthetic data. Furthermore, most previous works use the conventional random training-testing split, which may not be well-suited for predicting path loss for unknown positions of the transmitter or receiver involved.

Concerning the ML techniques, we have proposed several models with usually good performance for PL regression based on the tabular dataset produced by measurement campaigns carried out in Rio de Janeiro, Brazil, and thus representing real environments. We considered PL models for mmWave frequencies links in indoor and outdoor environments and multi-frequency macrocell coverage in the sub-6 GHz. For each environment, and considering the characteristics of the measurement campaigns, we have tailored different generalization tests to emphasize and assess the applicability of each trained model. Furthermore, we also dealt with the more dynamic propagation environment resulting from V2I and V2V wideband channels, considering measurement campaign data. For the V2V channel, a deep learning-based PL model has been proposed using a pre-trained model (ResNet18), i.e., based on transfer learning to counteract the effects of using a small dataset.

In Table 7.1, we summarize the results obtained for the different scenarios in terms of propagation problem complexity, data size for training, generalization improvement over the empirical path loss model (EPLM), and evaluation of additional measurement data.

Table 7.1: Summary results for the different environments.

Scenario	**Propagation problem complexity**	**Data size for training**	**Generalization improvement over the EPLM**	**Additional measurement data**
mmWave indoor	**High** ·Multi-frequency and multi-walls scenario. ·All the links presented NLOS condition.	**Medium** ·Higher than 700 PL instances.	**Major** ·Performance gain higher than 4 dB in the average RMSE (for the generalization test of unknown transmitter).	The proposed ML-based PL model presented a good performance in the tailored generalization tests.
mmWave outdoor	**Medium** ·Multi-frequency scenario ·Most of the links presented LOS condition.	**Small** ·Lower than 250 PL instances.	**Low** · Performance gain lower than 0.1 dB in the average RMSE (for the generalization test of unknown receivers).	Additional measurement data from different scenarios still need to be addressed to further evaluate the proposed ML-based PL model.
Macrocell coverage	**High** ·Multi-frequency scenario. ·Larger Tx-Rx distance. ·Dense buildings and vegetation.	**High** ·Higher than 12,000 PL instances.	**Major** ·Performance gain higher than 4 dB in the average RMSE (for the three frequencies, unknown streets).	To further investigate the generalization capacity of the ML-based PL model, an additional dataset can be employed.
V2I	**High** ·Multi-frequency scenario ·Low transmission antenna heigth. ·Most of the links presented NLOS condition.	**Medium** ·Higher than 700 PL instances.	**Major** · Performance gain higher than 6 dB in the average RMSE (for the three frequencies, unknown streets).	Using more extensive datasets may improve the performances of the V2I and V2V models.
V2V	**High** ·Higher dynamic channel. · Most of the links presented NLOS condition.	**Medium** ·Higher than 700 PL instances.	**Medium** ·Performance gain lower than 3 dB in RMSE for unknown streets.	

The hyperparameters Θ of each proposed ML model have been carefully chosen according to their performance, using cross-validation for reliable model selection. We have employed performance indicators such as the RMSE, MAPE, σ, and R^2 to compare the ML techniques and the proposed empirical models. We defined the propagation problem complexity of the addressed scenarios such as low, medium and high in terms of multi-frequency prediction and obstruction of the direct link. Regarding the data size used for training, we classified the obtained volume of data such as small, medium or high. We discuss the table results below, along with the conclusions drawn for each scenario addressed.

In Chapter 3, the PL modeling for the mmWave indoor links has addressed a multi-frequency scenario ranging from 26.5 GHz to 40 GHz. Considering the multi-frequency and multi-wall conditions, along with the presence of NLOS conditions in all the links, this propagation problem presented a high level of complexity. Different transmitters and receiver positions were considered to fully coverage in the building floor obtaining a volume data of medium size. Four ML models have been proposed: ANN, SVR, RF, and GTB; and their performance were compared against adjusted empirical models. When using the conventional random train-test split (80% for training and 20% for testing), the proposed empirical model ABGnw achieved the best generalization performance among the empirical models, with an RMSE of 9.11 dB, an MAPE of 6.52%, an σ of 5.70 dB and R^2 of 0.61. Although worse than the ABGnw, the CIFnw achieved a performance with an RMSE of 10.13 dB, an MAPE of 6.98%, an σ of 6.58 dB, and R^2 of 0.52.

An extensive comparison has been conducted using the proposed methodology for coalition selection for all the ML models. Furthermore, since the GTB-based PL model presents the best performance among the machines pre- and post-hoc interpretation tools were applied and analyzed over its response. The final optimized GTB model using the predictors f, d, and n_w attained a generalization performance with an RMSE of 4.29 dB, an MAPE of 3.03%, an σ of 2.75 dB and R^2 of 0.92; with a performance gain over the ABGnw of 4.82 dB in RMSE, 3.49% in MAPE, 2.95 dB in σ and 0.31 in R^2. In the generalization test considering an unknown transmitter point to best fits a real-world scenario, the GTB-based PL model achieved an average generalization performance with an RMSE of 5.15 dB from evaluation from the five models when using a set of transmitters for training and the test employed a different one. Therefore, as presented in Table 7.1, we considered the improvements in generalization as major. For additional generalization tests, more data might be desired.

Chapter 4 also dealt with mmWave links, in a multi-frequency case ranging from 27 GHz to 40 GHz, but outdoors. The measurement campaign data encompassed one transmitter and different receiver positions around the campus of PUC-Rio leading to a medium propagation complexity in terms of most LOS conditions and shorter transmission distances. However, the measurement campaign provided a small volume of data in comparison with the indoor case. The influence of the height difference between the transmitter and receiver and vegetation has been assessed for this environment. When adding those predictors to the ABG and CIF models leading to the ABG$\Delta_h, v_{\text{depth}}$ and CIF$\Delta_h, v_{\text{depth}}$ models, that show improved performance over the original ones. When using the conventional random train-test split (80% for training and 20% for testing), one observes a performance gain of 0.17 dB in RMSE, 0.24% in MAPE, and 0.03 in σ, and 0.05 in R^2 for the ABG$\Delta_h, v_{\text{depth}}$ and improvements of 0.37 dB in RMSE, 0.30% in MAPE, 0.17 dB in σ, and 0.11 in R^2 for the CIF$\Delta_h, v_{\text{depth}}$.

Among the ML models, the GTB with the predictors f, d, Δ_h and v_{depth} in the coalition achieves the best generalization performance with an RMSE of 1.39 dB, an MAPE of 0.94%, an σ of 0.90 dB, and R^2 of 0.90; with a performance gain over the ABG$\Delta_h, v_{\text{depth}}$ of 1.13 dB in RMSE, 0.69% in MAPE, 0.74 dB in σ, and 0.24 in R^2 on the testing set. In the generalization test considering unknown receiver points, the GTB-based PL model achieved an average generalization performance with an RMSE of 2.05 dB from the evaluation of the three models when using a set of receivers (seventeen) for training and testing with four unknown points. However, since the empirical models perform significantly well, the gain given by the ML models showed minor improvements. Additional data for training and testing could be desired to further investigation in this environment.

In Chapter 5, we have tackled the PL modeling for macrocell coverage in the sub-6 GHz (750 MHz, 2.5 GHz, and 3.5 GHz). As described in Table 7.1, the propagation complexity level is considered high in this environment. We have characterized the entire environment profile to extract the most significant predictors in this scenario. The conducted campaigns considered two fixed positions of the transmitters, and measurements were collected from two routes. In comparison to the mmWave scenarios, this measurement campaign provided a high volume of data. The GTB regression has shown the best generalization capabilities among the ML models. Using the predictors coalition containing b_{depth}, d, f, $\overline{g}_h$, Δ_h, and v_{depth}, the final optimized model using one route for training and the other for testing to assess generalization performance on a similar street, achieved an average RMSE on the testing set

of 9.42 dB (considering the performance prediction in the three frequencies), an average MAPE of 6.72%, an average σ of 5.66 dB, and an average of R^2 0.72. Between the empirical models, the best results were obtained for the Okumura-Hata, with an average RMSE on the testing set of 14.24 dB, an average MAPE of 11.45%, an average σ of 7.87 dB, and an average of R^2 0.38. In terms of gain performance, the ML model presented a major improvement in comparison with the empirical models.

Generally, among the machine learning models, those based on trees such as RF and GTB presented superior performances for PL model design using tabular datasets, and the GTB model has presented the best agreement between the predicted and the testing PL values. Also, the interpretability results showed that the GTB model relies most heavily on the main effects of each predictor in the coalition, irrespective of the frequency band and environment. Also, examining the overall interpretability complexity of the final optimized GTB models in the mmWave and sub-6 GHz bands, we have observed a lower complexity for the sub-6 GHz GTB-based PL model. The reduced complexity in the sub-6 GHz band can be attributed to more linear effects of the predictors on the response, given the operating frequency and distance of the links.

Chapter 6 has addressed the vehicular to infrastructure (V2I) in the sub-6 GHz frequency (735 MHz, 2.5 GHz, and 3.5 GHz). The data from the measurement campaigns employed had two routes using a low-height antenna transmitter for V2I communications representing a high level of propagation complexity, with a volume of data of size medium. Among the ML models, the GTB model presented the best generalization capacity for the V2I scenario. Using the coalition composed of f, n_b, σ_b, n_v, σ_v, σ_g, and d, the final optimized model (using one route for training and the other for testing to assess generalization performance on a similar street) achieved the best performance over the log-distance model with an average RMSE on the testing set of 6.14 dB (considering the performance prediction in the three frequencies), an average MAPE of 9.31%, and average σ of 3.38 dB. The log-distance achieved an average RMSE of 13.35 dB, an average MAPE of 21.61%, and average σ of 4.32 dB. The GTB showed the best generalization performance for the 735 MHz link, which presented the more similar PL distribution in both routes, for the links operating at 2.54 GHz and 3.5 GHz, it presents the lowest generalization. We noted a major improvement of the ML model in comparison with the empirical model.

Finally, continuing vehicular scenario channels, Chapter 6 also considered V2V links at 5.8 GHz. This scenario is characterized by a high level of

propagation complexity in terms of a higher dynamic channel and most of the links presented NLOS conditions. In this case, we proposed a deep learning model using a pre-trained CNN (i.e., using transfer learning), instead of training the CNN from scratch. The proposed approach leads to the best predictive performance among the ML models in the unknown street scenario. The pre-trained CNN model achieved a performance generalization with an RMSE of 6.97 dB, MAPE of 9.18%, σ of 3.85 dB and R^2 of 0.10. The performance gain over the log-distance is 2.77 dB in RMSE, 3.62% in MAPE, 1.9 dB in σ and 0.67 in R^2; and the performance gain over the RF model with the predictors d_x,d_y,v_x, and v_y, which attained the best performance among the tabular data-based models is 0.56 dB in RMSE, 0.32% in MAPE, 0.62 dB in σ and 0.16 in R^2. The results indicate that the proposed approach points to potential gains of using CNN for V2V PL prediction, considering the small dataset, with a medium performance improvement in comparison with the empirical model. For the vehicular channels, using more data may improve the performance of the ML models.

Summarizing, in this thesis we have proposed and evaluated ML models and associated methodologies for predictor coalition selection and database split for the different frequency bands and environments. Therefore, besides the obtained models, our contributions also encompass the following aspects.

The most relevant predictors for each radio propagation problem were selected using a predictor coalition methodology. This approach is adequate for radio propagation channel modeling since the dimensional data associated with the number of predictors and the output are usually small, which allows for examining each predictor's importance in improving the predictive performance.

We have interpreted the mappings provided by the selected coalitions (a significant subset of predictors) using global interpretability indicators such as IAS and $\overline{\text{MEC}}$ to quantify the global linear and non-linear effects of the predictors on the predicted path loss. Higher values of IAS and $\overline{\text{MEC}}$ in the ML models suggest a more complex model and harder to interpret due to the higher non-linear effects of predictors on path loss and higher interactions between predictors. The response of those indicators has been further examined using visual tools such as the ALE curves. Other tools presented in this thesis include evaluating the interaction between pairs of predictors and permuted feature importance values in the subset. The analysis methodology has shown that the predictor coalition methodology consistently has presented the ability to select the most relevant predictors to improve path loss prediction without compromising the model's predictive performance,

reducing the model complexity and leading to an improved interpretation of the model irrespective of the frequency and environment where it is applied.

Finally, we assessed the effects of using the conventional train-test random split and a tailored train-test split, based on the particular characteristics of the measurement campaigns, on model performance. As expected, models using train-test random split presented the best PL prediction since the testing set is contaminated by some samples collected very close to those used for training the model. However, the tailored train-test division best assesses the generalization capacity of the trained ML models for their practical application when they are trained using similar but not the same scenario where the model shall work. The proposed train-test split results have shown that the presented ML models present good generalization capacity and can be useful for path loss prediction for the different radio link categories addressed. The ML models have demonstrated the ability to capture the patterns and relationships between the predictors and path loss that challenge traditional empirical models.

7.1 Future Work

For the continuation of the work on ML-based PL prediction models, we can list problems related to each chapter of this thesis.

Upon examining the predictors obtained in each selected subset for each environment investigated, specifically for the mmWave Outdoor, macrocell, and V2I cases, similarities have been observed in the coalition members, such as frequency, distance, height difference between the transmitter and receiver, and predictors related to variability of the building height and vegetation height along the links. Therefore, one can ask if it is possible to build a single ML-based PL prediction model that allows the integration of heterogeneous urban cells considering different frequencies and scenarios. Looking to answer the above question, the selected subset obtained for the macrocell in the sub-6 GHz scenario and V2I could be interchanged to investigate the resulting prediction accuracy, leading to the use of a general subset of predictors that maximizes in all the environments the prediction of the model.

To further investigate the generalization capacity of the GTB-based PL model for macrocell coverage proposed in Chapter 5, the additional dataset available from a wideband measurement campaign, illustrated in Figure 7.1, can be employed. This measurement campaign was carried out in Rio de Janeiro, Brazil, in November 2017 for the frequency bands of 700 MHz, 2.5 GHz, and 3.5 GHz [158].

Figure 7.1: Wideband measurement campaign for macrocell coverage carried out in Rio de Janeiro Brazil, in November 2017 [157].

To adapt the proposed CNN model to urban macrocell links prediction, considering the longer Tx-Rx links involved in that scenario, image extraction as depicted in Figure 7.2 can be explored. In this technique, the parameter w can be varied to extract an image encompassing the propagation environment along the direct path from transmitter to receiver. However, the large distance between the transmitter and the receiver may lead to information loss when resizing the image; thus, countermeasures should be taken.

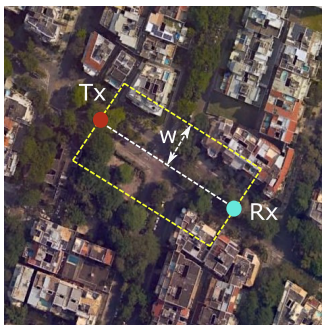

Figure 7.2: Proposed technique to extract images for macrocell coverage.

For the wideband measurement campaign data, obtained from the vehicular environments shown in Chapter 6, the prediction of the power profile delay could also be investigated as a strategy to derive other channel parameters such as the delay spread, as described in Subsection 2.1.3, which is critical for the design of wideband radio channels due to the effect of multipath propagation. Joint modeling of path loss and delay spread for wideband wireless systems can also be investigated using the proposed ML techniques. Although, other ML algorithms, such as recurrent neural networks (RNN) and Long Short Term Memory (LSTM), could also be of interest to this problem.

7.2 Published Works

The author has contributed the following published works:

- Y. Nunez, L. Lovisolo, L. da Silva Mello and C. Orihuela, "On the Interpretability of Machine Learning Regression for Path Loss Prediction of Millimeter-wave Links." Expert Systems with Applications 215 (2023): 119324.
- Y. Nunez, L. Lovisolo, L. da Silva Mello and C. Orihuela, "Path Loss Prediction of Millimeter-wave using Machine Learning Techniques." 2022 IEEE Latin-American Conference on Communications (LATINCOM). IEEE, 2022.
- Y. Nunez, L. da Silva Mello and C. Orihuela "Path Loss Prediction for 5G Millimeter Waves Propagation based in Artificial Neural Networks" 2020, Simpósio Brasileiro de Micro-ondas e Optoeletrônica (MOMAG), 2020.

Bibliography

[1] S. Mattisson, "Overview of 5g requirements and future wireless networks," in *ESSCIRC 2017-43rd IEEE European Solid State Circuits Conference*. IEEE, 2017, pp. 1–6.

[2] Y. Yifei and Z. Longming, "Application scenarios and enabling technologies of 5g," *China Communications*, vol. 11, no. 11, pp. 69–79, 2014.

[3] 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, 2021.

[4] A. Sufyan, K. B. Khan, O. A. Khashan, T. Mir, and U. Mir, "From 5g to beyond 5g: A comprehensive survey of wireless network evolution, challenges, and promising technologies," *Electronics*, vol. 12, no. 10, p. 2200, 2023.

[5] F.-L. Luo, *Machine learning for future wireless communications*, 1st ed. John Wiley & Sons, 2020.

[6] S. Sung, W. Choi, H. Kim, and J.-i. Jung, "Deep learning-based path loss prediction for fifth-generation new radio vehicle communications," *IEEE Access*, 2023.

[7] S. K. Ibrahim, M. J. Singh, S. S. Al-Bawri, H. H. Ibrahim, M. T. Islam, M. S. Islam, A. Alzamil, and W. M. Abdulkawi, "Design, challenges and developments for 5g massive mimo antenna systems at sub 6-ghz band: A review," *Nanomaterials*, vol. 13, no. 3, p. 520, 2023.

[8] C.-L. Cheng, S. Kim, and A. Zajić, "Comparison of path loss models for indoor 30 ghz, 140 ghz, and 300 ghz channels," in *2017 11th European Conference on Antennas and Propagation (EUCAP)*. IEEE, 2017, pp. 716–720.

[9] M. Q. Khan, A. Gaber, P. Schulz, and G. Fettweis, "Machine learning for millimeter wave and terahertz beam management: A survey and open challenges," *IEEE Access*, vol. 11, pp. 11 880–11 902, 2023.

[10] C. F. Mecklenbrauker, A. F. Molisch, J. Karedal, F. Tufvesson, A. Paier, L. Bernadó, T. Zemen, O. Klemp, and N. Czink, "Vehicular channel characterization and its implications for wireless system design and performance," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1189–1212, 2011.

[11] Y. Yang, Z. Gao, Y. Ma, B. Cao, and D. He, "Machine learning enabling analog beam selection for concurrent transmissions in millimeter-wave v2v communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 9185–9189, 2020.

[12] L. Rubio, V. M. Rodrigo-Peñarrocha, J. Reig, and H. Fernández, "Investigation of the path loss propagation for v2v communications in the opposite direction," in *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*. IEEE, 2016, pp. 1685–1686.

[13] P. Qi, Y. Zhang, Z. Yuan, L. Yu, P. Tang, and J. Zhang, "Channel modeling based on 3d scenario information for v2i communications," in *2021 15th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2021, pp. 1–5.

[14] S.-K. Noh, P.-j. Kim, and J.-H. Yoon, "Doppler effect on v2i path loss and v2v channel models," in *2016 international conference on information and communication technology convergence (ICTC)*. IEEE, 2016, pp. 898–902.

[15] J. Turner, A. Shahriman, A. Harun, S. Murad, M. Isa, R. Ismail, D. Ndzi, M. Hashim, Z. Razlan, W. Wan *et al.*, "Modelling on impact of building obstruction for v2i communication link in micro cellular environment," in *Journal of Physics: Conference Series*, vol. 1755, no. 1. IOP Publishing, 2021, p. 012031.

[16] F. Pinzel, J. Holfeld, A. Olunczek, P. Balzer, and O. Michler, "V2v-and v2x-communication data within a distributed computing platform for adaptive radio channel modelling," in *2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, 2019, pp. 1–6.

[17] M. Yangi, B. Ai, R. He, D. Yao, J. Li, B. Zhang, Q. Wang, D. Fei, and M. Ni, "Path loss characteristics for vehicle-to-infrastructure channel in urban and suburban scenarios at 5.9 ghz," in *2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS)*. IEEE, 2017, pp. 1–4.

[18] 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.

[19] A. J. Campuzano Candel, H. A. Fernández González, D. Balaguer Andrés, A. Vila Jiménez, B. Bernardo Clemente, V. M. Rodrigo Peñarrocha, J. Reig, A. Valero-Nogueira, and L. Rubio Arjona, "Vehicular-to-vehicular channel characterization and measurement results," in *Waves*, vol. 4. Instituto de Telecomunicaciones y Aplicaciones Multimedia (iTEAM), 2012, pp. 15–24.

[20] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.

[21] "2023 ericsson mobility report," https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2023, accessed: 21/07/2023.

[22] S. Sun, G. R. MacCartney, and T. S. Rappaport, "Millimeter-wave distance-dependent large-scale propagation measurements and path loss models for outdoor and indoor 5g systems," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2016, pp. 1–5.

[23] Y. Corre, T. Tenoux, J. Stéphan, F. Letourneux, and Y. Lostanlen, "Analysis of outdoor propagation and multi-cell coverage from ray-based simulations in sub-6ghz and mmwave bands," in *2016 10th European Conference on Antennas and Propagation (EuCAP)*. IEEE, 2016, pp. 1–5.

[24] T. S. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Pearson Education, 2015.

[25] 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.

[26] V. V. Ratnam, H. Chen, S. Pawar, B. Zhang, C. J. Zhang, Y.-J. Kim, S. Lee, M. Cho, and S.-R. Yoon, "Fadenet: Deep learning-based mm-wave large-scale channel fading prediction and its applications," *IEEE Access*, vol. 9, pp. 3278–3290, 2020.

[27] O. Onubogu, K. Ziri-Castro, D. Jayalath, S. Demmel, and H. Suzuki, "Doppler and pathloss characterization for vehicle-to-vehicle communications at 5.8 ghz," in *2014 Australasian Telecommunication Networks and Applications Conference (ATNAC)*. IEEE, 2014, pp. 58–64.

[28] W. Li, X. Hu, and T. Jiang, "Path loss models for ieee 802.15. 4 vehicle-to-infrastructure communications in rural areas," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3865–3875, 2018.

[29] 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.

[30] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*. prentice hall PTR New Jersey, 1996, vol. 2.

[31] C. F. Rodrigues, L. Lovisolo, and L. Muratori, "On the parameters of itu-r p. 1546 propagation model for dual-polarization links," *IEEE Transactions on Broadcasting*, vol. 66, no. 1, pp. 56–65, 2019.

[32] A. Gupta, J. Du, D. Chizhik, R. A. Valenzuela, and M. Sellathurai, "Machine learning-based urban canyon path loss prediction using 28 ghz manhattan measurements," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 6, pp. 4096–4111, 2022.

[33] N. R. Leonor, S. Faria, G. Ramos, P. V. G. Castellanos, C. Rodríguez, L. da Silva Mello, and R. F. Caldeirinha, "Site-specific radio propagation model for macrocell coverage at sub-6 ghz frequencies," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 10, pp. 9706–9715, 2022.

[34] S. Sun, T. S. Rappaport, M. Shafi, P. Tang, J. Zhang, and P. J. Smith, "Propagation models and performance evaluation for 5g millimeter-wave bands," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8422–8439, 2018.

[35] S. Sun, T. S. Rappaport, T. A. Thomas, A. Ghosh, H. C. Nguyen, I. Z. Kovács, I. Rodriguez, O. Koymen, and A. Partyka, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5g wireless communications," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 2843–2860, 2016.

[36] M. B. Majed, T. A. Rahman, O. A. Aziz, M. N. Hindia, and E. Hanafi, "Channel characterization and path loss modeling in indoor environment

at 4.5, 28, and 38 ghz for 5g cellular networks," *International Journal of Antennas and Propagation*, vol. 2018, 2018.

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

[38] Y. Okumura, "Field strength and its variability in vhf and uhf land-mobile radio service," *Rev. Electr. Commun. Lab.*, vol. 16, pp. 825–873, 1968.

[39] W. C. Lee, *Mobile communications engineering: theory and applications*. McGraw-Hill Education, 1998.

[40] A. Seretis and C. D. Sarris, "An overview of machine learning techniques for radiowave propagation modeling," *arXiv preprint arXiv:2101.11760*, 2021.

[41] T. M. Mitchell *et al.*, *Machine learning*. McGraw-hill New York, 1997.

[42] M. Sugiyama, *Introduction to statistical machine learning*. Morgan Kaufmann, 2015.

[43] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[44] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[45] 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.

[46] I. Popescu, D. Nikitopoulos, I. Nafornita, and P. Constantinou, "Ann prediction models for indoor environment," in *2006 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*. IEEE, 2006, pp. 366–371.

[47] 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.* Citeseer, 2014, pp. 1–5.

[48] G. Yang, Y. Zhang, Z. He, J. Wen, Z. Ji, and Y. Li, "Machine-learning-based prediction methods for path loss and delay spread in air-to-ground millimetre-wave channels," *IET Microwaves, Antennas & Propagation*, vol. 13, no. 8, pp. 1113–1121, 2019.

[49] 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)*. IEEE, 2019, pp. 1–6.

[50] J. Wen, Y. Zhang, G. Yang, Z. He, and W. Zhang, "Path loss prediction based on machine learning methods for aircraft cabin environments," *IEEE Access*, vol. 7, pp. 159 251–159 261, 2019.

[51] D. Karra, S. K. Goudos, G. V. Tsoulos, and G. Athanasiadou, "Prediction of received signal power in mobile communications using different machine learning algorithms: A comparative study," in *2019 Panhellenic Conference on Electronics & Telecommunications (PACET)*. IEEE, 2019, pp. 1–4.

[52] X. Zhang, X. Shu, B. Zhang, J. Ren, L. Zhou, and X. Chen, "Cellular network radio propagation modeling with deep convolutional neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on knowledge discovery & data mining*, 2020, pp. 2378–2386.

[53] M. Z. Alam, H. F. Ates, T. Baykas, and B. K. Gunturk, "Analysis of deep learning based path loss prediction from satellite images," in *2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, 2021, pp. 1–4.

[54] P. Wang and H. Lee, "Meta-learning approaches for indoor path loss modeling of 5g communications in smart factories," *ICT Express*, 2022.

[55] S. P. Sotiroudis, S. K. Goudos, and K. Siakavara, "Deep learning for radio propagation: Using image-driven regression to estimate path loss in urban areas," *ICT Express*, vol. 6, no. 3, pp. 160–165, 2020.

[56] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[57] K. Hopf and S. Reifenrath, "Filter methods for feature selection in supervised machine learning applications–review and benchmark," *arXiv preprint arXiv:2111.12140*, 2021.

[58] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS)*. IEEE, 2016, pp. 1–5.

[59] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[60] M. Verleysen, F. Rossi, and D. François, "Advances in feature selection with mutual information," in *Similarity-Based Clustering*. Springer, 2009, pp. 52–69.

[61] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "Efficient greedy feature selection for unsupervised learning," *Knowledge and information systems*, vol. 35, no. 2, pp. 285–310, 2013.

[62] G. Casalicchio, C. Molnar, and B. Bischl, "Visualizing the feature importance for black box models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2018, pp. 655–670.

[63] 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.

[64] R.-T. Juang, "Path loss modelling based on path profile in urban propagation environments," *IET Communications*, vol. 16, no. 6, pp. 685–694, 2022.

[65] A. N. Parveen, H. H. Inbarani, and E. S. Kumar, "Performance analysis of unsupervised feature selection methods," in *2012 International Conference on Computing, Communication and Applications*. IEEE, 2012, pp. 1–7.

[66] U. Challita, H. Ryden, and H. Tullberg, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 12–18, 2020.

[67] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[68] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[69] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning–a brief history, state-of-the-art and challenges," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020,a, pp. 417–431.

[70] P. M. Ramya, M. Boban, C. Zhou, and S. Stańczak, "Using learning methods for v2v path loss prediction," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–6.

[71] C. Diakhate, "Propagation channel modeling at centimeter–and–millimeter–wave frequencies in 5g urban micro–cellular context," Ph.D. dissertation, Université Paris-Saclay (ComUE), 2019.

[72] S. R. Theodore *et al.*, "Wireless communications: principles and practice," 2002.

[73] H. Arslan and T. Yucek, "Delay spread estimation for wireless communication systems," in *Proceedings of the Eighth IEEE Symposium on Computers and Communications. ISCC 2003.* IEEE, 2003, pp. 282–287.

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

[75] K. Halford and M. Webster, "Multipath measurement in wireless lan," *Intersil Application Note AN9895*, vol. 1, 2001.

[76] Y. Yu, Y. Liu, W.-J. Lu, and H.-B. Zhu, "Measurement and empirical modelling of root mean square delay spread in indoor femtocells scenarios," *Iet Communications*, vol. 11, no. 13, pp. 2125–2131, 2017.

[77] A. Bharti, R. Adeogun, X. Cai, W. Fan, F.-X. Briol, L. Clavier, and T. Pedersen, "Joint modeling of received power, mean delay, and delay spread for wideband radio channels," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 8, pp. 4871–4882, 2021.

[78] S. I. Popoola, A. Jefia, A. A. Atayero, O. Kingsley, N. Faruk, O. F. Oseni, and R. O. Abolade, "Determination of neural network parameters for path loss prediction in very high frequency wireless channel," *IEEE access*, vol. 7, pp. 150 462–150 483, 2019.

[79] R. He, Y. Gong, W. Bai, Y. Li, and X. Wang, "Random forests based path loss prediction in mobile communication systems," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*. IEEE, 2020, pp. 1246–1250.

[80] H. Singh, S. Gupta, C. Dhawan, and A. Mishra, "Path loss prediction in smart campus environment: Machine learning-based approaches," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[81] 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.

[82] A. B. Zineb and M. Ayadi, "A multi-wall and multi-frequency indoor path loss prediction model using artificial neural networks," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 987–996, 2016.

[83] M. Ayadi and A. B. Zineb, "Body shadowing and furniture effects for accuracy improvement of indoor wave propagation models," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 5999–6006, 2014.

[84] L. Azpilicueta, M. Rawat, K. Rawat, F. M. Ghannouchi, and F. Falcone, "A ray launching-neural network approach for radio wave propagation analysis in complex indoor environments," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 5, pp. 2777–2786, 2014.

[85] 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.

[86] M. Piacentini and F. Rinaldi, "Path loss prediction in urban environment using learning machines and dimensionality reduction techniques," *Computational Management Science*, vol. 8, no. 4, pp. 371–385, 2011.

[87] Y. Egi and C. E. Otero, "Machine-learning and 3d point-cloud based signal power path loss model for the deployment of wireless communication systems," *IEEE Access*, vol. 7, pp. 42 507–42 517, 2019.

[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] S. Duangsuwan and M. M. Maw, "Comparison of path loss prediction models for uav and iot air-to-ground communication system in rural precision farming environment." *J. Commun.*, vol. 16, no. 2, pp. 60–66, 2021.

[90] 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.

[91] S. Ojo, A. Sari, and T. P. Ojo, "Path loss modeling: A machine learning based approach using support vector regression and radial basis function models," *Open Journal of Applied Sciences*, vol. 12, no. 6, pp. 990–1010, 2022.

[92] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[93] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[94] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[95] S. M. Aldossari and K.-C. Chen, "Machine learning for wireless communication channel modeling: An overview," *Wireless Personal Communications*, vol. 106, no. 1, pp. 41–70, 2019.

[96] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, 1996.

[97] P. J. Braspenning, F. Thuijsman, and A. J. M. M. Weijters, *Artificial neural networks: an introduction to ANN theory and practice*. Springer Science & Business Media, 1995, vol. 931.

[98] Z.-F. Fu and J. He, *Modal analysis*. Elsevier, 2001.

[99] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[100] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[101] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in neural information processing systems*, 1992, pp. 950–957.

[102] R. M. Golden, *Statistical Machine Learning: A Unified Framework*. Chapman and Hall/CRC, 2020.

[103] M. Awad and R. Khanna, *Efficient learning machines: theories, concepts, and applications for engineers and system designers*. Springer nature, 2015.

[104] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[105] M. H. Law and J. T.-Y. Kwok, "Bayesian support vector regression," in *International Workshop on Artificial Intelligence and Statistics*. PMLR, 2001, pp. 162–167.

[106] A. S. Abba, K. I. Musa, A. Umar, N. Saleh, H. M. Khamis, and M. K. Dauda, "Transfer learning strategy for satellite image classification using deep convolutional neural network," *International Journal of Engineering Technologies and Mangament*, vol. 5, no. 4, pp. 1–14, 2020.

[107] H. Ishwaran, "The effect of splitting on random forests," *Machine learning*, vol. 99, no. 1, pp. 75–118, 2015.

[108] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Cart," *Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA*, 1984.

[109] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[110] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[111] 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)*. IEEE, 2019, pp. 1–5.

[112] 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.

[113] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of medical systems*, vol. 44, pp. 1–16, 2020.

[114] 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.

[115] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6530–6541, 2019.

[116] J. J. Luke, R. Joseph, and M. Balaji, "Impact of image size on accuracy and generalization of convolutional neural networks," *Int. J. Res. Anal. Rev.(IJRAR)*, vol. 6, no. 1, pp. 70–80, 2019.

[117] S. Jiang and V. M. Zavala, "Convolutional neural nets: Foundations, computations, and new applications," *arXiv preprint arXiv:2101.04869*, 2021.

[118] A. Bazaga, M. Roldan, C. Badosa, C. Jimenez-Mallebrera, and J. M. Porta, "A convolutional neural network for the automatic diagnosis of collagen vi-related muscular dystrophies," *Applied Soft Computing*, vol. 85, p. 105772, 2019.

[119] Q. Zhu, J. Chen, L. Zhu, X. Duan, and Y. Liu, "Wind speed prediction with spatio–temporal correlation: A deep learning approach," *Energies*, vol. 11, no. 4, p. 705, 2018.

[120] U. S. Sani, O. A. Malik, and D. T. C. Lai, "Improving path loss prediction using environmental feature extraction from satellite images: Hand-crafted vs. convolutional neural network," *Applied Sciences*, vol. 12, no. 15, p. 7685, 2022.

[121] M. Shaha and M. Pawar, "Transfer learning for image classification," in *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 2018, pp. 656–660.

[122] R. Ribani and M. Marengoni, "A survey of transfer learning for convolutional neural networks," in *2019 32nd SIBGRAPI conference on graphics, patterns and images tutorials (SIBGRAPI-T)*. IEEE, 2019, pp. 47–57.

[123] F. Souza, R. Araújo, S. Soares, and J. Mendes, "Variable selection based on mutual information for soft sensors applications," in *9 th Portuguese Conference on Automatic Control, Coimbra*. Citeseer, 2010.

[124] H. H. Yang and J. E. Moody, "Data visualization and feature selection: new algorithms for nongaussian data." in *NIPS*, vol. 12. Citeseer, 1999.

[125] R. Smith, "A mutual information approach to calculating nonlinearity," *Stat*, vol. 4, no. 1, pp. 291–303, 2015.

[126] H. Cheng, Z. Qin, W. Qian, and W. Liu, "Conditional mutual information based feature selection," in *2008 International Symposium on Knowledge Acquisition and Modeling*. IEEE, 2008, pp. 103–107.

[127] M. Beraha, A. M. Metelli, M. Papini, A. Tirinzoni, and M. Restelli, "Feature selection via mutual information: New theoretical insights," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.

[128] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[129] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[130] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020,b.

[131] C. Molnar, G. Casalicchio, and B. Bischl, "Quantifying model complexity via functional decomposition for better post-hoc interpretability," *arXiv preprint arXiv:1904.03867*, 2019.

[132] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 82, no. 4, pp. 1059–1086, 2020.

[133] M. Flora, C. Potvin, A. McGovern, and S. Handler, "Comparing explanation methods for traditional machine learning models part 2: Quantifying model explainability faithfulness and improvements with dimensionality reduction," *arXiv preprint arXiv:2211.10378*, 2022.

[134] C. E. O. Vargas, "Millimeter waves propagation for a range of frequencies from 26.5 ghz and 40 ghz," 2018.

[135] L. Mello, M. Pontes, I. Fagundes, M. Almeida, and F. Andrade, "Modified rain attenuation prediction method considering the effect of wind direction," *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, vol. 13, pp. 254–267, 2014.

[136] F. Andrade, P. Cruz, and L. da Silva Mello, "Evaluation of itu-r rain attenuation prediction methods for terrestrial links," in *2015 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*. IEEE, 2015, pp. 1–4.

[137] J. M. Keenan and A. J. Motley, "Radio coverage in buildings," *British telecom technology Journal*, vol. 8, no. 1, pp. 19–24, 1990.

[138] J. Lloret, J. J. López, C. Turró, and S. Flores, "A fast design model for indoor radio coverage in the 2.4 ghz wireless lan," in *1st International Symposium onWireless Communication Systems, 2004.* IEEE, 2004, pp. 408–412.

[139] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[140] P. H. Westfall and A. L. Arias, *Understanding regression analysis: a conditional distribution approach*. CRC Press, 2020.

[141] H. M. Sani, C. Lei, and D. Neagu, "Computational complexity analysis of decision tree algorithms," in *Artificial Intelligence XXXV: 38th SGAI International Conference on Artificial Intelligence, AI 2018, Cambridge, UK, December 11–13, 2018, Proceedings 38*. Springer, 2018, pp. 191–197.

[142] K. H. Knuth, "Optimal data-based binning for histograms," *arXiv preprint physics/0605197*, 2006.

[143] D. Gencaga, N. K. Malakar, and D. J. Lary, "Survey on the estimation of mutual information methods as a measure of dependency versus correlation analysis," in *AIP Conference Proceedings*, no. 1. American Institute of Physics, 2014, pp. 80–87.

[144] A. Kraskov, H. Stögbauer, and P. Grassberger, "Erratum: estimating mutual information [phys. rev. e 69, 066138 (2004)]," *Physical Review E*, vol. 83, no. 1, p. 019903, 2011.

[145] Y.-G. Lim, Y. J. Cho, M. Sim, Y. Kim, C.-B. Chae, and R. A. Valenzuela, "Map-based millimeter-wave channel models: An overview, hybrid modeling, data, and learning," *arXiv preprint arXiv:1711.09052*, 2017.

[146] Y.-G. Lim, Y. J. Cho, M. S. Sim, Y. Kim, C.-B. Chae, and R. A. Valenzuela, "Map-based millimeter-wave channel models: An overview, data for b5g evaluation and machine learning," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 54–62, 2020.

[147] N. Palizban, S. Szyszkowicz, and H. Yanikomeroglu, "Automation of millimeter wave network planning for outdoor coverage in dense urban areas using wall-mounted base stations," *IEEE Wireless Communications Letters*, vol. 6, no. 2, pp. 206–209, 2017.

[148] A. A. Budalal and M. R. Islam, "Path loss models for outdoor environment—with a focus on rain attenuation impact on short-range millimeter-wave links," *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, p. 100106, 2023.

[149] A. M. Al-Samman, T. A. Rahman, M. N. Hindia, A. Daho, and E. Hanafi, "Path loss model for outdoor parking environments at 28 ghz and 38 ghz for 5g wireless networks," *Symmetry*, vol. 10, no. 12, p. 672, 2018.

[150] G. L. Ramos, C. Vargas, L. da Silva Mello, P. Pereira, R. Vieira, S. Gonçalves, and C. Rego, "Measurement and prediction of short-range path loss between 27 and 40 ghz in university campus scenarios," *Journal of Communication and Information Systems*, vol. 36, no. 1, pp. 184–191, 2021.

[151] C. E. O. Vargas, M. M. Silva, J. J. A. Arnez, and L. da Silva Mello, "Initial results of millimeter wave outdoor propagation measurements in a campus environment," in *2018 IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC)*. IEEE, 2018, pp. 901–904.

[152] Q. Guo and Y. Long, "Pade second-order parabolic equation modeling for propagation over irregular terrain," *IEEE antennas and wireless propagation letters*, vol. 16, pp. 2852–2855, 2017.

[153] Y. L. de Jong, M. H. Koelen, and M. H. Herben, "A building-transmission model for improved propagation prediction in urban microcells," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 2, pp. 490–502, 2004.

[154] T. Jawhly and R. C. Tiwari, "The special case of egli and hata model optimization using least-square approximation method," *SN Applied Sciences*, vol. 2, pp. 1–10, 2020.

[155] S. Hawinkel, W. Waegeman, and S. Maere, "Out-of-sample r 2: estimation and inference," *The American Statistician*, no. just-accepted, pp. 1–16, 2023.

[156] M. M. Silva, L. da Silva Mello, R. C. V. Rodríguez, M. Almeida, and P. G. Castellanos, "Urban mobile channel delay spread measurements at 700 mhz, 2.5 ghz and 3.5 ghz," in *12th European Conference on Antennas and Propagation (EuCAP 2018)*. IET, 2018, pp. 1–4.

[157] M. Yang, B. Ai, R. He, C. Huang, Z. Ma, Z. Zhong, J. Wang, L. Pei, Y. Li, and J. Li, "Machine-learning-based fast angle-of-arrival recognition for vehicular communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1592–1605, 2021.

[158] M. M. Silva, "Caracterização do canal de propagação ponto-Área e veículo a veículo (v2v) na faixa de 5,8 ghz," 2018.

[159] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[160] "Georeferencing sheets and scanned maps," https://www.qgistutorials.com/en/docs/3/georeferencing_basics.html, accessed: 27/07/2023.

[161] E. Westra, *Python geospatial development*. Packt Publishing Ltd, 2013.

A
Hyperparameter Grid Search of the ML Models Design for Indoor mmWave

Figure A.1 shows a comparison between the average CV RMSE for the training and validation set regarding some hyperparameters values from the grid search described in Chapter 3, Subsection 3.6.1 for the model selection in the mmWave indoor environment. For the ANN model, Figure A.1.(a), a comparison is performed among different activation functions, including Logistic, Tanh and ReLU varying the number of neurons in the hidden layer. The training and validation curves show that the lowest CV RMSE in the validation set is for the activation function ReLU. For the SVR model, Figure A.1.(b) shows the CV RMSE curves varying the hyperparameters ϵ, number of support vectors, and C. For the RF model, Figure A.1.(c), the hyperparameters varied are the number of trees and maximum depth (md). The curves show that higher maximum depth lead to lower CV RMSE values. Lastly, the GTB model is tested with different learning rates (lr); after 246 trees, the CV RMSE values in the validation set are very similar for the evaluated learning rate values.

Figure A.2 presents the average CV RMSE at the grid points for the four ML models in the indoor mmWave from Chapter 3, Subsection 3.6.1. Each graph also highlights the grid point presenting the best performance as described in Subsection 3.6.1.

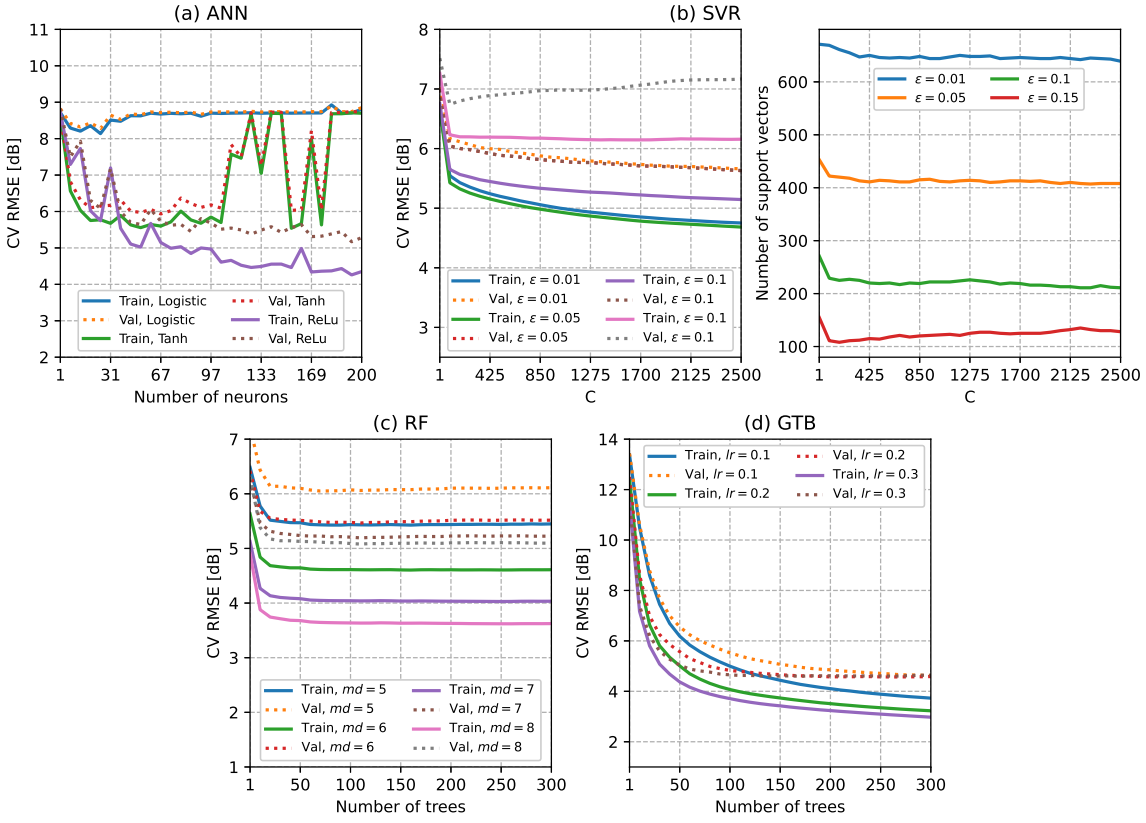

Figure A.1: Curve of training and validation set in the CV technique for the models: (a) ANN, (b) SVR, (c) RF and (d) GTB.

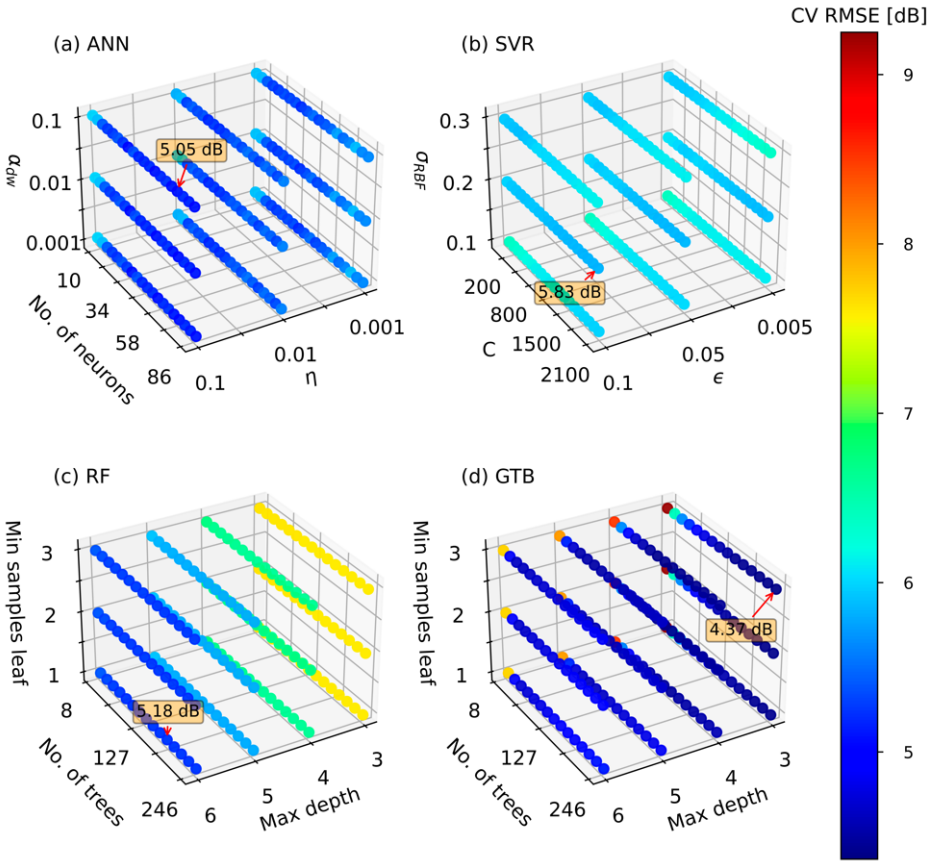

Figure A.2: Grid search of the hyperparameter optimization for the four models: (a) ANN, (b) SVR, (c) RF, and (d) GTB. Each graph presents the RMSE (color) according to the three different hyperparameters. The highlighted box presents the lowest RMSE value in each graph, and the arrow points to the correspondent hyperparameter point.

B
Interpretability Results for the ML Models in the mmWave Indoor Environment

The following sections discuss the interpretability results for the models ANN, SVR and RF in the different coalitions for the mmWave indoor environment.

B.1
Interpretability Results for the ANN-based Path Loss Model

We first look at the pre-training indices reported in Chapter 3, Table 3.7. The highest JMI was observed for the subset d_x, d_y (0.31), followed by d_x, d_y, f (0.24). However, adding the predictor n_w, the JMI is slightly reduced to 0.23. The CMI values among coalitions remain similar. The best one-predictor ANN-based path loss model is provided by d_x (RMSE: 9.18 dB, MAPE: 6.51%, σ: 5.72 dB and R^2: 0.59). Including d_y as a predictor significantly improved the model's performance (RMSE: 6.90 dB, MAPE: 4.75%, σ: 4.52 dB, R^2: 0.77). Adding f to the model input further improved performance (RMSE: 5.47 dB, MAPE: 3.40%, σ: 3.98 dB, R^2: 0.85). Lastly, the lowest performance improvements are for the predictors n_w and d, respectively.

For the ANN model, the predictors of distance d,d_x, and d_y show the most significant importance to improve path loss as presented in the PFI values (see Table 3.8) for the coalitions when they are added. However, the PL prediction mainly relies on the stronger interactions between them, specifically for d_x,d_y, and d_x,d as seen in Table B.1, and summarized in the IAS values for the different subsets in Table 3.7. Other strong interactions between pairs of predictors appear between distance and the number of walls, specifically in d_x,n_w, and d,n_w. Higher IAS and lower $\overline{\text{MEC}}$ values have been observed for the ANN model compared to the machines based on decision trees. The lower $\overline{\text{MEC}}$ values point to less non-linear effects of the predictors on the output, more markedly in distance and frequency compared with the RF and GTB models.

In addition, looking at the correlation that d and d_x share in Figure 3.3, the results suggest that d and d_x provide the same information to the model, where the model still perform well without d_x. Thus, in the permutation process

(as seen in Table 3.8), when calculating PFI, the model still has access to the information through the remaining d or d_x. As a result, the rank of predictor importance for the subset of four and five predictors remains the same, and their PFI values are similar.

Table B.1: Interaction strength between two predictors (2D-ALE) for the ANN-based PL model, for each number of predictors used for machine design, the first row presents the pair having the lowest interaction, and the pairs follow in increasing interaction order.

Interaction between two predictors (2D-ALE)							
Subset of predictors							
2		3		4		5	
d_x,d_y	1.540	d_y,f	0.025	f,n_w	0.020	f,n_w	0.020
		d_x,f	0.190	d_y,f	0.060	d_y,f	0.027
		d_x,d_y	10.38	d_y,n_w	0.080	d_x,f	0.111
				d_x,f	0.090	d_y,f	0.162
				d_x,n_w	0.700	f,d	0.184
				d_x,d_y	3.930	d_y,d	0.244
						d_x,n_w	0.440
						n_w,d	0.522
						d_x,d_y	0.735
						d_x,d	0.757

B.2 Interpretability Results for the SVR-based Path Loss Model

It was observed similar JMI values in the coalitions compared to the ANN model. The hightest JMI is for the subset of three predictors n_w,d_y and d_x (0.34). In adittion, when adding d_x to the coalition, the CMI value increase to 0.20. In the SVR-based path loss model, the best-one predictor showed that n_w provided the best performance (RMSE: 9.70 dB, MAPE: 7.01%, σ: 6.08 dB, R^2: 0.54), similar to the ANN model. Adding d_y resulted in a lower performance improvement compared to the ANN model. Including d_x improved the model's performance (RMSE: 6.34 dB, MAPE: 4.51%, σ: 33.91 dB, R^2: 0.80). Adding the last predictor d the improvement of the model's is meager (RMSE: 4.90 dB, MAPE: 3.21%, σ: 3.32 dB, R^2: 0.88).

In comparison with the ANN, RF and GTB, this machine presents the hightest IAS values from the subset of two until five predictors, which suggest that the PL prediction is highly dependent of the interaction between predictors. The $\overline{\text{MEC}}$ values remains similar compared to the ANN model. Examining the interaction between pair of predictors in Table B.2, the strongest interaction are presented between d_y and n_w in all the coalitions. As seen in the ALE

curves in Figure 3.4.(d), the individual effect of f on the path loss is similar compared to the ANN model and the predictor d in the SVR has a opposite effects in comparison with the ANN regressor. Among all the machines and coalitions, the predictor d and d_x show the highest accumulated local effects on the path loss.

Table B.2: Interaction strength between two predictors (2D-ALE) for the SVR-based PL model.

Interaction between two predictors (2D-ALE)							
Subset of predictors							
2		3		4		5	
n_w,d_y	0.478	n_w,d_x	0.419	n_w,f	0.041	n_w,f	0.101
		d_y,d_x	0.698	d_y,f	0.114	d_y,f	0.138
		n_w,d_y	0.946	d_x,f	0.149	n_w,d_x	0.189
				n_w,d_x	0.337	d_y,d_x	0.258
				d_y,d_x	0.809	n_w,d	0.289
				n_w,d_y	0.883	d_x,f	0.298
						n_w,d_y	0.324
						d,f	0.571
						d,d_y	0.621
						n_w,d_y	1.060

For the SVR regressor, we have observed that the model suffers a greater impact on the magnitude of PFI, as seen in Table 3.8. The phenomena occur because the PFI indicator relies on estimates of model errors and changes in prediction; thus, its magnitude value can be affected by the hyperparameter of regularization C. An example of that effect is shown in Figure B.1; when the C value is larger, the error penalty is more affected when the model's predictor is permuted or changed.

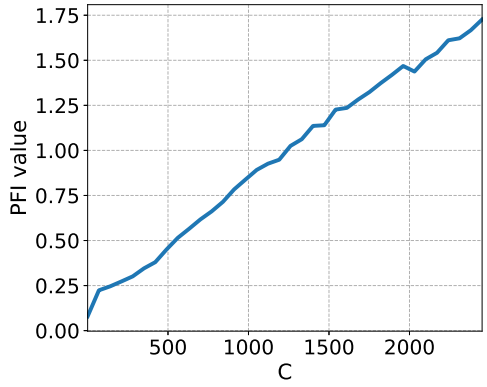

Figure B.1: Comparison between the value of the PFI and its response to vary C.

B.3
Interpretability Results for the RF-based Path loss Model

In the RF model, the order of the coalitions differed from ANN and SVR. The JMI values are similar compared to the them, and the CMI values remains similar compared to the GTB. For the RF-based path loss model, the best one-predictor is attained by d (RMSE: 7.10 dB, MAPE: 4.88%, σ: 4.65 dB, R^2: 0.75). Adding n_w improved the model's performance (RMSE: 6.22 dB, MAPE: 4.45%, σ: 3.78 dB, R^2: 0.81), which shows better performance in comparison when using two predictors for the ANN and SVR models. Lastly, the lowest performance improvements correspond to the predictors d_y and d_x, respectively.

Examining the global interaction between predictors summarized by the IAS values in the coalitions, higher values are obtained compared to the GTB, even considering the same coalitions in the subsets as seen in Table 3.7. Therefore, the performance prediction relies more on the interaction between predictors than their main effect. The strongest interaction occurs between n_w with d_y,d and f. Furthermore, the most significant predictor, as seen in Table 3.8 is for n_w, followed by d_y and f, as also seen for the GTB model. In addition, the $\overline{\text{MEC}}$ values are lower compared to the GTB model, but with similar ALE curves for the predictors d, n_w, and d_y, as seen in Figure 3.4.(a)(b)(d), respectively.

Table B.3: Interaction strength between two predictors (2D-ALE) for the RF-based PL model.

Interaction between two predictors (2D-ALE)							
Subset of predictors							
2		3		4		5	
d,n_w	0.184	d,n_w	0.013	d,d_y	0.004	d,f	0.013
		d,f	0.115	d,f	0.013	d,d_x	0.004
		n_w,f	0.161	d,n_w	0.054	n_w,f	0.014
				f,d_y	0.072	f,d_x	0.021
				n_w,f	0.129	d_x, d_y	0.021
				n_w,d_y	0.390	f,d_y	0.044
						n_w,d_x	0.053
						d,f	0.054
						d,n_w	0.055
						n_w,d_y	0.339

C
Evaluation of the Pre-trained CNN model

To assess the possibility of using the pre-trained ResNet18 model, we conducted experiments on land use classification tasks. We employed two distinct datasets, EuroSAT and NWPU-RESISC45. Figure C.1 shows some representative examples belonging to these datasets. The parameters of the first layers of the ResNet18 (which are said to extract the predictors or features) are not trained; only the (last) fully connected layer (classifier) is trained to classify images among 45 classes.

The classifier incorporates three hidden layers with 4096 neurons, as shown in Figure C.2. During training the learning rate is set to 0.0001, the batch size to 64, the activation function is the ReLU, the optimizer is the Adam loss and early-stopping. The fully connected layer for land classification was trained using a random split to 80%/20% for training/testing. The experiments were carried out in Python with the Pytorch framework on a workstation with an Intel Core i7 8th Gen processor, 64 GB of RAM, and the GPU GeForce RTX 2060. The classification performance is shown in Figure C.3.

The results show that using the EuroSAT dataset, the model achieves an accuracy of 91.12% and 90.89% on the training and testing sets, respectively. Moreover, for the NWPU-RESISC45, the model achieves an accuracy of 81.38% and 80.16% on the training and testing sets, respectively. As seen in the above results, the predictor extractor of the pre-trained ResNet18 can extract patterns from the satellite images.

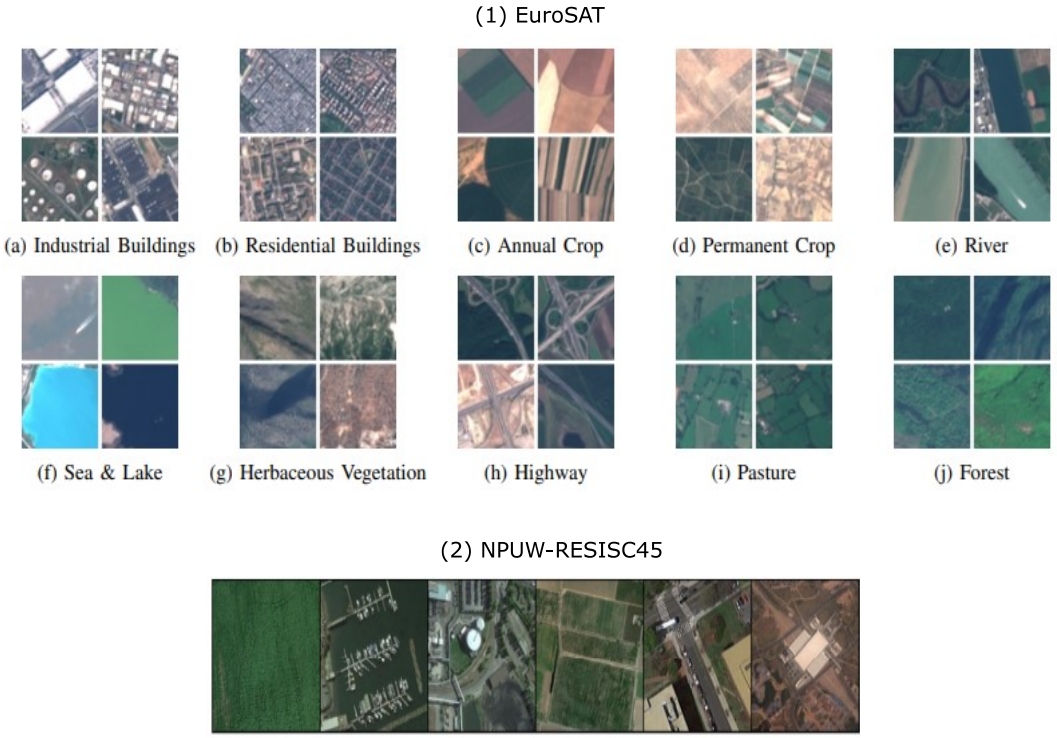

Figure C.1: A snapshot of some images from the dataset (1) EuroSAT [158] and (2) NWPU-RESISC45 [115] .

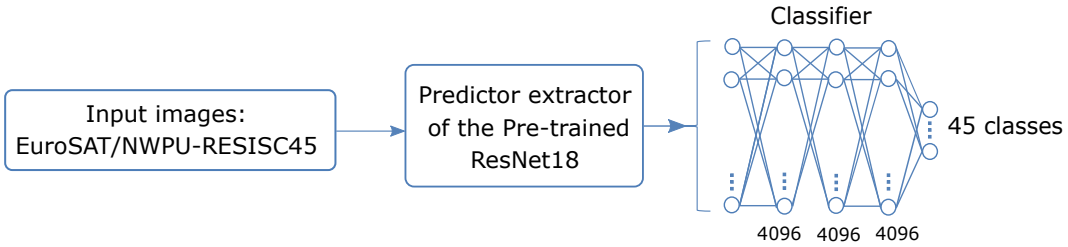

Figure C.2: Pre-trained CNN model for land use classification.

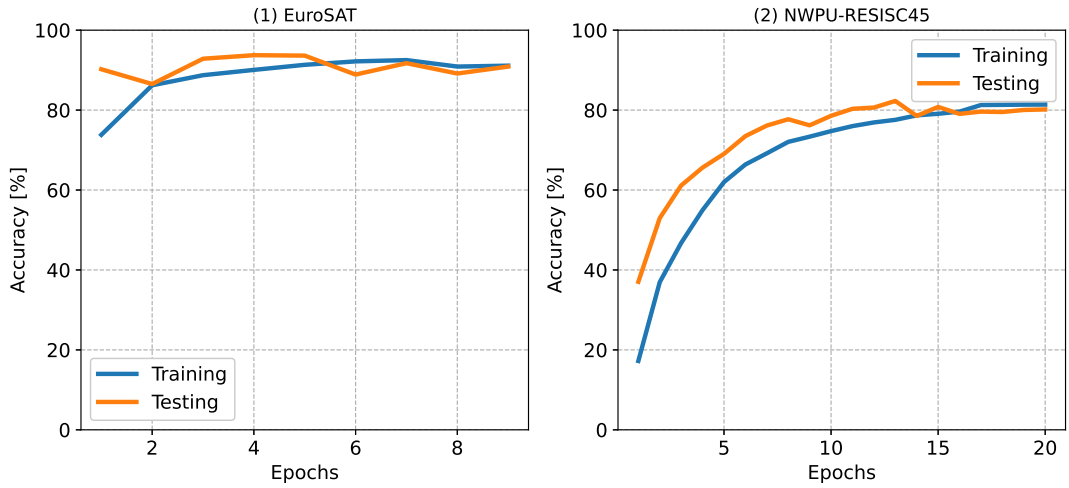

Figure C.3: Accuracy of the land use classifier using the pre-trained ResNet18 using EuroSAT (left) and NWPU-RESISC45 (right).