Machine Learning-Based Line-Of-Sight Prediction in Urban Manhattan-Like Environments

Nicola Di Cicco*, Simone Del Prete†, Silvi Kodra†, Marina Barbiroli† Franco Fuschini†, Enrico M. Vitucci†, Vittorio Degli Esposti†, Massimo Tornatore*,
*Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Italy
†Department of Electrical, Electronic and Information Engineering (DEI), CNIT, University of Bologna, Italy

***Abstract*—This paper considers the problem of predicting whether or not a transmitter and a receiver are in Line-of-Sight (LOS) condition. While this problem can be easily solved using a digital urban database and applying ray tracing, we consider the scenario in which only few high-level features descriptive of the propagation environment and of the radio link are available. LOS prediction is modelled as a binary classification Machine Learning problem, and a baseline classifier based on Gradient Boosting Decision Trees (GBDT) is proposed. A synthetic ray-tracing dataset of Manhattan-like topologies is generated for training and testing a GBDT classifier, and its generalization capabilities to both locations and environments unseen at training time are assessed. Results show that the GBDT model achieves good classification performance and provides accurate LOS probability modelling. By estimating feature importance, it can be concluded that the model learned simple decision rules that align with common sense.**

***Index Terms*—propagation modelling, ray tracing, line-of-sight probability, machine learning, datasets.**

I. Introduction

The presence of Line-of-Sight (LOS) condition between two radio link ends represents one of the basic properties of a propagation environment. LOS determines the fundamental characteristics of the radio channel, e.g. link budget, fading statistics, time and angle spreading, and heavily impacts on the choice of the optimal transmission and coding technique. With the use of the mm-wave spectrum for 5G and beyond systems [1], required to cope with the ever-increasing demand for higher bitrates, the LOS condition becomes even more important due to the higher blocking effect of obstacles.

As such, LOS probability has gained importance as a key property in wireless channel prediction and simulation. Several statistical propagation models, such as the 'WINNER' model [2] and the ITU-R Recommendation P.1411 [3] are based on the definition of different path loss formulation as a function of the LOS or Non-LOS (NLOS) condition. LOS probability is likewise important in spectrum-sharing studies, such as the ones conducted within the CEPT and ITU-R. In these studies, adjacent bands are allocated to services operating in the same geographical area, leading to design systems where minimum interference must be provided to the incumbent or protected users while maximizing the number of users simultaneously accessing the same spectrum [4].

Overall, the development of suitable models to determine LOS condition in urban environments based on general characteristics such as building density, street width, and link distance, is very valuable for all those cases where accurate information about the environment layout is unavailable, or would be too difficult or time-consuming to determine.

Often, LOS probability is estimated through an empirical model fitted from some measurement data, and the output is typically a decreasing exponential function with the distance. The LOS probability is provided for typical environments namely Indoor Hotspot, Urban Macro, Urban Micro, and Rural Macro, as described in [5]. However, these models do not consider the actual geometry of the environment, like the building height or position in an urban scenario, or the antenna height. 3GPP reported a study on channels from $0.5\,\mathrm{GHz}$ to $100\,\mathrm{GHz}$ [6]. The report describes different environments (e.g., indoor office, street canyon, etc.), but all the functions are simply a negative exponential of the distance, sometimes including the height of the antennas. Instead, in [7] a LOS probability model based on stochastic geometry is developed, taking into account the geometry of the environment through an average height and length of the building. However, the model is quite complex and difficult to apply in real systems. In [8], a statistical method based on an objective parameterization of the environment is proposed. Still, extracting the necessary environmental statistics from open-source data is not yet completed.

Recently, Machine Learning (ML) algorithms have drawn attention in the field of electromagnetic propagation [9], leading to some initial successful applications. In the context of electromagnetic propagation, empirical and stochastic models often rely on closed-form formulas, whereas ML-based methods attempt to learn an arbitrarily complex nonlinear function from raw measurements. Another significant advantage of ML models is the inference speed: while the training phase may be computationally expensive (often due to very large datasets), querying the output of a trained model is typically computationally light. In this paper, ML techniques are utilized to retrieve a LOS probability model, using synthetic data from Ray Tracing (RT) as training and testing datasets.

The remainder of this paper is organized as follows. In Section II applications of ML in the context of electromagnetic propagation are briefly surveyed. In Section III the dataset generation procedure is outlined. In Section IV the LOS classification problem is described. In Section V our numerical results are presented and discussed. Section VI concludes the paper with main takeaways and ideas for future work.

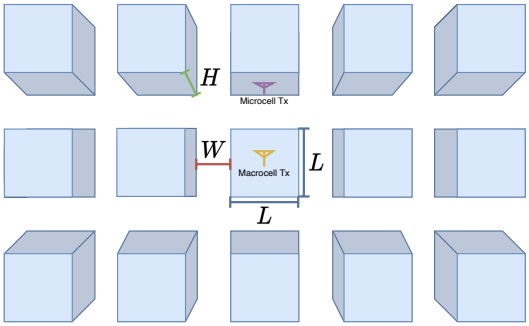


Fig. 1. Manhattan-like urban topology with street width W, building length L and building height H.

II. Related Work

ML in radio wave propagation has been traditionally employed for path loss (PL) estimation or for user localization [9]. For example, in [10] the RSSI data gathered from actual measurements are used for training an ML regression model for predicting the PL of a potential new device in the network. In [11] satellite images and measurements are used to train a deep neural network for PL regression. Results show more accurate path-loss prediction compared to a baseline RT algorithm and the 3GPP model [6]. In [12] an ML-based approach is developed for indoor localization via WiFi fingerprints. A similar approach is employed in [13] to achieve in-region localization. Finally, in [14] regression of probability for an indoor wireless link is shown to achieve values comparable to classical models.

In this paper, instead of regressing LOS probability as in [14], LOS prediction for individual radio links is considered. Predicting the LOS state for individual radio links yields more granular information than LOS probability, thus providing more flexibility for downstream applications. Indeed, the proposed approach can be employed for estimating the LOS probability, but can also be used for choosing the proper PL model (LOS or NLOS) given a specific radio link [8].

III. Dataset Generation

Several synthetic databases of Manhattan-like urban topologies have been generated according to the input file format required by the RT tool developed at the University of Bologna [15]. The LOS condition is assessed by checking the existence of an unobstructed direct ray between the Tx and the Rx: if the direct ray reaches the Rx, the Rx is in LOS condition.

We consider Manhattan-like urban topologies, as illustrated in Fig. 1. Buildings are assumed to be squared parallelepipeds with variable heights and lengths. Different simulation scenarios were defined based on building length, building height, and street width. Moreover, both urban macrocells and microcells are considered to evaluate the effectiveness of the ML algorithm in two diversified scenarios. In the macrocell case, the Tx is placed Δ meters above one of the buildings, while for microcells the Tx is placed below rooftop level, at 3m from the ground. The Rxs are always placed at ground level along the streets. The environmental parameters used for generating the Manhattan urban topologies for both macrocells and microcells RT simulations are reported in Table I.

TABLE I
Summary of the environmental parameters

Parameter	Values (macrocell)	Values (microcell)
Building height	15, 20, 25, 30 m	20 m
Building length	20, 30, 40, 50 m	20, 30, 40, 50 m
Street width	10, 15, 20, 25 m	10, 15, 20, 25 m
Δ / Tx height	2, 4, 6, 8 m	3 m

IV. Machine Learning for LOS Prediction

The LOS prediction problem has been formulated as a binary classification problem. Specifically, the goal is to train an ML model that, given a set of input features characterizing a Tx-Rx pair and the propagation environment, predicts whether or not the Tx and the Rx are in LOS condition.

In the considered scenario the model does not have access to the whole map, but only to a few descriptive features about the radio link and the Manhattan topology. Formally, a training, validation and test datasets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$, respectively, are considered. $\mathcal{D}_{\text{train}} = \{\mathbf{x_i}, y_i\}_{i=1}^{N}$ is a set of N training samples, where $\mathbf{x_i}$ are the features of the i-th sample, and $y_i \in 0, 1$ is a binary label (in this study LOS is set as the positive class). The following input features are adopted for simulations:

- the 3D Tx-Rx distance;
- the coordinates of the transmitter x_{Tx}, y_{Tx}, z_{Tx};
- the coordinates of the receiver x_{Rx} and y_{Rx}. The receiver height z_{Rx} is omitted as it is always assumed to be equal to $1.5m$, and is therefore uninformative;
- the building length, building height, and street width of the Manhattan-like topology.

In ML literature, data that can be organized in rows (in this case, one for each Tx-Rx pair) and columns (one for each feature) is referred to as "tabular". Empirically, the ML models that most of the time yield the best performance on tabular data are Gradient Boosting Decision Tree (GBDT) models [16]. Generally speaking, Gradient Boosting methods are function approximation algorithms that can optimize any differentiable loss function [17]. Hence, XGBoost [18] was used for implementing the binary classifier. Specifically, the GBDT model is trained by minimizing the logistic loss function averaged over all samples in the training set, as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \qquad (1)$$

Where $\hat{y}_i$ is the predicted probability for the positive class (i.e., LOS, in our case) given input features $\mathbf{x}_i$.

V. Illustrative Numerical Results

A. Experimental Setup

Manhattan-like topologies are randomly sampled from all possible combinations of parameters in Table I for building the training, validation, and test sets. Ten random urban topologies are sampled for training and two for validation, whereas the

remainder is kept for testing. Training is performed on a small fraction of our database, as the final goal is to assess the generalization capabilities of a model trained on a modest number of urban topologies. The loss on the validation set is monitored during the training phase, and at the end, the model providing the best validation loss is kept.

Three different training scenarios are considered:

1) **Macrocell only**: the model is trained only with data coming from RT simulations of macrocells.
2) **Microcell only**: the model is trained only with RT simulations of microcells.
3) **Macrocell+Microcell**: the model is trained with RT simulations of both microcells and macrocells.

As Tx-Rx pairs were uniformly generated over the Manhattan topologies, the number of NLOS samples in the datasets was significantly larger than the number of LOS samples (with a ratio of LOS over NLOS approximately equal to 0.03). It can be observed that feeding the classifier with a severely imbalanced dataset would lead to unsatisfactory performance, as the model will tend to overly favor the majority class (i.e., NLOS). Therefore, different simple countermeasures for imbalanced datasets (e.g., undersampling and oversampling) have been experimented with. After a grid search procedure, uniformly undersampling the NLOS samples down to twice the number of the LOS samples has been observed to provide the best performance on the validation set. Note that the validation and test datasets were not undersampled, as they must represent the true data distribution.

The model performance has been evaluated on Manhattan topologies not used for training. The Area Under the Precision-Recall Curve (AUCPR) has been adopted as a performance metric. AUCPR was chosen because the true data distribution is heavily skewed towards NLOS. As such, straightforward metrics such as the classification accuracy may strongly overestimate the model performance, e.g., a model that always predicts NLOS will show a deceitfully high accuracy because of the vast majority of samples being NLOS.

Formally, precision and recall are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where TP is the number of true positives (i.e., correctly classified LOS samples), FP is the number of false positives (i.e., NLOS samples mistaken as LOS), and FN is the number of false negatives (i.e., LOS samples mistaken as NLOS).

As the GBDT model outputs probabilities for LOS and NLOS, different values of precision and recall can be achieved by modifying the decision threshold, realizing the Precision-Recall curve. Tuning the decision threshold controls the trade-off between precision and recall (e.g., a model that always predicts LOS will have recall equal to 1 but precision near 0). Computing the area under the precision-recall curve (i.e., the AUCPR) provides therefore a fair and comprehensive score (upper-bounded by 1) of the model's performance. An illustrative PR curve and its AUCPR for "Macrocell only" training and macrocell test set are illustrated in Fig. 2.

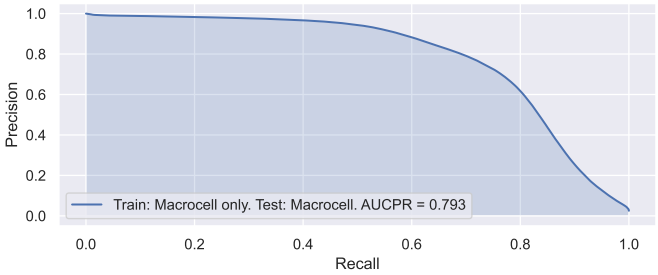


Fig. 2. Illustrative PR curve (solid line) and its corresponding AUCPR (shaded area) for a single train-test split in a "Macrocell only" training scenario and a macrocell test set.

TABLE II
VALUES AND 95% C.I. OF THE MEAN AUCPR ON MACROCELLS AND MICROCELLS TEST SETS

Training data	AUCPR - Macrocell	AUCPR - Microcell
Macrocell	$\mathbf{0.800 \pm 0.012}$	$\mathbf{0.845 \pm 0.023}$
Microcell	0.317 ± 0.003	0.756 ± 0.030
Macrocell+Microcell	0.742 ± 0.012	0.807 ± 0.020

B. Classification Performance

Mean and 95% confidence intervals for AUCPR are reported in Table II on both macrocells and microcells test sets. Mean values and confidence intervals are computed over fifty different combinations of training, validation, and test sets.

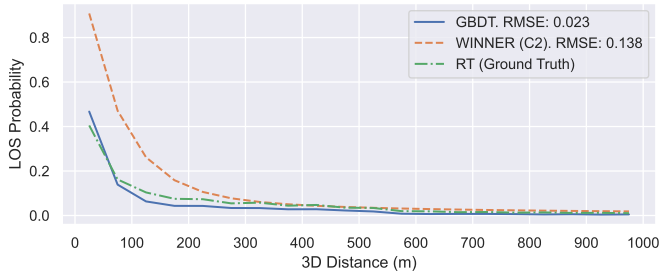
First, a model trained only on macrocell data achieves good generalization on both macrocell and microcell predictions on Manhattan topologies unseen during training. Conversely, a model trained only on microcell data is unable to generalize as well on macrocell data, exhibiting subpar performance. Finally, as expected, a model trained on macrocell and microcell data, expectedly, performs well on both macrocells and microcells.

Overall, the model yielding the best performance on the test set is the model trained only macrocell data. The model trained on macrocell-only data achieves higher AUCPR on microcells than the model trained specifically only on microcells[1]. This can be due to the lack of environmental diversity in the transmitter height in the microcell data, which drives the model to overfit. Conversely, the higher degree of environmental diversity provided by the macrocell data allows the model to learn more generalizable decision rules.

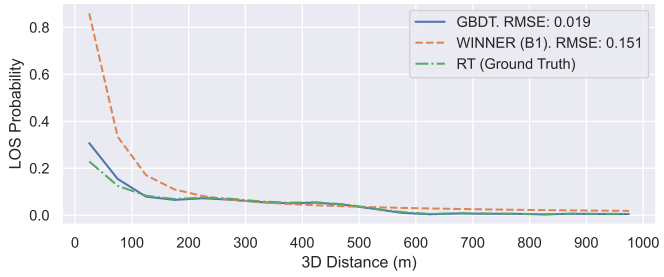
C. LOS Probability Modelling

As an illustrative application for our GBDT model, we show its effectiveness in providing approximate LOS probability models for Manhattan topologies unseen during training. We estimate the ground-truth LOS probability by running a Monte-Carlo simulation over an environment and averaging points over circular crowns with a $50m$ radius. On our hardware, running a Ray Tracing Monte-Carlo simulation for one Manhattan environment (namely, for 560000 Tx-Rx positions) took approximately six minutes, while our GBDT

[1]It is worth noting that AUCPR $\in [0.8, 0.85]$ signals a skilled classifier, far better than random guessing. As a reference, a dummy random classifier would achieve ≈ 0.025 AUCPR for both macrocells and microcells.



(a) Macrocell LOS Probability



(b) Microcell LOS probability

Fig. 3. LOS probability modeling on an illustrative Manhattan topology not used for training (building length: 30m, street width: 15m, building height: 20m, Tx height: 26m macrocell/3m microcell). WINNER [2] overestimates the LOS probability at low 3D distances for a narrow street width.

model evaluates the same number of Tx-Rx locations in less than one second on a Macbook Pro M1 CPU[2].

In Fig. 3 we present an illustrative comparison between our GBDT classifier, the WINNER [2] LOS probability analytical model, and the ground-truth from RT simulations for both macrocell and microcell data. We observe that our GBDT classifier provides a more truthful approximation of the ground-truth LOS probability compared to the analytical WINNER model. Overall, the GBDT classifier achieves an average test set RMSE equal to 0.020 and 0.029 for macrocells and microcells, respectively. In contrast, the WINNER model achieves an average 0.087 and 0.081 test set RMSE for macrocells (C2) and microcells (B1), respectively. While the WINNER formula is a function of the 3D distance only and can be in principle applied to any urban topology, ML allows building environment-specific, but also more accurate propagation models, exploiting richer features. Overall, our GBDT classifier stands in a middle ground between Ray Tracing (computationally expensive, but maximally accurate) and closed-form analytical models (potentially inaccurate, but computationally cheap).

D. Evaluating Feature Importance

Another angle for testing an ML model is to quantify the impact of the input features on the final prediction. If the criteria adopted by the model qualitatively align with rules dictated by common sense, we can deduce that the model learned reasonable decision rules.

[2]We underline that GBDT models are amenable to parallelization on both CPUs and GPUs, as Ray Tracing algorithms are.

SHapley Additive exPlanations (SHAP) is one of the most popular approaches for estimating feature attributions [19] and has been applied with success for distilling valuable insights in ML models trained on real physical layer traces [20]. SHAP aims to estimate the SHAP values given the input features. The magnitude of SHAP values conveys the feature importance, whereas their sign conveys whether the feature value drives the decision towards the positive (LOS) or the negative (NLOS) class.

Fig. 4 plots the SHAP values for each input feature as a function of the feature values. The most important feature for the final prediction are the Tx-Rx distance and the position coordinates of the Tx and the Rx. In particular, the influence of the Tx-Rx distance on the final prediction aligns with common sense: indeed, one expects that the greater the Tx-Rx distance, the less likely the two samples will be in LOS. The impact of the x and y coordinates of the transmitter and the receiver on the final prediction is not as straightforward as the Tx-Rx distance. Still, one can observe as follows: qualitatively, the model will tend towards NLOS if the two coordinates (either x_{Tx}, x_{Rx} or y_{Tx}, y_{Rx}) take very different values, vice-versa if they take similar values. In other words, if Tx and Rx are aligned as the streets in the urban layout, they are assigned a higher likelihood of being in LOS, which again aligns with common sense. These considerations will be further investigated by analyzing correlations between SHAP values. Furthermore, while being attributed less importance than the other coordinate features, the influence of the transmitter height on the predictions qualitatively aligns with common sense, i.e., the higher the transmitter, the greater the likelihood of being in LOS with the receiver.

The features characterizing the propagation environment are attributed the least importance for the final decision. The street width has the highest importance among the environment-specific features, and its influence on the final decision is still in agreement with common sense. Indeed, as the street width becomes larger, the greater the likelihood of the Tx and the Rx being in LOS. Similarly, as the building length becomes narrower, the model will tend to favor LOS over NLOS. The building height is attributed the least importance, but we can observe that the model will tend to favor NLOS for higher buildings, which is again consistent with common sense. Likely, the building height is attributed the least importance because (for macrocells) the building height is strongly correlated with z_{Tx}. When presented with heavily correlated features, decision trees will place the most importance on only one of them.

While the plot in Fig. 4 highlights the contribution of individual features, the final prediction results from a (generally nonlinear) interaction between features. It is therefore of great interest to assess, at least qualitatively, which and how much different features interact. To this end, Fig. 5 illustrates the absolute correlation coefficient between the absolute SHAP values (i.e., feature importance).

As conjectured, the importance of the Tx-Rx x and y coordinates are strongly correlated. As the coordinates of the

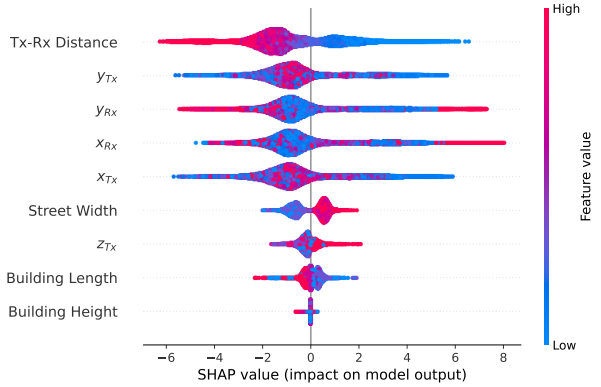


Fig. 4. Summary plot of the GBDT classifier displaying feature importance as a function of feature values for an illustrative macrocell test set.

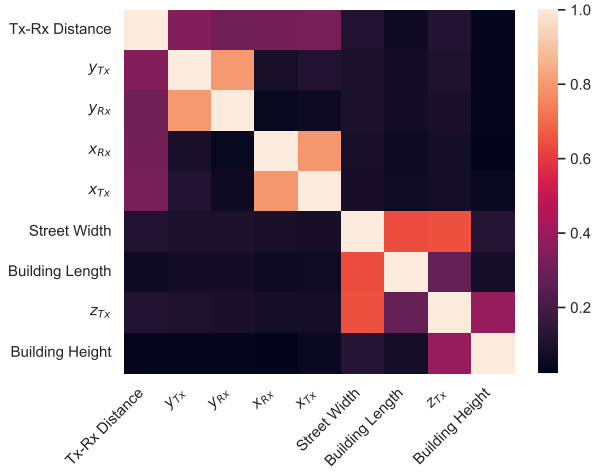


Fig. 5. Correlation between absolute SHAP values (i.e., feature importance) of the GBDT classifier for an illustrative macrocell test set.

Tx and Rx are independent, feature importance correlation means that the model successfully learned to take into account the relative Tx-Rx positions before making a decision.

The most concerning insight is perhaps the complete lack of correlation between the position-specific feature importance and the environment-specific feature importance. While the model correctly captured the influence of the environmental parameters on the LOS/NLOS states, it was unfortunately unable to capture subtler relationships between the environmental parameters and the Tx-Rx locations. These limitations, made apparent by feature importance estimation, are likely the primary source of inaccuracies in the model.

VI. Conclusion

We proposed a baseline GBDT classification model for determining whether or not a transmitter and a receiver are in LOS condition based only on a few high-level descriptive features of the radio link. We illustrated that our GBDT classifier achieves fair performance despite the lack of information, providing more truthful LOS probability approximations than the closed-form WINNER model. By analyzing feature importance, we qualitatively assessed that the baseline classifier learned simple decision rules aligned with common sense, while unfortunately failing to fully grasp more complex spatial relationships. Future work will investigate modeling the LOS condition based on unstructured data (e.g., images), and the knowledge transfer between LOS/NLOS classification models in radically different propagation environments.

References

[1] Y. Kim *et al.*, "New radio (NR) and its evolution toward 5G-advanced," *IEEE Wirel. Commun.*, vol. 26, no. 3, pp. 2–7, Jun. 2019.

[2] P. Kyösti *et al.*, "IST-4-027756 WINNER II D1.1.2 v1.2 - WINNER II channel models," *Inf. Soc. Technol*, vol. 11, 02 2008.

[3] ITU-R, "Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz," 2021.

[4] C. R. Anderson, "An integrated terrain and clutter propagation model for 1.7 GHz and 3.5 GHz spectrum sharing," *IEEE Trans. Antennas Propag.*, vol. 70, no. 7, pp. 5804–5818, Jul. 2022.

[5] "guidelines for evaluation of radio interface technologies for IMT-2020 - ITU."

[6] "3GPP TR 38.901 version 16.1.0 release 16: Study on channel model for frequencies from 0.5 to 100 GHz."

[7] X. Liu, J. Xu, and H. Tang, "Analysis of frequency-dependent line-of-sight probability in 3-D environment," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1732–1735, Aug. 2018.

[8] R. Rudd, "Statistical modelling of short-range interference paths," in *Proc. of 2022 16th European Conference on Antennas and Propagation (EuCAP)*, 2022, pp. 1–4.

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

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

[11] 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.

[12] J.-W. Jang and S.-N. Hong, "Indoor localization with WiFi fingerprinting using convolutional neural network," in *Proc. of 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2018, pp. 753–758.

[13] A. Brighente, F. Formaggio, M. Centenaro, G. M. Di Nunzio, and S. Tomasin, "Location-verification and network planning via machine learning approaches," in *Proc. of 2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, 2019, pp. 1–7.

[14] W. Yang, J. Zhang, and J. Zhang, "Machine learning based indoor line-of-sight probability prediction," in *Proc. of 2019 International Symposium on Antennas and Propagation (ISAP)*, 2019, pp. 1–3.

[15] E. M. Vitucci, V. Degli-Esposti, F. Fuschini, J. Lu, M. Barbiroli, J. Wu, M. Zoli, J. Zhu, and H. Bertoni, "Ray tracing RF field prediction: An unforgiving validation," *Int. J. Antennas Propag.*, vol. 2015, pp. 1–11, Aug. 2015.

[16] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," in *Proc. of 8th ICML Workshop on Automated Machine Learning (AutoML)*, 2021.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *Ann. Stat.*, vol. 29, no. 5, pp. 1189 – 1232, Oct. 2001. [Online]. Available: https://doi.org/10.1214/aos/1013203451

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, p. 785–794.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 4768–4777.

[20] O. Karandin *et al.*, "If not here, there. explaining machine learning models for fault localization in optical networks," in *Proc. of 2022 International Conference on Optical Network Design and Modeling (ONDM)*, 2022, pp. 1–3.