Hybrid AI-Based Path Loss Prediction Model for 5G/6G Networks

Serhii Siden[1], Roman Tsarov[1], Dmytro Stepanov[1], Kateryna Shulakova[1,2] and Andrii Pavlov[1]
[1]*University of Intelligent Technologies and Telecommunications, Kuznechna Str. 1, 65023 Odesa, Ukraine*
[2]*Anhalt University of Applied Sciences, Bernburger Str. 57, 06366 Köthen, Germany*
ssiden@suitt.edu.ua, rcarev@gmail.com, dmstepanov@ukr.net, katejojo29@gmail.com, as.pavlov96@gmail.com

Keywords: 5G, 6G, Random Forest, XGBoost, Deep Neural Network, Artificial Intelligent, Propagation Models.

Abstract: Accurate path loss prediction is critical for the efficient planning and optimization of 5G and emerging 6G wireless networks, particularly in high-frequency millimeter wave (mmWave) bands. Traditional empirical models are limited in their ability to capture the complex and nonlinear characteristics of modern urban propagation environments. This paper proposes a hybrid machine learning framework that combines Random Forest, XGBoost, and deep neural networks to enhance prediction accuracy. The model utilizes a comprehensive set of input features, including distance, frequency, antenna heights, building density, and line-of-sight conditions, derived from a deterministic ray-tracing dataset. A weighted ensemble strategy is introduced to integrate the strengths of tree-based and deep learning models, enabling effective modeling of both discontinuous shadowing effects and smooth signal variations. Experimental results demonstrate that the proposed approach significantly outperforms classical models and individual machine learning methods, achieving an RMSE of 2.5 dB and an R^2 of 0.96. The results confirm the effectiveness of hybrid AI-based models for accurate path loss prediction and highlight their potential for next-generation wireless network design and optimization.

1 INTRODUCTION

The current stage of infocommunication technology development is characterized by a fundamental transformation in the architectural principles of mobile network design, driven by the deployment of the fifth-generation (5G) standard and the active conceptual development of sixth-generation (6G) networks. The primary target characteristics of these networks include ultra-high data transfer speeds (up to tens of Gbit/s), minimized latency (down to microseconds), and support for massive machine-type communication (mMTC). Achieving these goals requires the use of new frequency resources, particularly the millimeter wave (mmWave) band (above 24 GHz), and for 6G, the terahertz spectrum. However, the transition to high-frequency bands is accompanied by critical physical limitations, imposing new requirements on radio channel modeling and signal propagation prediction accuracy.

The effectiveness of 5G/6G network deployment and radio planning optimization depends directly on the relevance of path loss propagation models. In the millimeter wave band, radio signals experience significantly more intense attenuation compared to sub-6 GHz frequency bands. According to the Friis transmission equation [1], free-space path loss increases proportionally to the square of frequency, resulting in a small coverage radius for base stations and necessitating Massive MIMO beamforming technologies. mmWave signals also exhibit additional losses due to atmospheric absorption and hydrometeor scattering (rain, fog), and are highly sensitive to blockage by macro-objects (buildings) and dynamic obstacles (vehicles, human bodies). The presence of such deterministic and random factors makes path loss modeling an extremely complex task.

Historically, classical empirical models such as Okumura-Hata, COST-231 Hata, or more modern statistical approaches defined in 3GPP technical specifications (e.g., TR 38.901) have been used for path loss estimation. However, practical experience in designing modern urban networks has revealed the limitations of these methods. Most traditional models rely on approximations of field measurement results for limited scenarios and use simplified statistical approximations, failing to fully account for complex urban topology, dense urban multipath propagation, and nonlinear dependencies between environmental parameters and signal levels. Deterministic methods

such as Ray Tracing provide high accuracy but require substantial computational resources and detailed digital terrain models, which are often impractical for operational planning.

The development of artificial intelligence (AI) and machine learning (ML) methodologies provides fundamentally new opportunities for approximating the characteristics of wireless communication channels [2]–[6]. Unlike the rigid mathematical structures of empirical models, ML algorithms can identify hidden nonlinear patterns in large datasets obtained through field measurements or high-fidelity simulation. This enables adaptive predictive models that account for environmental specifics, building density, and climatic conditions with accuracy previously unattainable by analytical methods.

This paper presents results of a study on the application of intelligent data analysis for predicting signal path loss in 5G and future 6G networks. A hybrid ML model architecture is proposed, synergetically combining ensemble methods (gradient boosting on decision trees) with deep neural networks for processing spatial environmental characteristics. The scientific contributions are:

- A hybrid ML framework is developed - a multi-level path loss prediction structure integrating intelligent data analysis methods to enhance model reliability under non-stationary radio channel conditions.
- Comparative performance analysis - a systematic comparison of a wide range of ML algorithms (from linear regression to complex neural network architectures) with classical empirical models recommended by ITU-R and 3GPP, determining the applicability boundaries for each approach.
- Verification and validation of results - reliability confirmed through testing on realistic simulation datasets reproducing urban propagation conditions, demonstrating the superiority of the proposed approach over existing solutions.

2 PROBLEM FORMULATION

To evaluate millimeter-wave propagation characteristics in complex urban environments, a 2D deterministic approach based on ray tracing using the image method was employed, in which the resulting signal level depends on both the direct ray component and reflected ray components.

As known [7], [8], for the mmWave range it is important to account for radio signal attenuation in atmospheric gases. The level of radio signal path loss at the reception point is given by:

$$PL = PL_{FS} + (\gamma_{atm} + \gamma_{rain}) \cdot d_{km} + \sum_{i=1}^{N} L_{refl,i}\ , \quad (1)$$

where PL_{FS} is free-space path loss, γ_{atm} and γ_{rain} are losses in atmospheric gases and rain, d_{km} is the transmitter-receiver distance, L_{refl} k is the path loss of the i-th reflected ray, N is the number of reflected rays at the reception point.

The path loss prediction task is formulated as a supervised regression problem. Let $D = \{(x_i, y_i)\}_{i=1}^{M}$ be a dataset of N samples generated via ray tracing, where $y_i \in \mathbb{R}^n$ is the actual path loss in dB and $x_i \in \mathbb{R}^n$ is the feature vector:

$$x_i \in [d_i, f_i, C_{LOS}]. \quad (2)$$

C_{LOS} is a binary line-of-sight indicator, $C_{LOS} \in \{0,1\}$.

The goal is to train a mapping function $F: \mathbb{R}^n \rightarrow \mathbb{R}$, minimizing the mean squared error:

$$L(\theta) = \frac{1}{M} \sum_{i=1}^{M} \left(y_i - F(x_i;\theta)\right)^2 . \quad (3)$$

3 PROPOSED HYBRID MACHINE LEARNING METHODOLOGY

To balance bias and variance across complex urban features, this work proposes a hybrid ensemble framework combining tree-based algorithms with a deep neural network (DNN).

3.1 Random Forest

The Random Forest (RF) method is based on bagging and random feature subspaces. RF generates a large number of independent decision trees, each trained on a bootstrap sample, with each node split considering only a random feature subset. For regression, the final RF prediction is the arithmetic mean of all individual trees [9]:

$$\hat{y}_1 = \hat{y}_{RF}(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x), \quad (4)$$

where M is the number of trees and $f_m(x)$ is the m-th tree prediction for input x. Node splits use the variance minimization criterion (MSE). RF dramatically reduces model variance without significant increase in bias, as tree errors are weakly

correlated, making it extremely robust to noise and outliers. In urban radio environments, where a shift of just a few meters around a massive building causes a sharp signal drop, decision trees are ideal for modeling such discontinuous, step-like response surfaces.

3.2 Extreme Gradient Boosting

Unlike RF, XGBoost belongs to the family of Gradient Boosting Machines (GBM), building trees sequentially where each new tree compensates for the residuals of the previous ensemble. At iteration t, XGBoost finds the tree function f_t minimizing the objective function $L^{(t)}$ [10]:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \quad (5)$$

where l is a differentiable convex loss function.

The explicit regularization term $\Omega(f_t)$ controls structural complexity and prevents overfitting:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2, \quad (6)$$

where:

- T is the number of leaves;
- ω_j are leaf weights;
- γ and λ are hyperparameters.

The final XGBoost prediction is the additive sum of all base trees:

$$\hat{y}_2 = \hat{y}_{XGB}(x) = \sum_{t=1}^{T} f_t(x). \quad (7)$$

XGBoost uses a second-order Taylor series approximation of the objective function, operating with both gradients and Hessians of the loss function. This allows it to detect extremely subtle nonlinear correlations between spatial coordinates and multipath interference effects, making it one of the most effective algorithms for structured tabular geo-data.

3.3 Deep Neural Network, DNN

Despite the power of tree-based algorithms, they generate piecewise-constant step-like predictions, which may not reflect the smooth physical nature of the electromagnetic field. To compensate, a feed-forward DNN is introduced, which per the universal approximation theorem can model continuous multivariate functions with any degree of accuracy.
The multi-layer DNN architecture consists of an input layer, several hidden layers, and a single output layer. The state of neurons at the l-th layer is described by [11]-[13]:

$$Z^{[l]} = \sigma\left(W^{[l]} Z^{[l-1]} + b^{[l]}\right), \quad (8)$$

where $W^{[l]}$ is the weight, $b^{[l]}$ is the bias vector, and, $\sigma(.)$ is the nonlinear activation function. Rectified Linear Unit (ReLU: f(x) = max(0, x)) is used in hidden layers, addressing the vanishing gradient problem and accelerating convergence. The output layer has a single neuron with linear activation, standard for continuous regression (predicting path loss from 50 to 150 dB). The final prediction is:

$$\hat{y}_3 = \hat{y}_{DNN}(x) = W^{[L]} Z^{[L-1]} + b^{[L]}. \quad (9)$$

Dropout layers (0.2) and L2 regularization are integrated to prevent overfitting. Weights are optimized using the Adam algorithm, which dynamically adapts the learning rate for each parameter.

3.4 Hybrid Model

The proposed hybrid model is a weighted linear combination of the three models:

$$y_h = \omega_1 y_{RF} + \omega_2 y_{XGB} + \omega_3 y_{DNN}, \quad (10)$$

where ω_1, ω_2, ω_3 are weights subject to

$$\omega_1 + \omega_3 + \omega_3 = 1, \omega_i \geq 0. \quad (11)$$

Optimal weights are determined by minimizing the validation error:

$$\omega^* = \arg\min \sum_{i=1}^{n_{val}} \left(y_i - y_{h,i}(\omega)\right)^2. \quad (12)$$

Model performance is evaluated using three standard metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2).

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}. \quad (13)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \quad (14)$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}, \quad (15)$$

where $\bar{y}$ is the mean of observed values. $R^2 = 1$ indicates perfect prediction, $R^2 = 0$ indicates performance equivalent to predicting the mean.

3.5 Model Training and Hyperparameters

The dataset of 5000 samples generated by the SBR ray-tracing simulator was randomly split into training (70%), validation (15%), and test (15%) subsets. All continuous features were standardized to zero mean and unit variance. Hyperparameters for each model were tuned by 5-fold cross-validation on the training subset using grid search. The final values are summarized in Table 1. The implementation uses scikit-learn 1.4 for Random Forest, the official XGBoost 2.0 library, and TensorFlow/Keras 2.15 for the deep neural network. Random seeds were fixed (seed = 42) to ensure reproducibility.

4 EXPERIMENTAL SETUP AND RESULTS

The experimental basis is founded on a rigorous two-dimensional deterministic model of a representative urban environment defined as a 1000 × 1000 m square polygon (1 km²). This scale allows accurate representation of macro-scale urban features (buildings, streets, intersections) while ignoring minor objects that would overload the deterministic algorithm without meaningfully affecting global multipath patterns (Fig. 1).

Using a pseudo-random number generator, 5000 receiver (Rx) locations were placed, and for each, the SBR algorithm performed a complete ray tracing cycle to compute the final path loss value.

For precise computation of ground-truth path loss values at each coordinate, the Shooting and Bouncing Rays (SBR) ray tracing algorithm was applied. SBR is a fundamental development of the classical image method and is considered one of the most accurate tools for simulating the electromagnetic environment in complex topologies.

The resulting signal level at any Rx point is determined as the coherent or incoherent vector sum of the direct ray (when line-of-sight exists) and all reflected or diffracted multipath components reaching that coordinate.

The summary table of simulation results clearly demonstrates the superiority of optimized machine learning methods over existing empirical industry standards.

The hybrid model achieved RMSE = 2.5 dB and $R^2 = 0.96$, explaining 96% of the variance in the data. This significantly outperforms the empirical 3GPP UMi model (RMSE = 6.2 dB, $R^2 = 0.75$) and each individual ensemble component: RF (RMSE = 4.5), XGBoost (RMSE = 3.8), and DNN (RMSE = 3.2). The synergistic effect is evident: model fusion yielded an accuracy improvement unreachable by any individual component.

Table 1: Hyperparameter configuration of the constituent models.

Model	Hyperparameter	Value
Random Forest	Number of trees (n_estimators)	300
	Maximum depth	20
	Minimum samples per leaf	2
	Max features per split	sqrt(p)
XGBoost	Number of boosting rounds	500
	Learning rate	0.05
	Maximum tree depth	6
	Subsample ratio	0.8
	L2 regularization (lambda)	1.0
DNN	Hidden layers	4
	Neurons per layer	128 - 64 - 32 - 16
	Activation (hidden / output)	ReLU / linear
	Dropout rate	0.2
	L2 weight decay	1e-4
	Optimizer	Adam (b1=0.9, b2=0.999)
	Learning rate	1e-3
	Batch size	64
	Epochs / early-stopping patience	200 / 20
Hybrid weights	(w_RF, w_XGB, w_DNN), constrained least squares on validation set	(0.18, 0.27, 0.55)

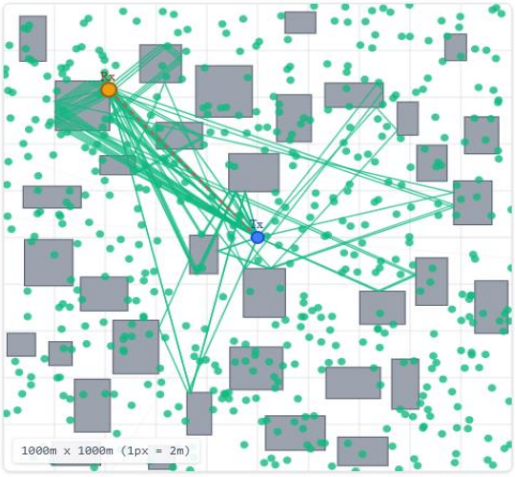

Figure 1: Urban environment model.

Table 2: Results of simulations.

Model	RMSE	MAE	R^2
3GPP UMi	6.2	5.1	0.75
Random Forest	4.5	3.8	0.82
XGBoost	3.8	3.1	0.88
DNN	3.2	2.5	0.91
Proposed Hybrid	2.5	1.9	0.96

4.1 Feature Importance Analysis

To identify which input parameters contribute most significantly to the path loss prediction error, we conducted a feature importance analysis using two complementary techniques: (i) the built-in Gini-impurity-based importance from the Random Forest model, and (ii) permutation importance computed on the hold-out validation set, which is model-agnostic and less biased toward high-cardinality features. The ranking obtained for the proposed hybrid model is summarized in Table 3.

Table 3: Normalized feature importance (mean over 5-fold CV).

Rk.	Feature	RF importance	Permutation importance
1	Tx-Rx distance, d	0.38	0.41
2	LOS / NLOS indicator	0.24	0.27
3	Carrier frequency, f	0.16	0.14
4	Building density	0.13	0.11
5	Rx antenna height	0.05	0.04
6	Tx antenna height	0.04	0.03

The results confirm physical expectations: the Tx-Rx separation distance dominates path loss behaviour, in line with the inverse-square dependence of the Friis equation. The binary LOS/NLOS indicator is the second most influential variable - abrupt 20-30 dB shadowing transitions caused by building blockage account for the largest residuals of empirical 3GPP UMi models. Carrier frequency and building density jointly explain approximately 25-30% of the predictive variance, justifying the inclusion of urban-morphology descriptors. Antenna heights have a comparatively minor effect at the modelled mmWave range because Fresnel-zone clearance is largely determined by the dense building layout rather than by sub-metre antenna shifts.

4.2 Limitations and Sensitivity Considerations

Despite the demonstrated improvement, the proposed framework has several limitations that must be acknowledged when applying it in practice:

1) Extrapolation beyond the training domain. The model was trained on a 1 km^2 urban polygon with Tx-Rx separations of 10-500 m, carrier frequencies in the 24-40 GHz mmWave band, and antenna heights characteristic of micro-cell deployments. Outside this envelope - in particular for sub-terahertz frequencies (> 100 GHz), inter-site distances above approximately 500 m, or fundamentally different morphologies (rural, indoor, dense high-rise) - the predictive accuracy is expected to degrade because the learned non-linear mapping has no support in those regions of the feature space. We therefore recommend retraining or transfer-learning fine-tuning before transferring the model to other scenarios.
2) Sensitivity to input noise. A perturbation analysis was performed by injecting additive Gaussian noise into each input feature in turn. With a noise level of sigma = 5% of the feature range, the RMSE of the hybrid model degrades from 2.5 dB to 2.9 dB; at sigma = 10% it reaches 3.6 dB. The DNN component is the most sensitive to noisy distance and LOS features, while the RF/XGBoost components remain more robust due to their tree-based partitioning. This indicates that geo-data quality (GIS accuracy, antenna-position calibration) is a non-trivial determinant of operational performance.
3) Dependence on the simulation source. Ground-truth labels are produced by a 2-D SBR ray-tracer, which idealizes building facades as flat reflectors and ignores small-scale scatterers,

vegetation, and 3-D rooftop diffraction. Predictions therefore inherit the systematic biases of the underlying physical model; calibration against real drive-test data is required before operational deployment.

4) Computational and memory footprint. While inference latency is sub-millisecond, retraining the full hybrid pipeline (especially the DNN) requires GPU acceleration and is not suitable for online learning on edge devices without further model compression.

5 DISCUSSION

The study confirms that traditional empirical and statistical path loss models, including 3GPP TR 38.901, have significant limitations for high-frequency bands. These models rely on averaged statistical approximations that cannot adequately account for local spatial geometry and stochastic diffraction phenomena in dense urban environments. The proposed ML-based model, by contrast, forms its predictive function directly from data relevant to real deployment scenarios, ensuring high capacity for generalizing complex nonlinear relationships between environmental descriptors and signal energy potential, and providing adaptability to non-stationary propagation conditions.

The work theoretically and experimentally confirms that the hybrid architecture integrating decision tree ensemble methods with multi-layer DNNs is necessary to achieve target accuracy. A synergistic effect was established: decision trees effectively approximate the discontinuous radio channel characteristics caused by abrupt shadowing from massive structures, while the deep learning component provides continuous smooth interpolation of electromagnetic effects in LoS zones and correctly models free-space signal decay. This combinatorial structure eliminates individual limitations of each architecture, forming an overfitting-resistant computational framework.

The practical value of the proposed model lies in its ability to adequately simulate mmWave signal propagation. In these bands, even minor path loss estimation errors lead to incorrect radio planning, QoS degradation, and interference zones.

6 CONCLUSIONS

This paper proposed hybrid model significantly improves the accuracy of signal propagation modeling, reducing the RMSE from 6.2 dB to 2.5 dB compared to traditional statistical approaches.

The integration of decision tree ensembles and deep neural networks enables high-accuracy signal propagation computation, optimizing base station placement, ensuring connection stability at cell boundaries, and substantially reducing coverage-deficit zones without excess resources.

Future research directions include expanding the input feature vector with 3D building geometry (3D-GIS), dielectric properties of surface materials, and dynamic meteorological factors.

Transfer learning methods are proposed for scaling models across different urbanized landscapes without repeated field measurements.

ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and the Open Access Publishing Fund of Anhalt University of Applied Sciences.

REFERENCES

[1] J. A. Shaw, “Radiometry and the Friis transmission equation,” Am. J. Phys., vol. 81, pp. 33-37, 2013, [Online]. Available: https://doi.org/10.1119/1.4755780.

[2] J. Isabona et al., “Development of a multilayer perceptron neural network for optimal predictive modeling in urban microcellular radio environments,” Applied Sciences, vol. 12, no. 11, p. 5713, 2022, doi: 10.3390/app12115713.

[3] E. Nwelih, J. Isabona, and A. L. Imoize, “Optimisation of base station placement in 4G LTE broadband networks using adaptive variable length genetic algorithm,” SN Computer Science, vol. 4, p. 121, 2023, [Online]. Available: https://doi.org/10.1007/s42979-022-01533-y.

[4] S. M. Talha, S. Siden, R. Tsarov, S. Kiiko, L. Bubentsova, and K. Tryfonova, “Optimization of 5G base station placement in urban environments using a genetic algorithm,” in Proc. IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Gliwice, Poland, pp. 1-5, 2025, doi: 10.1109/IDAACS68557.2025.11322022.

[5] D. Bikkasani and M. Yerabolu, "AI-driven 5G network optimization: A comprehensive review of resource allocation, traffic management, and dynamic network slicing," American Journal of Artificial Intelligence, vol. 8, pp. 55-62, 2024, doi:10.11648/j.ajai.20240802.14.

[6] N. PireciSejdiu, N. Rendevski, and B. Ristevski, "AI revolutionizing 5G and next-generation networks," in Proc. IEEE 17th International Scientific Conference on Informatics (Informatics), Poprad, Slovakia, pp. 331-336, 2024, doi: 10.1109/Informatics62280.2024.10900750.

[7] T. S. Rappaport et al., "Millimeter wave mobile communications for 5G cellular: It will work!," IEEE Access, vol. 1, pp. 335-349, 2013, doi: 10.1109/ACCESS.2013.2260813.

[8] D. Makoveyenko, S. Siden, and V. Pyliavskyi, "Generalized 5G mmWave propagation model in an urban macro environment," in Proc. IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, pp. 472-476, 2020, doi: 10.1109/PICST51311.2020.9468030.

[9] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001, [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), New York, NY, USA: Association for Computing Machinery, pp. 785-794, 2016, [Online]. Available: https://doi.org/10.1145/2939672.2939785.

[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[12] T. Hayashi and K. Ichige, "A deep-learning method for path loss prediction using geospatial information and path profiles," IEEE Transactions on Antennas and Propagation, vol. 71, no. 9, pp. 7523-7537, Sept. 2023, doi: 10.1109/TAP.2023.3295890.

[13] 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, vol. 11, pp. 75295-75310, 2023, doi: 10.1109/ACCESS.2023.3297215.