

Simulation-Enhanced Data Augmentation for Machine Learning Pathloss Prediction

Ahmed P. Mohamed*, Byunghyun Lee*, Yaguang Zhang*, Max Hollingsworth[†], C. Robert Anderson[‡],
James V. Krogmeier*, David J. Love*

*Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907

[†]University of Colorado, Boulder, CO 80309

[‡]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061

Email: {mohame23, lee4093, ygzhang, jvk, djlove}@purdue.edu, maho5047@colorado.edu, chanders@vt.edu

Abstract—Machine learning (ML) offers a promising solution to pathloss prediction. However, its effectiveness can be degraded by the limited availability of data. To alleviate these challenges, this paper introduces a novel simulation-enhanced data augmentation method for machine learning (ML) pathloss prediction. Our method integrates synthetic data generated from a cellular coverage simulator and independently collected real-world datasets. These datasets were collected through an extensive measurement campaign in different environments, including farms, hilly terrains, and residential areas. This comprehensive data collection provides vital ground truth for model training. A set of channel features was engineered, including geographical attributes derived from LiDAR datasets. These features were then used to train our prediction model, incorporating the highly efficient and robust gradient boosting ML algorithm, CatBoost. The integration of synthetic data, as demonstrated in our study, significantly improves the generalizability of the model in different environments, achieving a remarkable improvement of approximately 12 dB in terms of mean absolute error for the best-case scenario. Moreover, our analysis reveals that even a small fraction of measurements added to the simulation training set, with proper data balance, can significantly enhance the model's performance.

I. INTRODUCTION

Radio signals experience pathloss as they propagate to a receiver. Pathloss refers to the attenuation of the signal of a communication link between the transmitter and the receiver. Accurately estimating pathloss is fundamental for coverage estimation and interference analysis, which are key to effective network planning. Additionally, accurate pathloss prediction enables improved mobility management, such as handover and quality-of-service prediction. However, predicting pathloss is challenging because it depends on various factors, including propagation distance, geometry (e.g., buildings and trees), antenna pattern and carrier frequency [1].

The fundamental model for the prediction of pathloss is the Friis equation, which calculates the loss of transmission in the free space based on the distance and carrier frequency [2]. Terrain-based models improve accuracy by incorporating topological features. The irregular terrain model (ITM), often referred to as the Longley-Rice model, aims to predict pathloss considering terrain profiles [3]. The extended Hata (eHata) model adds environment categories, such as urban, suburban or rural settings, to account for different endpoint propagation environments [4]. However, in many practical scenarios where

multiple environments are combined, these models may fail to provide an accurate pathloss prediction.

Stochastic models add a random variable to a deterministic pathloss model to describe the randomness (e.g., scattering and multipath effect) in a wireless link. The most widely used model is the log-normal shadowing model, which accounts for shadowing with a Gaussian random variable. Furthermore, the COST 231 model [5], 3GPP spatial channel model [6], and QuadRiGa [7] empirically model the shadowing distribution. These models are useful in that they can be easily employed without high complexity and provide a general understanding of signal propagation characteristics in various environments. However, since these models do not consider the exact surroundings and environments of the transmitter and receiver, they may not capture site-specific features.

Ray-tracing models simulate the propagation of electromagnetic waves using deterministic modeling of reflection, diffraction, and scattering [8], [9]. Ray-tracing has been used extensively for research purposes as it offers detailed information about the structure of a wireless channel, including angle of arrival, angle of departure, and path delays. Despite this, the computational cost of ray-tracing becomes unaffordable as the simulation scale increases, making it less suitable for large-scale implementations.

To overcome these issues, learning-based models have attracted significant research interest [10]–[14]. Despite their ability to learn site-specific features, the prediction performance of learning-based models is largely based on the availability of extensive high-quality data. Nonetheless, the collection of such datasets poses challenges as it involves considerable human labor and sophisticated measurement tools. Although driving tests are often used to streamline this process, they are not practical in locations such as farms and school campuses, which can bias the data sets to publicly accessible roads. Due to the lack of data, a learning-based prediction model may struggle to generalize these unfamiliar radio environments. This is a major issue for rural wireless communication applications, where diverse propagation environments coexist in a wide area [15].

In the machine learning (ML) domain, simulation-assisted data augmentation has been actively investigated due to its cost-effectiveness and convenience [16], [17]. However, the use of simulation-aided data augmentation remains relatively unexplored in the domain of wireless communications, and more specifically, in pathloss prediction. In [18], the authors

This work is supported by the National Science Foundation under grants EEC-1941529, CNS-2212565, and CNS-2225578. (Ahmed P. Mohamed and Byunghyun Lee contributed equally to this work.)

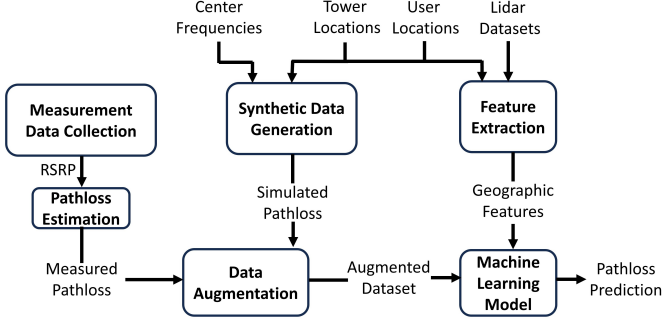


Figure 1: Flowchart of the proposed simulation-enhanced data augmentation process.

used simulation data to cover inaccessible geographic regions for site-specific channel modeling. However, this work focused on augmenting a partial dataset for a single geographic area rather than augmenting a dataset with different environments.

In this paper, we introduce a novel data augmentation method for ML pathloss prediction, which incorporates both real and synthetic data. Our goal is to improve the generalizability and reliability of the pathloss prediction model by enriching the dataset with synthetic data, especially in cases where measurement data are limited. The proposed data augmentation, as shown in Fig. 1, has three main components: (i) measurement data collection, (ii) synthetic data generation, and (iii) feature extraction. For the collection of measurement data, we conducted a measurement campaign in three different environments (rural, residential, and hilly). We then produce synthetic data using the state-of-the-art large-scale pathloss simulator [19]. This simulator utilizes high-resolution LiDAR data to extract features along line-of-sight (LoS) paths, enabling the prediction of large-scale fading. We then develop features based on domain knowledge and extract them from LiDAR-based geographic datasets. We evaluated the prediction performance of the augmented dataset in various scenarios. The results show that the combination of synthetic and real data can improve the prediction performance for unseen environments at a slight cost of loss of accuracy for known environments.

II. SIMULATION-ENHANCED DATA AUGMENTATION

Learning-based models often suffer from inaccurate predictions in unfamiliar or unseen environments. To address this problem, this paper aims to build a robust model that can perform well not only for known areas but also for unseen scenarios by incorporating synthetic data. Specifically, we detail the data processing procedure for merging synthetic data with measurements, as well as the data collection procedure. In this section, we will outline our proposed data augmentation methodology and then describe the procedures for creating both measurement and synthetic datasets.

A. Methodology

Fig. 1 illustrates the overall flow of the proposed simulation-enhanced data augmentation. The three main components of the proposed simulation-enhanced data augmentation can be described as follows:

- **Measurement Data Collection:** Measurement data are collected from mobile phones in different environments.

These measurements can be used to obtain pathloss values and associated data, such as location, cell information, and carrier frequency.

- **Synthetic Data Generation:** The simulation parameters such as user location and carrier frequency should be determined based on what the measurement dataset lacks, to enhance the dataset. For example, if the dataset lacks rural environment data, simulations for rural environments can be performed using rural geographic data and macrocell parameters.
- **Feature Extraction:** After preparing the measurement and synthetic datasets, we extract the features of the data points from the geographic datasets, which are predominantly site-specific. We carefully design features that reflect the geometry and surroundings of the transmitter and receiver using domain knowledge. Then, we generate features for every geographic point in the area of interest and pair them with the corresponding pathloss values collected in the previous step.

B. Creating Measurement Datasets

In this section, we will describe how we collected datasets in different environments and the process of converting the Reference Signal Received Power (RSRP) measurements into path loss values.

1) Collecting RSRP Measurements

We carried out a comprehensive data collection in the city of West Lafayette, located in the state of Indiana in the USA. We collected data from three different environments: farm, hilly and residential areas, as illustrated in Fig. 2. We conducted 4G LTE measurements using three commodity Android phones, Samsung Galaxy S8, S20 and S21. In addition, we used a cellular network monitoring app called G-NetTrack. Our measurements include RSRP measures, cell identifiers (IDs), E-UTRA Absolute Radio Frequency Channel Number (EARFCN), and Global Positioning System (GPS) coordinates. Note that EARFCN represents the LTE band and the center frequency of the serving cell. To acquire cell-site information, we used a mobile application called CellMapper [20] and a database known as Antenna Search [21]. CellMapper is a crowd-sourcing application that offers useful cell-specific information, such as cell types, cell IDs, uplink/downlink carrier frequency, and cell addresses. We determined the addresses of the serving cells using the cell IDs from our measurements. However, since CellMapper only provides approximate addresses, we took an additional step using Antenna Search. Specifically, we searched for the approximate addresses obtained from CellMapper on Antenna Search and extracted detailed tower information such as GPS coordinates and tower height.

2) Converting RSRP into Pathloss

In 4G LTE, RSRP is the average power received in the resource elements (RE) that carry cell-specific reference signals (CRS). The RSRP in dBm can be expressed as [19], [22]

$$RSRP = -PL + \Delta \quad (1)$$

where PL is the pathloss in dB and Δ is an offset. The offset Δ represents the cumulative effect of unknown site-

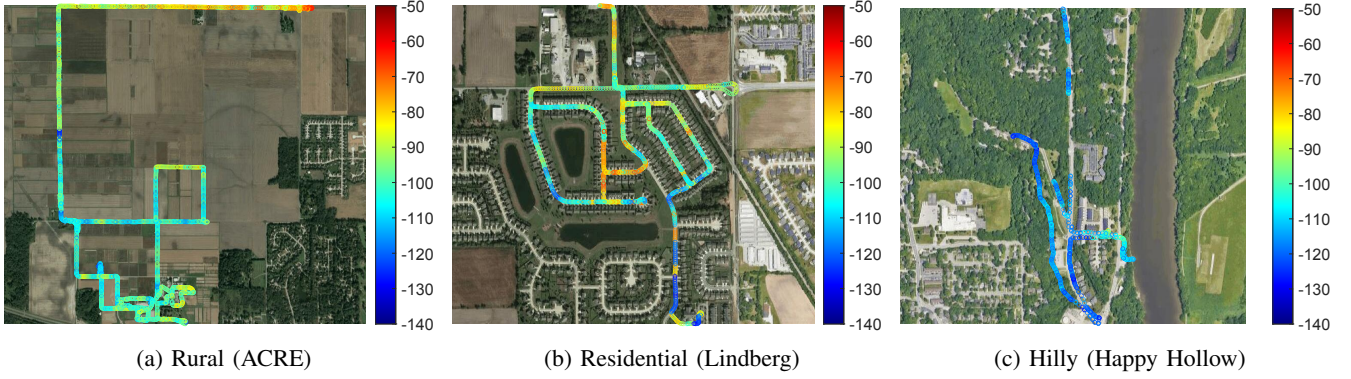


Figure 2: RSRP in dBm derived from the data collected during the measurement campaign.

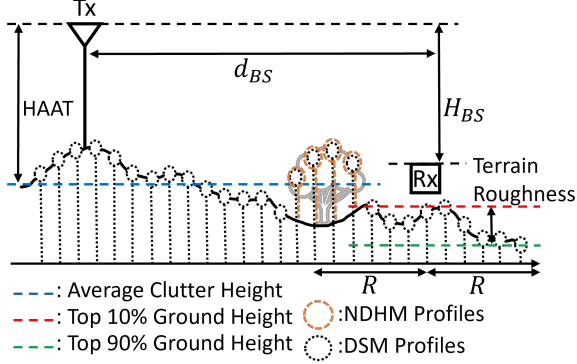


Figure 3: Illustration of the engineered features used as inputs for ML algorithm.

specific parameters, including transmit power, antenna gain, and cable loss. Since these parameters are in general not available, we have to estimate the offset Δ of each site to compute the path loss. To this end, we compared RSRP in the measurements and pathloss in the synthetic data point-to-point. The difference between the ℓ th RSRP measure and its corresponding simulated point can be written as Δ_ℓ . Then, the offset Δ can be obtained by taking the average of the differences as

$$\Delta = \frac{1}{N_s} \sum_{\ell=1}^{N_s} \Delta_\ell, \quad (2)$$

where N_s is the number of RSRP samples. In practice, the averaging process can cause additional errors. Therefore, it would be desirable to derive the exact offsets if site-specific parameter information is available. Once the offsets are calculated, the pathloss can be derived by subtracting the values of RSRP from the offsets, that is, $PL = \Delta - RSRP$.

C. Creating Synthetic Datasets

We utilized a comprehensive cellular coverage simulator [19] to estimate generate pathloss for various scenarios, based on high-accuracy LiDAR data with meter-level resolution. The simulation area for each scenario was designed to cover all relevant measurement points and their surrounding environments. Typical LTE network parameters and equipment specifications were applied (Section II-B2).

As introduced in [19], two sets of simulation results are available for analysis: pathloss based on the eHata model and

cumulative blockage distance based on the 60% clearance test of the first Fresnel zone. The locations of the cell towers were determined by a manual process that involved cross-referencing of data from various publicly accessible tools and sources (Section II-B). All carrier frequencies for each scenario, as observed in the corresponding measurement data set, were included in the simulation.

III. FEATURE EXTRACTION

In this section, we develop and extract the features of the pathloss prediction model. Given the significant impact of feature engineering on ML prediction performance, we carefully define and extract geographical features using geographical datasets.

A. Geographic Dataset

To extract geographic features from the LiDAR datasets, we used Indiana's statewide LiDAR datasets collected in 2018 and Digital Surface Model (DSM) and Normalized Digital Height Model (NDHM) for geographic profiles [23]. The DSM provides the elevation above sea level for each point, representing the vertical height of the tallest objects, including trees and buildings at those points. In contrast, NDHM provides information about the elevation above ground level, which indicates the height of the clutter (e.g., buildings and trees) at each point. The DSM and NDHM have a spatial resolution of 5 feet. With these two datasets, we can determine the elevation of the ground and the height of the surface of each point.

B. Feature Engineering

For radio attributes, we focus on the center frequency, given its significant influence on pathloss [1]. In terms of geographic attributes, we have incorporated both endpoint- and path-based characteristics, including the relative height of the serving cell to each reception point and the elevation angle. Fig. 3 illustrates the radio and geographical features. The description of the features is given below.

- **Carrier Frequency:** Represents the center frequency at which the communication signal is transmitted. This parameter can be extracted directly from the attribute EARFCN in the measurement dataset.
- **Serving base stations (BS) Distance (d_{BS}):** Defined as the distance between a receiving point and its service BS.
- **Relative BS Height (H_{BS}):** The height of the serving BS relative to a receive point.

- **Average Clutter Height (H_C):** Average relative height of clutter of neighboring points. Neighboring points are defined as points within a radius circle R centered on a received point. In this study, we choose to use $R = 50$ m.
- **Terrain Roughness:** This parameter characterizes the roughness of the terrain, which is defined as the difference between the top 10% and 90% ground elevation of the neighboring points.
- **Transmitter Height Above Average Terrain (Tx-HAAT):** The difference between the BS height and the average clutter height. This parameter provides information about the height of BS above the surrounding terrain.
- **Ratio α [12]:** The ratio α is defined as $\frac{H_{BS}-H_C}{d_{BS}}$. This ratio is used to characterize the elevation angle between the serving BS and a receiving point.

IV. PERFORMANCE EVALUATION

For the performance evaluation, collected measurements were used from three different areas. The collected datasets are composed of about 133,800 RSRP measurements, spanning three different scenarios:

- **ACRE** (71,068 samples): Rural farm area
- **Lindberg** (16,107 samples): Residential area.
- **Happy Hollow** (46,641 samples): Hilly suburban area

We created simulations for ACRE, Lindberg, and Happy Hollow based on the cell information acquired from the measurement campaign. We configured user location grids with 6300 points for ACRE, 9200 points for Lindberg, and 5900 points for Happy Hollow to fully cover the area of interest. We simulated each site with varying center frequencies from 731.5 MHz to 2538.2 MHz to account for different propagation characteristics. In the end, we obtained a synthetic dataset of approximately 651,000 data points, which improved our training datasets with different environments.

To evaluate the prediction performance of the ML models, we use mean absolute error (MAE), which is expressed as

$$MAE = \frac{1}{n} \sum_{i=1}^n |L_i - \hat{L}_i| \quad (3)$$

where n is the number of pathloss samples, L_i is the true pathloss, and \hat{L}_i is the predicted pathloss.

A. Comparison with Empirical Radio Propagation Models

We compared our ML-based pathloss prediction with traditional empirical models¹ that use an equation to calculate pathloss at any given location from the base station [24]. Specifically, we used the COST-231 Hata model and the Stanford University Interim (SUI) model as benchmarks [25]. The COST-231 Hata model includes adjustments for different environments, such as urban, suburban, and rural areas. The SUI model refines the predictions by offering correction factors for antenna heights of the base station and user equipment, along with constants that vary depending on the type of terrain. We applied terrain A corrections for the Lindberg and Happy Hollow datasets and terrain C corrections for the ACRE dataset, as specified for the SUI model [25].

¹The source code for this work is available publicly at <https://github.com/aprincemohamed/DeepLearningBasedCellularCoverageMap>.

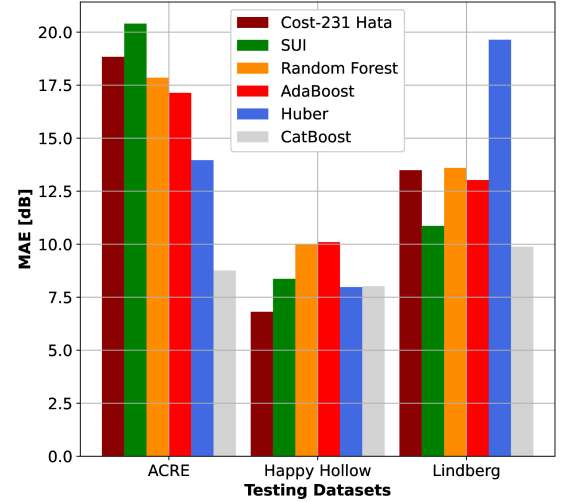


Figure 4: Prediction performance in the same environment, where training data consists of synthetic data structured to mimic the characteristics of the real test data environment.

Fig. 4 compares the prediction performance of the deterministic and ML models at different sites. We evaluated several ML algorithms, including Random Forest, AdaBoost, HuberRegressor, and CatBoost, to identify the one that offers the highest accuracy. The models were trained on a synthetic dataset and then validated against a real-world dataset. Specifically, for the Happy Hollow datasets, excluding the center frequency feature from training and testing improved the model's performance over using all features.

In general, it can be verified that CatBoost outperforms the other ML schemes and empirical models. The MAE achieved for the ACRE, Happy Hollow, and Lindberg data sets was approximately 8.76 dB, 8.02 dB, and 9.88 dB, respectively. It should be noted that synthetic data, coupled with the developed features, contributed significantly to favorable results. This is evidenced by consistently lower MAE values, which significantly undercut those obtained with empirical models. For the Happy Hollow dataset, the COST-231 Hata model achieved an MAE of 6.8 dB, which is marginally better than the 8.02 dB of CatBoost when using synthetic data. However, when CatBoost is trained on real data, specifically with a 50% split, it exhibits superior performance with an MAE reduced to approximately 5.22 dB, outperforming the COST-231 Hata model. We infer that CatBoost's superior performance can be attributed to its sophisticated ordered-boosting technique, which is especially effective in avoiding overfitting. Hence, we chose CatBoost as our main ML scheme for the following experiments.

These comparisons clearly highlight the advantage of using machine learning over empirical models. This approach circumvents the need for costly data collection processes by showing that our models, when trained with cost-effective synthetic data, still maintain high performance in real-world scenarios. This is particularly advantageous because it is achieved without reliance on complex or expensive data types, such as images, highlighting the practicality and efficiency of our methodology.

Training Data	MAE [dB] for Different Testing Datasets	
Scenario 1	ACRE	Happy Hollow
ACRE (R) + Happy Hollow (S)	4.97	7.7
ACRE (R)	4.33	8.7
Happy Hollow (S)	8.6	7.98
Happy Hollow (R) + ACRE (S)	8.74	5.55
Happy Hollow (R)	9.65	5.22
ACRE (S)	8.74	6.59
Scenario 2	Lindberg	ACRE
Lindberg (R) + ACRE (S)	6.32	8.75
Lindberg (R)	5.39	10.49
ACRE (S)	10.14	8.74
Scenario 3	Lindberg	Happy Hollow
Lindberg (R) + Happy Hollow (S)	9.5	7.79
Lindberg (R)	5.39	19.99
Happy Hollow (S)	10.23	7.98

Table I: Pathloss prediction accuracy (in MAE) comparing the proposed data augmentation (highlighted) to baselines. "R" denotes real data and "S" denotes simulated data. For each dataset, training includes all synthetic data and 50% of the real data, the remaining real data being used for evaluation.

B. Generalization Performance Evaluation

In Table I, we assess how effectively our proposed training method generalizes across different scenarios, specifically comparing its performance when using only measurement data versus scenarios that use only synthetic datasets. For the measurement data at each location, we partitioned the dataset into two equal halves, allocating 50% for training purposes and the remaining 50% for testing. Generally, cases that use only measurement data show good prediction accuracy in known areas; however, their performance decreases in unseen areas. In all scenarios, the cases that used only measurement data showed lower performance compared to those that only used simulation data to predict unseen areas. It can be seen that the proposed training method improved the accuracy of the prediction of unseen areas by incorporating synthetic data, although at the cost of some accuracy in known areas. Specifically, the ACRE (R) + Happy Hollow (S) case improved the prediction accuracy of Happy Hollow by 1.0 dB compared to the ACRE (R) case. On the other hand, the accuracy of the ACRE prediction decreased slightly by 0.64 dB. Similarly, the Happy Hollow (R) + ACRE (S) case improved the prediction accuracy of ACRE by 1.09 dB, with a minor loss of the prediction accuracy of 0.33 dB for Happy Hollow. A similar trend can be observed in Scenario 2. The proposed data augmentation improved the prediction accuracy for ACRE by 2.26 dB at the cost of a 1.07 dB accuracy loss for Lindberg.

In Scenario 3, the only real data-based scheme achieves the prediction accuracies of 19.99 dB and 5.38 dB for Happy Hollow and Lindberg, respectively. The large difference in the prediction results for Happy Hollow (hilly) and Lindberg (residential) may suggest that they have significantly different propagation environments. However, when the synthetic data for Happy Hollow was added to the dataset, the prediction accuracy for Happy Hollow was enhanced by approximately 12.2 dB. This improvement came with a trade-off, leading to a loss 4.11 dB in the accuracy of the Lindberg prediction, which is consistent with previous findings.

These results suggest that the incorporation of synthetic data

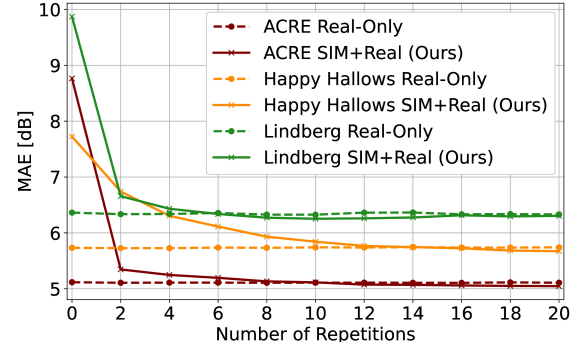


Figure 5: MAE vs 5% Real Data Repetitions in Training Set.

into training improves prediction accuracy in almost all test scenarios, as seen in the lower dB values for the combinations of real and synthetic data compared to real data alone or synthetic data alone. This aligns with our earlier discussion on the potential advantages of leveraging synthetic data to address the generalization challenge posed by data scarcity.

C. Training Optimization with Limited Measurement Data

Next, we evaluate the prediction performance when available measurement data is limited. In real-world applications, the challenge of data scarcity is common. To navigate this, we use a nominal 5% of available measurement data in the training process. To evaluate the impact of simulation-aided data augmentation, we compared our proposed method with a baseline that uses only measurement data. In our approach, for each site, we randomly sampled only 5% of the entire measurement dataset for training and used the remaining 95% for testing. To counterbalance the limited measurement data, we augment it with synthetic data, exposing the model to a broader range of scenarios. The synthetic dataset, which is significantly larger than the real data, risks overfitting the model to its characteristics. To address this issue, we implemented a strategy of repeating the measurement dataset multiple times in each epoch, ensuring a balanced distribution of real and synthetic data during the model training phase.

Fig. 5 shows the prediction accuracy for different scenarios with varying numbers of repetitions. Clearly, our approach, which involves repeated exposure to a small subset of measurement data, results in significant improvements in predictive accuracy. Specifically, for ACRE datasets, employing 20 repetitions of the 5% real data subset, the model achieved an MAE of 5.05 dB, a significant improvement over the MAE of 8.77 dB observed when the model was trained only using synthetic data. Interestingly, the figure also highlights the benefits of our repetitive method compared to using only real data without repetition, which yielded a slightly higher MAE of 5.11 dB using 16 repetitions. Similarly, for the Happy Hollow site, the MAE was reduced to 5.67 dB with 20 repetitions, compared to 7.72 dB using only synthetic data and slightly better than 5.74 dB observed with 18 repetitions of only real data. Furthermore, for the Lindberg dataset, where the model reported an MAE of 6.25 dB with 10 repetitions, significantly outperforming the MAE of the synthetic data of 9.87 dB, and also showing a slight advantage over the MAE of 6.32 dB achieved with 10 repetitions of only real data.

By addressing data imbalance, we improved the prediction accuracy of our model, even with limited measurements in training sets. This underscores the efficacy of our simulation-aided data augmentation, which proves its potential as a solution to measurement scarcity in real-world applications. Our approach enhances performance beyond what is achieved with either synthetic or real data alone. It showcases the advantages of a balanced mix of both, thus optimizing predictive accuracy across various datasets.

V. CONCLUSION

This paper addressed a simulation-enhanced data augmentation method to overcome the data shortage problem in the prediction of pathloss via ML. We collected data for three different environments using commodity mobile phones and augmented the dataset using synthetic data to enhance diversity. For the ML prediction model, we engineered and extracted geographic features from LiDAR datasets. We conducted a rigorous performance evaluation to show the effectiveness of the proposed method. The experimental results demonstrated that the proposed data augmentation method effectively enhances the generalization performance of the prediction model. In summary, our research offers a practical and effective strategy for the prediction of pathloss, highlighting the potential of simulation-aided ML to mitigate data scarcity challenges and improve the efficiency of network deployment.

REFERENCES

- [1] C. Phillips, D. Sicker, and D. Grunwald, "A Survey of Wireless Path Loss Prediction and Coverage Mapping Methods," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 255–270, 2013.
- [2] H. Friis, "A Note on a Simple Transmission Formula," *Proc. of the IRE*, vol. 34, no. 5, pp. 254–256, May 1946.
- [3] A. G. Longley and P. L. Rice, "Prediction of tropospheric radio transmission loss over irregular terrain. A computer method-1968," U.S. Department of Commerce, Environmental Sciences Service Administration, Institute for Telecommunication Sciences, Tech. Rep. ERL 79-ITS-67, 1968, <https://its.ntia.gov/publications/2784.aspx>.
- [4] E. F. Drocella, J. Richards, R. Sole, F. Najmy, A. Lundy, and P. McKenna, "3.5 GHz exclusion zone analyses and methodology," U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences,

- Tech. Rep. TR-15-517, 2016, <https://its.ntia.gov/publications/details.aspx?pub=2805>.
- [5] "COST Action 231 - Publications Office of the EU," <https://op.europa.eu/en/publication-detail/-/publication/f2f42003-4028-4496-af95-beaa38fd475f>.
- [6] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical report (TR) 38.901, 01 2020, version 16.1.0.
- [7] S. Jaeckel, L. Raschkowski, K. Borner, and L. Thiele, "QuaDRiGa: A 3-D Multi-Cell Channel Model With Time Evolution for Enabling Virtual Field Trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.
- [8] M. Lawton and J. McGeehan, "The application of a deterministic ray launching algorithm for the prediction of radio channel characteristics in small-cell environments," *IEEE Trans. Veh. Technol.*, vol. 43, no. 4, pp. 955–969, Nov./1994.
- [9] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong, and T. Kürner, "The Design and Applications of High-Performance Ray-Tracing Simulation Platform for 5G and Beyond Wireless Communications: A Tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 10–27, 2019.
- [10] V. V. Ratnam *et al.*, "FadeNet: Deep Learning-Based mm-Wave Large-Scale Channel Fading Prediction and its Applications," *IEEE Access*, vol. 9, pp. 3278–3290, 2021.
- [11] A. Vanleer and C. R. Anderson, "Improving Propagation Model Predictions via Machine Learning with Engineered Features," in *2021 IEEE Military Commun. Conf. (MILCOM)*, Jan. 2021, pp. 420–425.
- [12] G. Reus-Muns, J. Du, D. Chizhik, R. Valenzuela, and K. R. Chowdhury, "Machine Learning-based mmWave Path Loss Prediction for Urban/Suburban Macro Sites," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, Feb. 2022, pp. 1429–1434.
- [13] A. Gupta, J. Du, D. Chizhik, R. A. Valenzuela, and M. Sellathurai, "Machine Learning-Based Urban Canyon Path Loss Prediction Using 28 GHz Manhattan Measurements," *IEEE Trans. Antennas Propag.*, vol. 70, no. 6, pp. 4096–4111, Jun. 2022.
- [14] X. Zhang, X. Shu, B. Zhang, J. Ren, L. Zhou, and X. Chen, "Cellular Network Radio Propagation Modeling with Deep Convolutional Neural Networks," in *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM, Aug. 2020, pp. 2378–2386.
- [15] Y. Zhang, D. J. Love, J. V. Krogmeier, C. R. Anderson, R. W. Heath, and D. R. Buckmaster, "Challenges and Opportunities of Future Rural Wireless Communications," *IEEE Commun. Mag.*, vol. 59, no. 12, pp. 16–22, Dec. 2021.
- [16] C. Tang, S. Vishwakarma, W. Li, R. Adve, S. Julier, and K. Chetty, "Augmenting experimental data with simulations to improve activity classification in healthcare monitoring," in *2021 IEEE radar conf. (RadarConf21)*. IEEE, 2021, pp. 1–6.
- [17] A. P. Mohamed, A. S. M. M. Jameel, and A. El Gamal, "Knowledge distillation for wireless edge learning," in *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2021, pp. 600–604.
- [18] Y. Zhang, J. A. Tan, B. M. Dorbert, C. R. Anderson, and J. V. Krogmeier, "Simulation-Aided Measurement-Based Channel Modeling for Propagation at 28 GHz in a Coniferous Forest," in *2020 IEEE Global Commun. Conf. (GLOBECOM)*. Taipei, Taiwan: IEEE, Dec. 2020, pp. 1–6.
- [19] Y. Zhang, J. V. Krogmeier, C. R. Anderson, and D. J. Love, "Large-Scale Cellular Coverage Simulation and Analyses for Follow-Me UAV Data Relay," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.
- [20] CellMapper, "AT&T Mobility (United States of America) - Cellular Coverage and Tower Map," <https://www.cellmapper.net/map?MCC=-1&MNC=-1>, <https://www.cellmapper.net/map>.
- [21] "AntennaSearch - Search for Cell Towers & Antennas," <https://www.antennasearch.com/>.
- [22] S. J. Maeng, H. Kwon, O. Ozdemir, and I. Güvenç, "Impact of 3D Antenna Radiation Pattern in UAV Air-to-Ground Path Loss Modeling and RSRP-based Localization in Rural Area," *IEEE Open J. Antennas and Propag.*, pp. 1–1, 2023.
- [23] J. Jung and S. Oh, "Indiana statewide digital surface model (2016-2019)," Feb. 2021.
- [24] U. Masood, H. Farooq, and A. Imran, "A machine learning based 3d propagation model for intelligent future cellular networks," in *2019 IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2019, pp. 1–6.
- [25] V. Abhayawardhana, I. Wassell, D. Crosby, M. Sellars, and M. Brown, "Comparison of empirical propagation path loss models for fixed wireless access systems," in *2005 IEEE 61st Veh. Technol. Conf.*, vol. 1. IEEE, 2005, pp. 73–77.