Refining Wireless Propagation Models using Domain-Informed GANs amid Data Scarcity

Waseem Raza*, Syed Basit Ali Zaidi‡, Muhammad Umar Bin Farooq*, Haneya Naeem Qureshi* and Ali Imran*‡
‡James Watt School of Engineering, University of Glasgow, United Kingdom.
*AI4Networks Research Center, School of Electrical & Computer Engineering, University of Oklahoma, OK, USA
Email: {waseem, umar.farooq, haneya, ali.imran}@ou.edu, s.zaidi.2@research.gla.ac.uk

***Abstract*—Data-driven Machine Learning (ML) based propagation models are essential for modern wireless network planning and optimization. However, their effectiveness is limited by scarse data conditions. Generative Adversarial Networks (GANs) often considered as a viable approach for data augmentation, struggle in these conditions because they also require large datasets for effective training. To address this challenge, we propose a novel approach that incorporates domain knowledge directly into GAN training. Using an analytical propagation equation based on 3GPP recommendations, we generate pseudo-random data to train a neural network, which then initializes the GAN generator network. This initialization improves the GAN's learning ability in extreme data scarcity. The framework enhances data generation quality by up to 52% and machine learning applicability by 60%, providing a robust solution to the scarse data problem in wireless network modeling with demonstrating the potential of integrating domain knowledge within ML methodologies.**

***Index Terms*—Propagation Modeling, Generative Adversarial Networks, Data Sparsity, Domain Knowledge, Weight Transfer.**

I. Introduction

Propagation modeling is crucial for the design and optimization of wireless networks, predicting signal strength, coverage, interference, and system reliability [1]. Models are typically categorized into empirical, deterministic, and stochastic types [2]. Empirical models use real-world data and statistical analysis, deterministic models apply physical laws and environmental geometry, and stochastic models rely on random processes and probabilistic channel parameters. Emerging data-driven and AI-based approaches offer promising enhancements to traditional methods by utilizing extensive datasets and advanced algorithms like machine learning and deep neural networks to capture complex wireless propagation patterns [2]. These methods train models, such as neural networks, to understand the intricate relationships between input variables (e.g., distance, frequency, antenna height) and outputs (e.g., signal strength or path loss) using data from actual network operations [3].

However, data-driven models face significant challenges, including high computational demands for real-time predictions and constrained data availability due to privacy issues [4]. A primary concern is data scarcity from resource-intensive signal measurement campaigns, leading to datasets that are limited or geographically restricted [5]. This scarcity impairs model generalization, particularly in diverse or evolving environments. Although data augmentation techniques like Generative Adversarial Networks (GANs) exist, they require extensive datasets, which exacerbates the data scarcity issue and risks mode collapse, where models reproduce only a narrow portion of the data spectrum [5], [6].

A. Related Work

The investigation of ML in the domain of wireless network propagation modeling and pathloss prediction spans various environments and methods [7]–[10]. Specifically, [7] confirms the effectiveness of DNNs for pathloss prediction in macrocells across diverse terrains, while [8] enhances ANN design using a composite differential evolution algorithm, which improves prediction precision. The incorporation of environmental variables into pathloss models via machine learning and DNNs is elaborated in [9], underlining their significance in heterogeneous networks. Additionally, [10] focuses on precise network coverage predictions and the interpretability of models, essential for their practical deployment. Progress in 3D propagation modeling is explored by [1], addressing intricate spatial dynamics and merging AI interpretability with detailed 3D modeling to develop advanced tools for autonomous network design.

Despite these advancements, ML-based propagation models face significant challenges, notably the scarcity of varied, large-scale datasets which skews model performance towards urban settings and diminishes effectiveness in rural or varied terrains [11]–[13]. To combat this, recent studies have utilized GANs for data generation and augmentation. For instance, [6] introduces a dual-phase learning framework with conditional GANs that enhance radio map estimation accuracy, particularly in under-documented outdoor environments. Similarly, [14] employs GANs to generate detailed path loss maps, treating path loss prediction as an image synthesis problem, which effectively manages complex urban layouts and diverse terrains. These approaches leverage GANs' ability to replicate spatial dependencies in images, demonstrating their utility in creating accurate radio maps from varied data sources.

Data-driven machine learning-based propagation modeling greatly benefits from incorporating diverse network and environmental parameters in tabular data formats, which, unlike image data, include both discrete and continuous variables without spatial correlations. This complexity and the typically scarce nature of tabular datasets in real scenarios pose significant challenges as they require extracting meaningful patterns from limited data points without image-like contextual

cues. To effectively train a typical GAN and achieve high-quality data generation, an ample supply of training data is crucial. However, to produce this abundance of data through synthetic generation, a well-trained GAN is required. This cyclical relationship creates a paradox where the scarcity of training data impedes the development of a robust GAN, while a lack of a robust GAN hinders the ability to generate synthetic data that could alleviate the data scarcity.

To overcome these obstacles, innovative adaptations of GAN architectures or alternative ML strategies are necessary to effectively utilize scarce tabular data for detailed propagation modeling. *To the best of the authors' knowledge, no existing work focusing on enhancing tabular data-based propagation modeling performance through domain informed GAN-based data augmentation.* Therefore, this work aims to address this gap as summarized in the following.

B. Contribution Summarized

The summary of the proposed work is discussed below.

- We develop a GAN-based synthetic data generation and augmentation framework to enhance data-driven propagation modeling performance in extremely scarce datasets. This framework incorporates the domain knowledge into GANs by leveraging the pseudo-random data generated from an analytical propagation modeling equation derived according to the 3GPP recommendations.
- After performing some sanity checks on the pseudo-random data, in the first stage, a DNN model is trained to learn the relationship between received signal strength, and crucial channel and antenna parameters such as distance, Base Station (BS) height, 3D antenna angles. Then, weights of the trained DNN model are transferred to GAN generator network, and serve as the initial weights in the GAN training on real scarce data in the second stage of proposed framework. The propose framework also includes the GAN synthetic data quality based validation, which directs the learning of the generator and discriminator networks to improve the quality of generated data.
- The proposed framework is evaluated for the quality of generated data measured as column shape and column pair metrics, and ML applicability of augmented data measured as the prediction performance of ML models, for different combinations of training and synthetic data sizes. It is demonstrated that, when compared with baseline schemes, i.e., Gaussian Copula, Conditional Tabular GAN (CT-GAN), the proposed approach exhibits up to 12.2%, and 52% improvement for GAN quality metrics, respectively. Similarly, the Root Mean Square Error (RMSE) and Adjusted R2 prediction performance on augmented data demonstrate up to 19.5% and 60.8% improvement from the baseline schemes.

Rest of the paper is organized as following. In section II, we discuss the formulation of domain knowledge based analytical equation, working of the proposed domain informed weight transfer GANs, and synthetic data generation, evaluation, augmentation, and ML based propagation model training modules of the proposed framework. The simulation setup and performance evaluation in terms of GAN data generation quality, ML applicability are discussed in section III.

II. Domain Informed GANs Framework

In this section, we discuss the formulation of domain knowledge based analytical equation, working of the proposed domain informed weight transfer GANs, and synthetic data generation, evaluation, augmentation, and ML based propagation model training modules of the proposed framework.

A. Domain knowledge-based analytical equation formulation

In order to utilize the domain knowledge in ML model training, we start by identifying the relevant analytical equation which should have significant critical parameters that accurately represent the physical and geometric attributes of the signal propagation environment. Also, these parameters of analytical equation should be present in the dataset typically used to train ML models for data driven propagation models. Although, the choice of these parameters depends on various factors, such as the considerations about the underlying environment, and operating frequency, our domain knowledge of the wireless propagation field helps to identify these features, which should include the distance between User Equipment (UE) and its serving base station (eNodeB), transmit power, horizontal and vertical antenna angles etc. The Reference Signal Received Power (RSRP) which serves as the received signal strength metric at a specific location is given by:

$$P_r[\mathrm{dBm}] = P_t - PL + G - L + X, \tag{1}$$

where P_t is transmitted power by the eNodeB, PL is path loss, G is antenna gain, L indicates attenuation from obstacles, and X covers additional losses. As per 3GPP TR 36.873 [15], the path loss for LOS is given as,

$$PL_{LOS}(d, f_c) = 22.0\log_{10}(d) + 28.0 + 20\log_{10}(f_c), \tag{2}$$

and for the NLOS it is given as,

$$\begin{aligned} &PL_{NLOS}(f_c, W, h, h_{bs}, h_{ue}, d) = 20\log_{10}(f_c) \\ &-7.1\log_{10}(W) + 14.9\log_{10}(h) - 18.76\log_{10}(h_{bs}) - 0.6(h_{ue}) \\ &\quad - 3.1\log_{10}(d)\log_{10}(h_{bs}) + 43.42\log_{10}(d) + 31.69, \end{aligned} \tag{3}$$

where h_{bs}, and h_{ue}, are the heights of BS and UE antennas, W is the street width, d is the distance between UE and BS, and f_c is the carrier frequency. Antenna gain G is defined as the product of the maximum antenna gain G_{max} and the antenna attenuation A_{att}, expressed as $G_{\max}(B_h, B_v, \zeta) = \zeta D = \zeta\frac{4\pi}{B_h B_v}$. The value of A_{att} is estimated following 3GPP guidelines: $A_{att} = \lambda_v \min[12(\frac{\phi_u - \phi_{tilt}}{B_v})^2, A_v] + \lambda_h \min[12(\frac{\theta_u - \theta_{azi}}{B_h})^2, A_h]$. This comprehensive analytical formula for RSRP calculation incorporates various aspects of signal propagation. B_h and B_v represent the horizontal and vertical half-power beam widths, θ_u, θ_{azi} are the azimuth angles, whereas, ϕ_u, ϕ_{tilt}, are the tilt angles for BS and UE. λ_h and λ_v represent the weighting factors for the beam pattern in both directions. The final equation (4) is obtained by combining these components as shown on top of page 3.

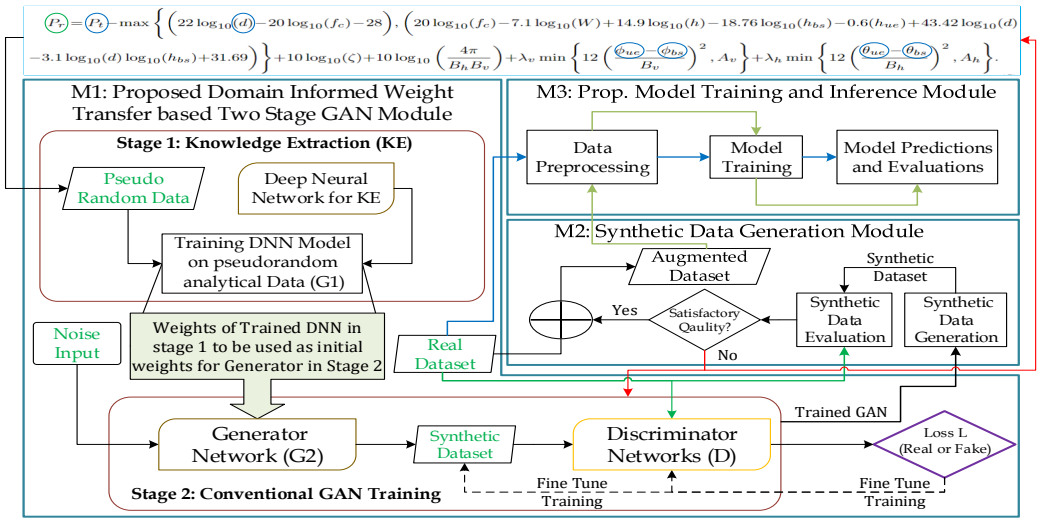

Fig. 1: The framework to explain the working of proposed Weight Transfer based domain informed GAN (WT-GAN) approach.

$$P_r(d, f_c, W, h, h_{ue}, h_{bs}, B_h, B_v, \phi_{ue}, \phi_{bs}, \theta_{ue}, \theta_{bs}, \lambda_h, \lambda_v, A_h, A_v) = P_t - \max[(22\log_{10}(d) - 20\log_{10}(f_c) - 28), (20\log_{10}(f_c) - 7.1\log_{10}(W) + 14.9\log_{10}(h) - 18.76\log_{10}(h_{bs}) - 0.6(h_{ue}) - 3.1\log_{10}(d)\log_{10}(h_{bs}) + 43.42\log_{10}(d) + 31.69)] + 10\log_{10}(\zeta) + 10\log_{10}(\frac{4\pi}{B_h B_v}) + \lambda_v \min[12(\frac{\phi_{ue} - \phi_{bs}}{B_v})^2, A_v] + \lambda_h \min[12(\frac{\theta_{ue} - \theta_{bs}}{B_h})^2, A_h]. \quad (4)$$

B. Proposed Domain Informed Weight Transfer GAN (M1)

In this section, we explain the working of proposed weight transfer based domain informed GANs depicted in Fig. 1. The key idea behind this approach is to utilize the analytical equation in (4) to generate pseudo-random data for different configuration parameters such as the transmit power, distance, and antenna angles. Then, we perform the exploratory data analysis, and some sanity checks, such as the value ranges and trends of different variables of this data with respect to individual parameters, to ensure that this data is truly capturing the propagation modeling relations between parameters and target variable. Then, in the first stage of proposed approach, we train a DNN model to learn the relationships in pseudo-random data from analytical equation. We term this model as G1, and the weights of the trained G1 are then utilized as the initial weights for the *Generator Network G2* in the second stage. This approach provides a mechanism to incorporate the domain knowledge into ML based propagation modeling, specifically to the generator of GANs by learning the knowledge of complex relationships and inter-dependencies between these parameters and target RSRP.

The second stage of the proposed approach is similar to conventional GAN which involves, two neural networks - a generator and a discriminator - that are trained simultaneously through a competitive process. However, in each training iteration, the generator is initiated by the weights from the trained model from stage 1. The generator creates synthetic data samples, while the discriminator evaluates them and provides feedback to the generator. This feedback loop continues until the generator produces data that is indistinguishable from real data on which GAN training is performed. It should be noted that the real dataset used for GAN training is different from the pseudo-random data generated from the equation in the sense that the former is obtained either from real networks or by simulating the realistic propagation modeling environment in a planning tool. However, it should have some common features with the parameters of pseudo random data to enable the effective learning and transfer of knowledge.

We also improve the GAN training process by incorporating the real time data generation quality based GAN validation into the training process. Since, the loss values of generators and discriminators exhibit fluctuations, it is difficult to track the learning or validations through their losses. Hence, we utilize the GAN data generation quality in each iteration as the validation metric, which not only guide the learning process but also serves as the metric for early stopping criteria to achieve the convergence in GAN training.

C. Data Preparation and Training of GANs

Now we discuss about obtaining the realistic data for GAN training following the system model and feature engineering.

1) System Model and Feature Engineering: We simulate a 3GPP compliant 5G network system in Atoll [16], across a 3 sectored multi-macrocell simulation area in Brussels [1]. Leveraging ray-tracing, it allows to accurately model the network topology, and incorporates the realistic antenna designs, terrain elevation data, urban infrastructure, and the Aster propagation model to which simulates phenomena like signal diffraction, reflections, and environmental attenuations. This sophisticated modeling ensures the raw data generated reflects realistic propagation scenarios. However, despite being high accurate, the Atoll is computationally inefficient to generate large datasets for propagation modeling, hence, its data is augmented by GAN generated synthetic data for better modeling.

The collected raw data has three important components: BS data, geographic information, and UE measurements. BS data encompasses antenna positioning, orientation etc., while geographic data offers insights into terrain elevation, building heights, and land usage. UE measurements include received signal strength and location data. To maximize the effectiveness of ML models, feature engineering is applied to convert raw data to right data. This involves transforming the raw data into meaningful features that capture propagation characteristics like distance, diffraction, and angular separation between BS and UE. These features involve distances (generally indoor and outdoor paths taken separately), clutter information, number of penetrations in the building, diffraction points, and angular separations of BS and UE. Readers are encouraged to Sections II-C of [1] for further details on the feature engineering methodology, and from these engineered features we have resorted to 9 highly impact features for our analysis, shown in top of Table I

2) GAN Training and Validation: We utilize the real dataset depicted in proposed framework in Fig. 1 for the training and real-time validation (dark green lines in Fig. 1) of proposed WT-GAN approach. To achieve the later, we utilize the trained GAN to generate synthetic data after each training iteration of GAN and compare it with validation data from real dataset. This validation metric is formed as the average of similarity and correlation metrics, i.e., column shape, and column pair trends discussed with details in III-B, from [17]. Along with weight transfer, this validation approach assists the GAN models to avoid being stuck in local optima, and be trained to generate the best quality synthetic data. Although, it is previously discussed and will be clear in result discussion that we consider extremely scarce data set for GAN training, however, we set aside an ample amount of data from real dataset for real-time GAN validation, post augmentation data evaluation, and model testing in inference phase.

D. Synthetic Data Generation Module (M2)

After completing the training of proposed WT-GAN, it is employed to generate synthetic data of required number of instances in M2 module. To further ensure the high quality of synthetic data, we employ an evaluation block where quality of synthetic data is evaluated by comparing it with real data set for column shape and column pair trends metrics. This is followed by a satisfactory quality check, an arbitrarily selected threshold using domain knowledge, to only allow the high quality synthetic data for data augmentation, otherwise requiring to redo the whole process (red lines in Fig. 1) of analytical equation formulation, stage 1 and stage 2 of proposed WT-GAN approach.

E. Model Training and Inference Modules (M3)

In this module, we aim to check the ML applicability of proposed approach, by training the DNN based regression model on augmented data for propagation modeling. Hence, this process involves the blocks for relevant preprocessing and model training, and model predictions & evaluation, and augmented data is passed through these blocks as shown by light green line in Fig. 1. Also, for the purpose of calibration and evaluation these processes are separately followed for real dataset as well (blue line in Fig. 1). Hence, model prediction performance for the real non augmented dataset serve as the benchmark for the augmented case. In the following section, we discuss the simulation setup and carryout the performance evaluation of the proposed framework.

Table I: Parameters for simulation setup and analytical propagation modeling.

Parameter Description	Symbol
Transmit Power (P_t)	$[10,\ 43]$ dBW
Distance (3D) (d_{3D})	$[10, 1000]$ m
Frequency and Bandwidth (f_c, W)	2.4 GHz, 20 MHz
BS Tilt Angle ($\phi_{\text{bs\_tilt}}$)	$[0,\ 6]^o$
UE Tilt Angle ($\phi_{\text{ue\_tilt}}$)	$[-86,\ -9]^o$
BS Azimuth Angle ($\theta_{\text{bs\_azim}}$)	$[40,\ -320]^o$
UE Azimuth Angle ($\theta_{\text{ue\_azim}}$)	$[0,\ -360]^o$
Horiz. and Vert. Antenna Gain (A_h, A_v)	30 dB
Horiz. and Vert. Beamwidth (B_h, B_v)	65^o
UE and BS Heights ($h_{\text{ue}}, h_{\text{bs}}$)	1m, 30 m
Path Loss Exponent (X)	50
Shadowing Factor (ζ)	0.7
DNN and GAN Models Architecture	5 by 50

III. Simulation Setup and Performance Evaluation

A. Experimental Setup

The experimental setup employs Atoll software to simulate a network environment that encapsulates a geographic area of $1000m \times 1000m$, served by 10 macrocells. The parameters governing the simulation are detailed in Table I. This area is segmented into discrete bins, within each of which the RSRP values are calculated for multiple users. The RSRP value for each bin is determined by averaging the RSRP values among all users located within that bin. A total of 1000 users are distributed across the simulation area following a Poisson distribution. The dataset encompasses 5000 instances, subsets of different sizes for GAN training, and for the purpose of testing, which includes GAN data generation quality assessment, and ML applicability performance evaluation of baseline and proposed schemes, about 2500 instances are kept aside.

B. Evaluation of GAN Data Generation Capability

In this section we focus on evaluating the data generation capability of proposed and base line schemes. However, before delving into discussion of the actual numbers and trends, we discuss the baseline schemes, and evaluation metrics in the following description. To evaluate the data generation capabilities of our proposed scheme, we compare it with two baseline methods: *Gaussian Copula* and *CT-GAN* from SDV [18], for two performance evaluation metrics: Column Shape, and Column Pair Trends from SDMetrics [17].

The *Gaussian Copula* is a statistical technique synthesizes data by first capturing the probability distributions of individual columns using an inverse cumulative distribution function transformation. It then learns the correlations between variables to construct a copula model—a multivariate distribution that encapsulates these relationships. The method generates synthetic data by sampling from this model, maintaining the

correlation structure of the original dataset. *CT-GAN* operates within a GAN framework, tailored for tabular data. It manages non-Gaussian distributions and imbalances in discrete columns effectively. Continuous variables are modeled using a variational Gaussian mixture model to accurately represent various distribution modes, and it includes a conditional generator for synthetic data based on discrete variable values. The "Column Shape" is measured as Kolmogorov-Smirnov complement based similarity between various columns of original and synthetic data, and average over all column is taken as a single metric. The "Column Pair Trends" metric is measured as Pearson correlation similarity and assesses the preservation of correlations, trends, and dependencies, found between pair of features in the original dataset. Ensuring accurate reflection of inter-feature relationships in synthetic data is crucial for predictive modeling of applications involving these dynamics.

Fig. 2 provides a comparative analysis of the synthetic data generation for the baseline and proposed methods, against hierarchical combinations of training and synthetic data sizes as shown on x-axis. This structured visualization offers a clear comparative perspective, allowing for in-depth analysis of each model's performance across different data. Our analysis of Column Shape metric in top subfigure indicates that the proposed approach outperforms the baseline schemes in all combinations of training and synthetic data sizes. Specifically, it shows an average improvement of 12.2% and 9.45% compared with CT-GAN and Gaussian Copula schemes, respectively. Further the comparison with respect to training and synthetic data combinations depicts that that larger training datasets exhibit noticeable improvement in this metric, by providing a more robust learning environment. Specifically, we get the best improvement of 1.1% with WT-GAN, and 0.69% with CT-GAN, when training data varies from 1K to 2K and synthetic data is 1K. Similarly, the performance improvement with Gaussian Copula is more prominent with 2.1% increase against training data increase from 1K to 2K, and synthetic data size of 2K. This finding supports the notion that more real-world data equips models to better replicate statistical nuances. However, increasing the size of synthetic data tends to degrade performance, suggesting a trade-off between data volume and model effectiveness. In this comparison, most significant fall of -2.53% is also observed with Gaussian Copula when synthetic data increases from 1K to 2K, and training data size of 1K. Hence, it can be concluded that a larger training dataset enhances a models capacity to discern complex patterns, while an increase in synthetic data exerts more strain on the models, potentially stretching their capability to maintain quality.

We compare the "Column Pair Trends" metric, in lower subfigure of Fig. 2. The proposed WT-GAN approach again outperforms the CT-GAN and Gaussian Copula schemes for all combinations of training and synthetic data sizes. Averaging the performance improvements over all these combination, WT-GAN exhibits about 16.4% improvement against CTGAN and 52.3% improvement against Gaussian Copula. Similar to previous metric, we also observe improved performance

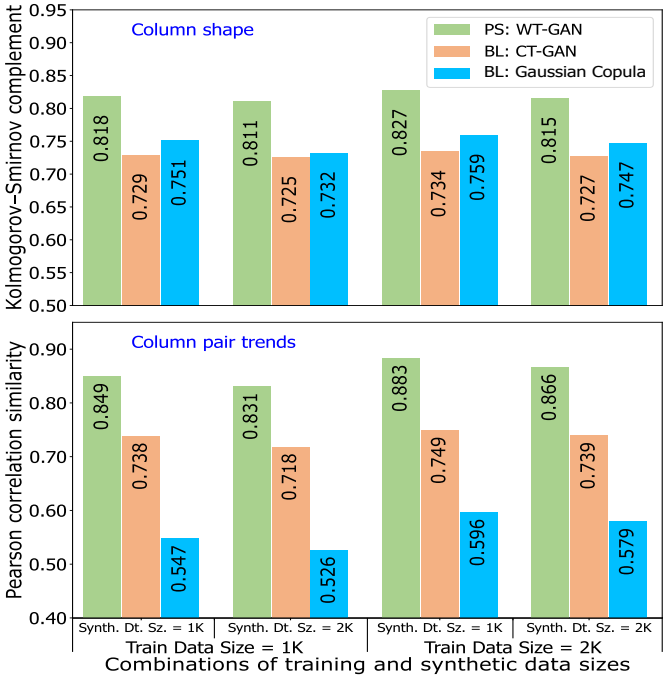

Fig. 2: Comparing the (a) *Column Shape* and (b) *Column Pair Trends* metrics for three comparing schemes (2 baseline and 1 proposed) against different combination of training and synthetic data sizes.

with increased training data, and performance decrease with increased synthetic data. However, one notable aspect is the significantly low values of Gaussian Copula indicating that this scheme fails to effectively capture the correlations between various columns/features of the dataset.

C. Evaluation in terms of Machine Learning Applicability

Synthetic data generation and augmentation is crucial in scenarios with limited data, but its effectiveness is contingent on its relevance and performance in machine learning tasks. Hence, we evaluate GAN-based data augmentation techniques by training a DNN regression model across three augmented data sizes and evaluating the results for two regression metrics, RMSE and Adjusted R2 score in Fig. 3. RMSE is a standard metric for regression models that measures the average magnitude of the errors between predicted and observed values. R2 is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variable. Adjusted R2 is a modified version of R2 that considers the number of predictors (independent variables) in the model.

For this comparison, we evaluate the performance of 'Original' non-augmented data (1K instances) and two augmented data types: 'Baseline: Augmented' using CT-GAN technique, and 'Proposed: Augmented' using WT-GAN approach. It is observed that for both metrics, GAN based data augmentation results in improved performance than the non augmented case shown by horizontal lines in Fig. 2. Specifically, for augmented data size of $2K$ we observe 8.8% improvement with baseline and 10.8% improvement with proposed approach, which respectively extends to 40.42% and 52% for these schemes for data size of $4K$. Also, in both of these metrics the proposed WT-GAN scheme outperforms the baseline schemes

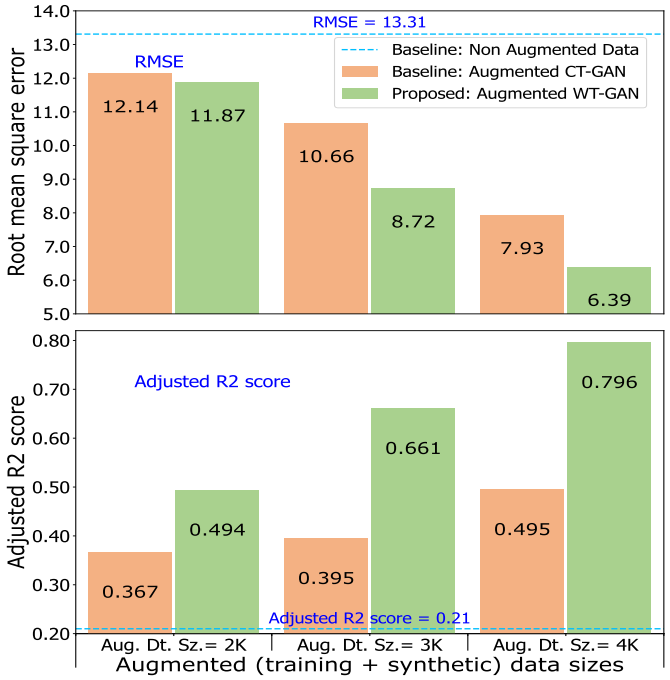

Fig. 3: Comparing the (a) RMSE and (b) Adjusted R2 scores for three comparing schemes (2 baseline and 1 proposed) against different combination of training and synthetic data sizes.

for all data augmentation cases, i.e., showing 2.2%, 18.2%, and 19.5% lower RMSE than baseline scheme for $2K$, $3K$, and $4K$ augmented data sizes, respectively. This comparison further validates the importance of data augmentation by the fact that with data size increase of $2K$ to $4K$, the RMSE performance improvement of proposed scheme gets better.

Similar to RMSE comparison, we observe the improvement of both augmented schemes with non-augmented counterpart with a significant margin for Adjusted R^2 scores metric, which keeps on increasing with the increase in data sizes as shown in bottom in Fig. 2. Also, the comparison between both augmented schemes, baseline and proposed, shows that proposed approach outperforms the baseline for all data sizes, showing the improvement of 34.9%, 67.6%, and 60.8%, for $2K$, $3K$, and $4K$ data sizes, respectively. These improvements highlight and validate the machine learning applicability of synthetic data generated by GAN based data augmentation.

IV. Conclusion and Future Work

This research offers a comprehensive solution to data scarcity in developing data-driven propagation models for wireless networks. Leveraging pseudo-random data from an analytical equation, we integrate domain knowledge into GAN training through DNN model weights. This approach enables high-quality data generation even with limited training datasets of 1K samples. Our results show significant improvements in GAN data generation quality and ML applicability, confirming the efficacy of this method under stringent data sparsity. This work highlights the potential of domain knowledge to overcome data scarcity, marking a major advancement in the field. In the future, we plan to expand this analysis by considering a broader range of parameters in the analytical equation, varying pseudo-random data sizes, and experimenting with different model architectures in both stages of the proposed framework.

V. Acknowledgement

This work is supported by the National Science Foundation through Grant Number 1923669.

References

[1] U. Masood, H. Farooq, A. Imran, and A. Abu-Dayya, "Interpretable AI-Based Large-Scale 3D Pathloss Prediction Model for Enabling Emerging Self-Driving Networks," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 3967–3984, 2023.

[2] 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, p. 2378–2386.

[3] Y. Liu and S. Mao, "DeepWiERL: Deep Reinforcement Learning Based Wireless Propagation Estimation," in *2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, p. 1–6.

[4] F. Lara, R. Lara-Cueva, M. Castillo, J. F. Arellano, and L. Top'on, "Modeling Wireless Propagation Channel: A Traditional Versus Machine Learning Approach," in *Emerging Research in Intelligent Systems*. Springer, 2022, p. 281–297.

[5] Y. Wei, M.-M. Zhao, and M.-J. Zhao, "Channel Distribution Learning: Model-Driven GAN-Based Channel Modeling for IRS-Aided Wireless Communication," *IEEE Transactions on Communications*, vol. 70, no. 7, pp. 4482–4497, 2022.

[6] S. Zhang, A. Wijesinghe, and Z. Ding, "RME-GAN: A Learning Framework for Radio Map Estimation Based on Conditional Generative Adversarial Network," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 18 016–18 027, 2023.

[7] M. Ribero, R. W. Heath, H. Vikalo, D. Chizhik, and R. A. Valenzuela, "Deep Learning Propagation Models over Irregular Rerrain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4519–4523.

[8] S. P. Sotiroudis, S. K. Goudos, K. A. Gotsis, K. Siakavara, and J. N. Sahalos, "Application of a Composite Differential Evolution Algorithm in Optimal Neural Network Design for Propagation Path-Loss Prediction in Mobile Communication Systems," *IEEE Antennas and Wireless Propagation Letters*, vol. 12, pp. 364–367, 2013.

[9] S. I. Popoola, E. Adetiba, A. A. Atayero, N. Faruk, and C. T. Calafate, "Optimal Model for Path Loss Predictions using Feed-Forward Neural Networks," *Cogent Engineering*, vol. 5, no. 1, p. 1444345, 2018.

[10] A. Ghasemi, "Data-driven prediction of cellular networks coverage: An interpretable machine-learning model," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2018, pp. 604–608.

[11] V. P. Rekkas, S. Sotiroudis, P. Sarigiannidis, S. Wan, G. K. Karagiannidis, and S. K. Goudos, "Machine Learning in Beyond 5G/6G Networks—State-of-the-Art and Future Trends," *Electronics*, vol. 10, p. 2786, 2021.

[12] M. Chen, Y.-C. Liang, and H.-H. Chen, "Machine Learning for Wireless Networks with Artificial Intelligence: A Comprehensive Survey," *IEEE Access*, vol. 6, pp. 36 114–36 134, 2018.

[13] H. N. Qureshi, U. Masood, M. Manalastas, S. M. A. Zaidi, H. Farooq, J. Forgeat, M. Bouton, S. Bothe, P. Karlsson, A. Rizwan *et al.*, "Towards Addressing Training Data Scarcity Challenge in Emerging Radio Access Networks: A Survey and Framework," *IEEE Communications Surveys & Tutorials*, 2023.

[14] A. Marey, M. Bal, H. F. Ates, and B. K. Gunturk, "PL-GAN: Path Loss Prediction Using Generative Adversarial Networks," *IEEE Access*, vol. 10, pp. 90 474–90 480, 2022.

[15] 3rd Generation Partnership Project, "TR36.814 v9.0.0: Further Advancements for E-UTRA Physical Layers Aspects (Release 9)," 3GPP, Sophia Antipolis, France, Technical Report TR36.814 v9.0.0, March 2010.

[16] "Atoll Radio Frequency Planning & Optimisation Software," https://www.forsk.com/atoll-overview, accessed: February 29, 2024.

[17] "SDMetrics: Synthetic Data Metrics," accessed: February 29, 2024. [Online]. Available: https://docs.sdv.dev/sdmetrics/

[18] "Synthetic Data Vault," accessed: February 29, 2024. [Online]. Available: https://docs.sdv.dev/sdv/