Towards Addressing the Spatial Sparsity of MDT Reports to Enable Zero Touch Network Automation

Joel Shodamola, Haneya Qureshi, Usama Masood and Ali Imran
AI4Networks Research Center, Dept. of Electrical & Computer Engineering, University of Oklahoma, Tulsa, USA
{joelshodams,haneya,usama.masood,ali.imran}@ou.edu

GLOBECOM 2021 - 2021 IEEE Global Communications Conference | 978-1-7281-8104-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GLOBECOM46510.2021.9686011

***Abstract*—Minimization of Drive Test (MDT) reports are a key enabler for Machine Learning (ML)-based zero-touch automation envisioned for emerging cellular networks. However, due to numerous factors, the MDT reports are spatially sparse in nature. This sparsity undermines the performance of ML models that are built on the MDT data to estimate and optimize network KPIs. In this paper, we present and evaluate a framework to address this challenge. We leverage generative models, specifically, Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) to augment the sparse multi-dimensional MDT data. Unlike image data where the quality of synthetic images produced by the generative models can be evaluated visually, establishing the authenticity of tabular synthetic data is a more complex problem. We address this problem by leveraging a tripartite approach: 1) We use several statistical measures to quantify the resemblance of synthetic data with original data. 2) We compare the performance of an ensemble learning model trained on augmented data, with that of trained on original data only 3) We benchmark the performance of the generative models with several classical ML models. This analysis is carried out for varying levels of sparsity and reveals insights about robustness of generative models against training data sparsity as well as on suitability of various methods for evaluating the quality of the generated synthetic tabular data. Results show GAN performs considerably better compared to other approaches. The presented solution thus can be used to overcome the sparsity problem in MDT reports thereby enabling ML-based automation use cases.**

***Index Terms*—Minimization of Drive Test (MDT), GAN, VAE, RSRP, machine learning, key performance indicator (KPI), network automation**

I. Introduction

With emerging cellular network technologies, the absolute functionality of data-driven autonomous operations like self-healing, self-configuration and self-optimization will depend on the availability of data. It is envisaged that operational complexity of the network will be an albatross for operators as this complexity is assumed to scale linearly with increase in network densification [1]. Consequently, the current manual and offline planning operations which depends wholly on collection of measurements report from drive tests are becoming more obsolete and ineffective. To mollify this complication, 3GPP introduced the minimization of drive test (MDT) [2] to reinforce autonomous solutions embedded in the features of self-organizing networks.

Data-driven autonomous solutions leverage on machine learning which has the capabilities of learning intuitive characteristics from these MDT reports. These reports contain user location and network quality of service which are quantified with certain key performance indicators (KPIs). The advantages that come with MDT reports range from reduction of human intervention, reduction in operational expenditure (OPEX) as well as reduction in time-inefficiency arising from offline configurations. However, the self-organizing networks functionality has not culminated to its expected use-case capacity predominantly because of lack of representative data.

The coverage estimation maps derived from the MDT reports are accompanied with a few challenges that impede the seamless operation of intelligent network operations. Among the existing challenges are geographical positioning errors, error due to quantization and data sparsity [3]. The focus of this study is to address the data sparsity challenge. Several factors contribute to data sparsity in cellular network domain, such as:

- **Sparsity due to smaller cells**: It is expected that user traffic under small cells will be less dense compared to macro cells [3]. Hence, the reports gotten from small cell users measurement will be scanty thus leading to a sparse coverage map.
- **MDT incompatibility of user equipment (UE)**: An important factor that contributes to sparse coverage mapping is that while some UE have inbuilt compatibility to upload MDT reports, some UE manufacturers have not implemented the features of MDT.
- **Data privacy**: Full ground truth is not attained for optimization and planning due to privacy concerns. This reason contributes to the sparsity of data reported for MDT-based optimization solutions.
- **Data sparsity from network operators**: Operators do not explore all possible combinations of network variables to avoid jeopardizing quality of service in live network. This results in non-availability of relevant or rich data for machine learning exploration.

.

Several studies have proposed spatial interpolation techniques as a remedy to solving the sparsity challenge as highlighted in Section I-A. These techniques, although proven to produce close estimates to the ground truth coverage map, have limitations, like their applicability to only stationary environments, which will pose a problem for a dynamic environments. Moreover, they are limited in the feature space used for prediction, for example, classical spatial interpolation techniques like inverse distance weighted or Kriging [4] rely on the distance feature only and do not capture additional features,

such as antenna tilt or azimuth angles.

A level of intelligence is required to absorb real-time intricacy of data and enhance pro-activity. To this end, several machine learning based solutions in the cellular network domain are also proposed in literature as highlighted in Section I-A. However, these solutions are predominantly limited to the using image data as features and not tabular data [5]–[8].

We investigate alternatives that involve regenerating entire coverage maps based on inherent correlation that exist in tabular MDT data and further augmenting it. For this purpose, we leverage two deep learning based generative models namely generative adversarial network (GAN) and variational auto-encoder (VAE).

Generative models are well known for their abilities to learn how data are generated. Using neural network as the bedrock, data is passed into the generative model for training and after some iterations or convergence point, a similar data is reproduced. Contrary to most applications of generative models for cellular network in literature that are used for image data [5]–[8], the use of generative models is presented for regression based analysis in this study. We present two applications of generative model to address the sparsity challenges as follows; coverage map generation and data augmentation.

A. Related Work

Prior to the advent of intelligence in cellular networks, most studies proposed using traditional interpolation techniques [4], [9], [10] to estimate and predict missing coverage values in a geo-spatial environment. Although these techniques tend to estimate and inset missing values, they do not scale well with dynamic environments because of the limiting application to stationary environments. In [11], authors used a spatial sampling technique to approximate traffic in a network with several base stations using an intuitive metric to select loads from a few base stations. They show that this technique eases the burden of data collection of an entire network. Authors in [12] and [13] acknowledge cell outage sparsity in data base station (BS) and propose a mathematical approach called Grey-prediction that uses differential equations to predict RSRP data in the control BSs that comes from periodic updates between users and data. In [14], authors give a comparative evaluation of different interpolation techniques including kriging for localization and radio frequency estimation. They conclude that natural-neighbor interpolation has a better performance in terms of robustness to increase shadowing. Authors in [15], address the sparsity of data that comes from small cell users by using SMOTE to address data imbalance and an ensemble learning solution to classify fault diagnosis network. They show that this method reduces communication cost that occurs as a result of overhead. A cost function is proposed in [16], where authors use the function to jointly optimize a use case of capacity and coverage optimization in both uplink and downlink. In their contribution, they formulate sparsity as a function of two factors. First, availability of data within a limited parameter range (tilt) without having knowledge of users location and secondly unknown dependence between network parameters and KPIs. Simulated Annealing is used to obtain upper bound of KPIs and coordinate descent is used for tilt search in this study.

In emerging networks, network automation utilizes deep learning models that require massive amount of data to determine inherent and existing inter-dependencies that can be used to drive self-optimization and future predictive patterns. One such technique used to address the data sparsity challenge is transfer learning. For example, in [17], authors use transfer learning to address a different domain in wireless network. To fully achieve optimize important local edge caching, they utilize transfer learning from a source domain to a target domain under sparse knowledge of users content in small cells.

From classic interpolation techniques [4], [9], [10], to sampling techniques [18], and most recently, generative models [19]–[21], several literature have come up with different propositions to address different data challenges like data imbalance, data irregularity and corruption, privacy concerns and most recent and relevant to our study, data sparsity. While little attention is focused on categorical, numerical and tabular data as is the case with most cellular networks domain, majority of the literature leveraging GANs for addressing data sparsity challenge focus on using GAN to recreate synthetic image and audio data similar to full ground truth [5]–[8]. Recently, GAN has also gained much acceptance in the medical space. For example, one study [22], addresses privacy concerns by generating and evaluating synthetic tabular data generation. With reference to cellular networks, authors in [8] applied a variant of Generative Adversarial Network (GAN) to generate radio frequency estimation maps from irregular maps, where a reconstruction error loss was formulated in addition to typical traditional GAN loss to enhance stability in the generator. Authors in [23], use integrated sampling, additive noise and variation autoencoder (VAE) to generate synthetic data for localization of both indoor and outdoor environments and conclude that the proposed augmentation techniques improve the accuracy of localization. However, studies using GANs for numerical tabular data in cellular context are very few. The center of focus of authors in [24] was to generate and predict cellular traffic data for smart city usage where vanilla GAN with LSTM (long short-term memory) networks was used for time-series data. They conclude that performance increased with augmentation rates and decreased with generative data quantities. Authors in [25] use a combination of generative and classification model to address imbalance data for cell outage detection. The closest relevant study to our work is from authors in [26], where GAN was used to augment call data records (CDR). However, this work focuses on using one dimensional data, whereas our work involves the use of higher dimensions that involves tilt, azimuth and distance parameters.

B. Contributions and Organization

To the best of authors' knowledge, this work is the first to study the efficacy of GANs using multi-dimensional tabular data in cellular networks context with varying sparsity levels.

The key contributions in this work can be summarized as follows:

- We use generative models to regenerate coverage maps from sparse tabular multi-dimensional cellular data consisting of features such as user to base station distance, user antenna tilt and azimuth angles.
- We study the effect of varying data sparsity levels on coverage estimation.
- We compare the results with traditional methods, such as sampling technique and classical machine learning (CML) predictive methods.
- In order to test the authenticity of the generated synthetic data, we use a three-fold statistical and modeling analysis approach consisting of (i) evaluating the authenticity of synthetic data produced using spearman's rank correlation coefficient (SCC) and joint plot (ii) observe the effect of synthetic data on another ensemble ML model (iii) comparing its performance in terms of RMSE to other state-of-the-art synthetic data generating models. This analysis is crucial to ascertain if the synthetic data generated has the same feature characteristics that exist in the ground truth or it is just generating some random noise.

The rest of this paper is thus organized: Section II involves detailed explanation of the proposed framework which includes data collection, prediction of map using different CML methods and description of augmentation techniques. In section III, we evaluate the performance of GAN comparing it with CML methods as well as a sampling technique. We finally conclude this paper with Section IV.

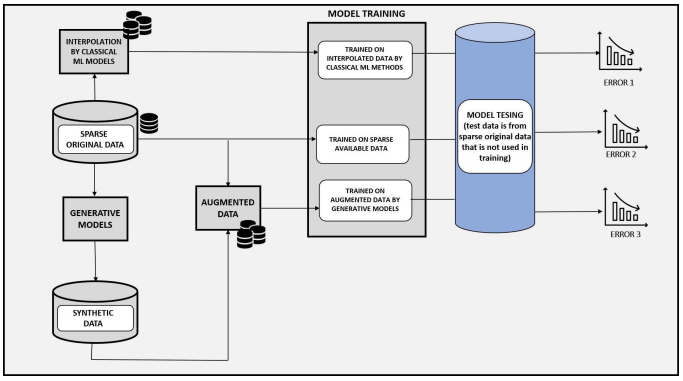


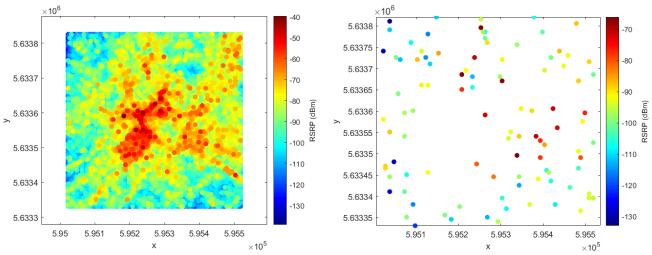
Figure 1: System framework.

II. System Model

Coverage estimation of a particular network is usually measured using reference signal received power (RSRP). This way operators can detect if there are coverage holes, blind coverage spots or poor coverage signals. Practically, the coverage area is usually divided into bins and hence, RSRP is measured based on users in each bin. Fig. 2a shows what a complete coverage map given a full ground truth would look like, however, in realistic scenarios due to the reasons highlighted in Section I, network operators do not have access to this map, hence, they are left with the task of deciphering values from a sparse map like the one presented in Fig. 2b. This work addresses this challenge by studying the ability of several CML models and deep learning based generative models to predict users received powers in the white spaces of Fig. 2b.

TABLE I: Network Measurements

System Parameters	Value
Carrier frequency	2100 MHz
Transmission power	43 dBm
Cell sectors	3 sectors per site
User distribution	Poisson
Main lobe antenna gain	18dBi
Pathloss propagation	Ray-tracing
BS height	30m
Tilt range	2.5°- 84.3°
Azimuth range	0.5°– 359°



(a) Full coverage map (b) Sparse coverage map

Figure 2: Full and sparse coverage map.

A. Data collection and coverage prediction

For this study, we utilize a commercial network planning tool with an avant-garde ray-tracing propagation model [27]. Through the data obtained from this tool, we acquire RSRP reports while leveraging on the Poisson distribution of users. To capture and reflect ground-truth of realistic coverage measurements, we inculcate the environment maps from a real world environment in the city of Brussels consisting of buildings, heights, clutters and terrain profiles. We consider one cell site location with three sectors having the same coordinates and coverage area divided into bin widths of 5m. Coverage estimation from multiple cell will be studied as part of future work. Further network measurements are listed in Table I.

For classical ML algorithms, we use four different ML algorithms namely: Random Forest, K-Nearest Neighbor, Support Vector and Linear Regression. Using the regression-based numerical data, we split into train and test, where the train represents sparse data with distance, tilt and azimuth values of each user. The training data is fed into each of the ML models and further used to predict the incomplete or white regions in the sparse map as shown in Figure 2b. For visualization purposes, we show the predicted coverage map of all models along with their RMSE values as seen in Figure 3 using 20 percent of full coverage data as available sparse data. Of all the listed models, we observe that K-Nearest Neighbor performed the best with the closest map and lowest RMSE value.

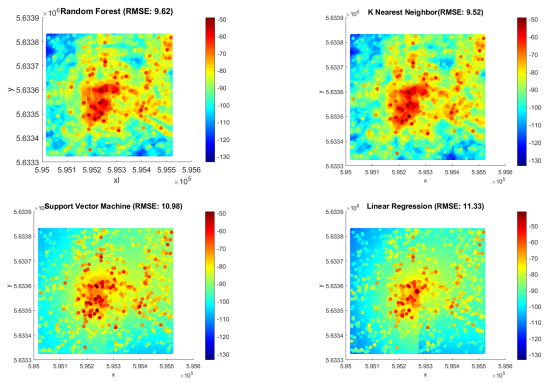


Figure 3: Comparison of selected machine learning algorithms for coverage estimation.

B. Data Augmentation

In this section, we discuss the effect of data augmentation in training a ML model, since it is established that the performance of a model to learn the network behaviour is dependent on the amount of representative data. Generative models are widely known for their ability to reproduce similar data to the ones fed into them through. They recreate how data is generated by sampling from the probabilistic models that exist in the data. Using neural network as their backbone they tend to learn more with the amount of data fed into them. Although, first application of generative models were generally used for image dataset using convolutional neural networks, recent applications are seen on tabular data. In our work, we employ and modify two types of generative models from the synthesizer in [28]; GANs and VAEs for data augmentation.

1) GAN: GAN is a type of unsupervised learning which comprises of two neural networks; generator and discriminator, where the former masters the distribution of training fed into it and maps out similar data from latent space, the latter validates the generated with real data. These two neural networks are modelled to play a mini-max game against each other. As the generator keeps learning the distribution of the original training data, it uses the parameters from the distribution to recreate similar samples from a Gaussian noise, z. Simultaneously, the discriminator acts a critic to differentiate between the true data and synthetic samples. The endpoint of this mini-max is usually user-defined or otherwise determined by a convergence point where the discriminator can no longer tell the difference between true and synthetic samples. This function is illustrated by the mathematical expressions in equations 1-3.

$$\mathbb{L}^{(D)} = -\mathbb{E}_{x\sim P_{ori}} log D(x) - \mathbb{E}_z log(1 - D(G(z))) \quad (1)$$

$$\mathbb{L}^{(G)} = -\mathbb{L}^{(D)} \quad (2)$$

$$where \mathbb{L}^{(G)} = -\mathbb{E}_z log(1 - D(G(z))) \quad (3)$$

where D(x) is the real data and G(z) is the generated data and the cross-entropy loss for correct classification given as $L^{(D)}$.

2) Variational autoencoders (VAE): Like GANs, VAE comprises of the encoder and decoder network. Where the encoder network compresses the input to a hidden latent structure of lower dimension, the decoder tends to reconstruct the distribution from the latent space back to the dimension of the input data. The general loss formulation of variation auto encoder is from an Evidence Lower Bound (ELBO) which consists of the reconstruction loss and KL divergence term as illustrated in the expression in equation 4.

$$\mathcal{L}_{vae} = \mathbb{E}_x[\mathbb{E}_z[\mathcal{D}(\mathcal{E}(x))] + \mathbb{KL}(\mathcal{N}(\mu, \sigma)||\mathcal{N}(0, I)] \quad (4)$$

Where $\mathcal{E}(x)$ and $\mathcal{D}(x)$ represent the Encoder and Decoder term respectively. The right term in the above equation tends to minimize the KL divergence of the latent distribution to a Gaussian distribution. Where the KL equation in the right-hand acts as a regularizer. Both models tend to learn and absorb the joint distribution that exists in the training data fed into them. In our work, we take the coordinates, distance to base station, tilt and azimuth of user equipment as input features as well as the corresponding RSRP values to be mapped for each parameter setting.

III. Evaluation and Results

Evaluating the quality of synthetic data to test if the synthetic data generated has the same feature characteristics that exist in the ground truth or it is just generating some random noise is crucial and an open research question. We evaluate GAN performance in three ways: (1) evaluating the authenticity of synthetic data produced using spearman's rank correlation coefficient (SCC) and joint plot. (2) Observe the effect of synthetic data on an ensemble ML model and (3) comparing its performance in terms of RMSE to other state-of-the-art synthetic data models.

Spearman's Correlation Coefficient is a metric used to measure the monotonicity of the relationship between two data or variables. In this work, we use SCC instead of Pearson Correlation (PC), because the latter tends to assume normal distribution of both variables as well as evaluate the linear relationship. SCC does not have this assumption and can capture nonlinear relationship that exists between the variables with less sensitivity to outliers. The SCC scores are between [-1 +1], where 0 means no correlation, +1 indicates direct proportionality and -1 indicates inverse proportionality between the variables or dataset. SCC is computed using the formula:

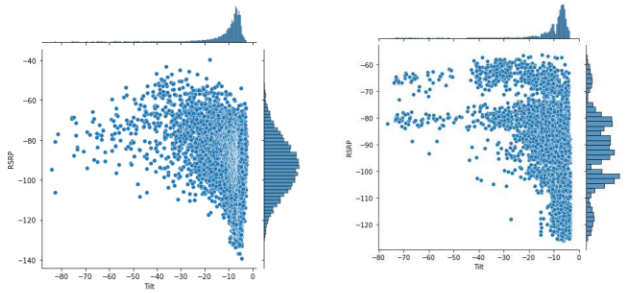
$$\mathbb{S}_{(r)} = \frac{6\Sigma r_i^2}{n(n^2-1)} \quad (5)$$

Where n is the sample size and r is the difference between the variable ranks of observation. To calculate the SCC value, we first compare the distance with the RSRP of the Original and further observe if there is a similar relationship in GAN. As seen from Table II, negative values, mean that increased distance from the base station, will yield a reduced RSRP value. GAN is able to reflect this relationship in both the

TABLE II: Spearman's correlation coefficient between RSRP and distance or tilt of original data and synthetic data from GAN.

Data Size	Original(Distance)	GAN (Distance)	Original (Tilt)	GAN (Tilt)
1%	-0.44	-0.27	-0.44	-0.35
5%	-0.56	-0.57	-0.51	-0.56
10%	-0.46	-0.38	-0.4	-0.29
50%	-0.46	-0.41	-0.40	-0.43
100%	-0.47	-0.67	-0.41	-0.68

distance and tilt feature. In addition, we visualize the behaviour and validate the authenticity of the synthetic data gotten from GAN, we observe using joint plots of RSRP and tilt values with as little as one percent training data. We point out that the training samples contained 8000 samples, where 1% represents 80 samples. As seen from Figure 4, GAN tends to produce similar distribution plots of synthetic data when trained with as little as 80 samples. We state that this is only true if the sparse data has relevant characteristics and information between the parameters and KPIs.



(a) Ground Truth

(b) Synthetic Data from GAN trained on 1 % data

Figure 4: Joint distribution between RSRP and tilt values.

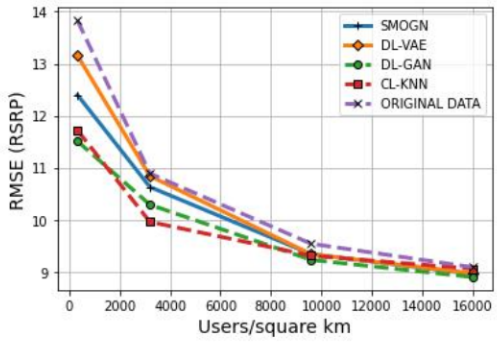


Figure 5: Comparison of data augmentation techniques.

To convey relevance in cellular network domain, we compute the performance of GAN in terms of users density. Each sparsity level having training sample size of (1,10,50,100)% has been converted to connote (320,3200,9600,16000) users per square km. We convey this data in terms of user density because network operators are not able to tell what percentage of data they have access to compared to ground-truth, however, based on MDT reports they can get the user density information.

As mentioned in Section II, we split the data into training and testing size of 0.8 and 0.2 respectively. The training data is further divided into several percentage sizes to represent various data sparsity levels. Each training data size is then fed into GAN to produce synthetic data. We used a batch size of 100 for each training sample in the exception of 1%. In addition, we ran a step-size iteration of 10 to 20 for each epoch depending on the size of the training sample. We further evaluate the response of an ensemble learner by observing its performance with respect to errors when training with sparse data and with data augmented from the GAN or its related models. We do this to benchmark the performance of augmented synthetic data gotten from GAN to other state-of-the-art techniques like SMOGN, VAE and CML. Here we selected the best classical machine learning model, CL-KNN from Section II. For generalization, we conducted a double cross validation of 10-fold also known as repeated fold. As observed from Figure 5, out of the two generative models, GAN achieved the lowest RMSE value for both the lowest and highest user density and is comparable in performance to the best performing algorithm from classical machine learning in Section II, CL-KNN. This shows that with very little but relevant data, GAN is able to generate similar data that can improve performance of machine learning model.

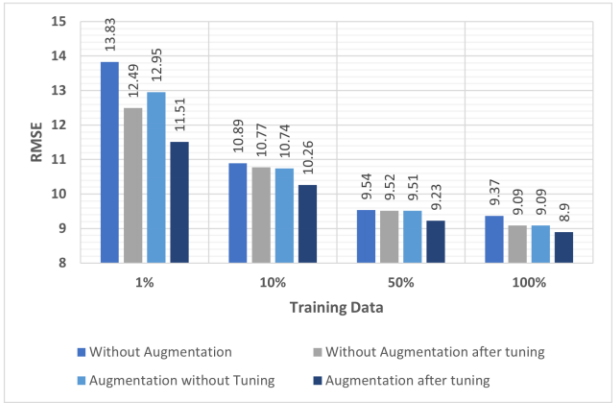


Figure 6: Effect of parameter tuning.

It is also worthy to observe the effect of parameter tuning to increase the performance of any choice of ML model. This is evident from Figure 6, as we compare the performance of XGBoost model with default parameters and tuned parameters using the synthetic data from GAN. For this purpose, we employed grid-search method to obtain the best parameter values for each of the sparsity level. The hyperparameters tuned were the maximum, depth, estimators, learning rate and gamma values. The optimum batch size increased with training size, while the epoch size ranged between 300 and 500. It was observed that the value of maximum depth increased proportionally to the size of the data. ML model hyperparameter optimization plays an important role in the field of artificial

intelligence as different models respond differently to the quantity and quality of data fed into them. Lastly, convergence of the generator in GAN becomes a major underlying issue particularly when the distribution to be modeled involves large dimensions. This challenge will be investigated and addressed in future works.

IV. Conclusion

In this study, we investigated the use of generative models to predict and augment sparse MDT reports for coverage estimation using multi-dimensional cellular data under varying sparsity levels. We evaluated the authenticity of the synthetic data generated by several generative models and classical machine learning models using statistical measures and observing the effect of synthetic data generated by them on another ensemble ML model. Results show that out of the two-deep learning-based generative models used in this study, GANs are able to better learn the intrinsic characteristics and improve AI-assisted data-driven network automation solutions even with little representative data as compared to several classical ML-based and traditional sampling approaches. MDT reports are key enabler for ML-based zero-touch automation, however their sparsity thwarts their practical use for ML-based reliable model training. The presented framework presents a method to overcome this challenge thereby paving the way for practical adaption of MDT reports for ML-based zero-touch automation.

Acknowledgment

This work is supported by the National Science Foundation under Grant Numbers 1923669 and Qatar National Research Fund under Grant No. NPRP12-S 0311-190302. The statements made herein are solely the responsibility of the authors. For more details about these projects please visit: http://www.ai4networks.com

References

[1] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: how to empower SON with big data for enabling 5G," *IEEE Network*, vol. 28, no. 6, pp. 27–33, 2014.

[2] V. . 3GPP, document TS 37.320, "Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2 Release 10," 2010.

[3] H. N. Qureshi, A. Imran, and A. Abu-Dayya, "Enhanced mdt-based performance estimation for ai driven optimization in future cellular networks," *IEEE Access*, vol. 8, pp. 161 406–161 426, 2020.

[4] J.-S. Ryu, M.-S. Kim, K.-J. Cha, T. H. Lee, and D.-H. Choi, "Kriging interpolation methods in geostatistics and dace model," *KSME International Journal*, vol. 16, no. 5, pp. 619–632, 2002.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[6] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[7] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1947–1962, 2018.

[8] Z. Li, J. Cao, H. Wang, and M. Zhao, "Sparsely self-supervised generative adversarial nets for radio frequency estimation," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2428–2442, 2019.

[9] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolating methods for image resampling," *IEEE Transactions on medical imaging*, vol. 2, no. 1, pp. 31–39, 1983.

[10] A. S. Agung Setianto and T. T. Tamia Triandini, "Comparison of kriging and inverse distance weighted (idw) interpolation methods in lineament extraction and analysis," *Journal of Southeast Asian Applied Geology*, vol. 5, no. 1, pp. 21–29, 2013.

[11] U. Paul, L. Ortiz, S. R. Das, G. Fusco, and M. M. Buddhikot, "Learning probabilistic models of cellular network traffic with applications to resource management," in *2014 IEEE International Symposium on Dynamic Spectrum Access Networks (DYSPAN)*. IEEE, 2014, pp. 82–91.

[12] O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "Cell outage detection in heterogeneous networks with separated control and data plane," in *European Wireless 2014; 20th European Wireless Conference*. VDE, 2014, pp. 1–6.

[13] O. Onireti, A. Zoha, J. Moysen, A. Imran, L. Giupponi, M. A. Imran, and A. Abu-Dayya, "A cell outage management framework for dense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, 2015.

[14] S. Üreten, A. Yongaçoğlu, and E. Petriu, "A comparison of interference cartography generation techniques in cognitive radio networks," in *2012 IEEE International Conference on Communications (ICC)*. IEEE, 2012, pp. 1879–1883.

[15] Y. Wang, K. Zhu, M. Sun, and Y. Deng, "An ensemble learning approach for fault diagnosis in self-organizing heterogeneous networks," *IEEE Access*, vol. 7, pp. 125 662–125 675, 2019.

[16] S. Berger, M. Simsek, A. Fehske, P. Zanier, I. Viering, and G. Fettweis, "Joint downlink and uplink tilt-based self-organization of coverage and capacity under sparse system knowledge," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2259–2273, 2015.

[17] W. Wang, Q. Liao, and Q. Zhang, "Cod: A cooperative cell outage detection architecture for self-organizing femtocell networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 6007–6014, 2014.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[20] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances in neural information processing systems*, 2016, pp. 2352–2360.

[21] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019, pp. 7335–7345.

[22] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," *BMC Medical Research Methodology*, vol. 20, pp. 1–40, 2020.

[23] H. Rizk, A. Shokry, and M. Youssef, "Effectiveness of data augmentation in cellular-based localization using deep learning," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2019, pp. 1–6.

[24] Z. Wang, J. Hu, G. Min, Z. Zhao, and J. Wang, "Data augmentation based cellular traffic prediction in edge computing enabled smart city," *IEEE Transactions on Industrial Informatics*, 2020.

[25] T. Zhang, K. Zhu, and D. Niyato, "A generative adversarial learning-based approach for cell outage detection in self-organizing cellular networks," *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 171–174, 2019.

[26] B. Hughes, S. Bothe, H. Farooq, and A. Imran, "Generative adversarial learning for machine learning empowered self organizing 5G networks," in *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2019, pp. 282–286.

[27] "Atoll, [online] available:https://www.forsk.com/."

[28] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv preprint arXiv:1811.11264*, 2018.