Synthetic CSI Dataset Generation for Urban Wireless Slices

Ahmed Aredah, EmadelDin A Mazied, Lamice Albaayno, and Taeesh Azal Assadi
Department of Computer Science, Virginia Tech
Blacksburg, VA 24061, USA
{ahmedaredah,emazied,lamicebaayno,taeeshazalassadi}@vt.edu

ABSTRACT

Advancing urban computing for human development relies on resilient wireless infrastructure to enable seamless connectivity among everything. Accurate forecasting of wireless channel quality contributes to achieving resilient wireless systems and robust planning of future wireless network infrastructure essential for urban development. However, the limited availability of realistic datasets presents a significant challenge for developing predictive models. In this report, we address this challenge by exploring synthetic CSI data generation models to augment wireless resource planning. We adopt a Generative Adversarial Network (GAN) model to synthesize data, utilizing Stochastic-geometry Channel Model (SCM) based simulated datasets for training. These simulated datasets are calibrated against a small set of realistic data to enhance their relevance and reliability.

KEYWORDS

Synthetic CSI datasets, RAN slicing, Generative Adversarial Networks

ACM Reference Format:
Ahmed Aredah, EmadelDin A Mazied, Lamice Albaayno, and Taeesh Azal Assadi. 2024. Synthetic CSI Dataset Generation for Urban Wireless Slices. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (CS-5834 Intro to Urban Computing).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Advances in urban computing for sustainable urban and suburban development rely on resilient wireless infrastructure as a key enabler of various urban verticals [35]. To meet the diverse demands for wireless services, particularly in urban areas, 3GPP[2] and the O-RAN Alliance[23] introduced RAN slicing, which ensures performance isolation among wireless services by preventing the degradation of one service from affecting others.

The deployment of RAN slicing hinges on understanding fluctuations in two categories of network datasets: i) wireless network conditions and ii) urban use-case demands. Accurate forecasting of wireless conditions is critical, as resource demands are directly tied to channel quality and capacity, which is determined based on capturing the statistical variations in Channel State Information (CSI).

Channel State Information (CSI)-a key metric providing a fine-grained view of the wireless channel-by which, when RAN slicing integrates this component into its framework, network performance is optimized to ensure high-quality, customized services for particular applications [33]. However, leveraging CSI effectively requires overcoming significant challenges associated with its generation and utilization.

One of the main challenges in making practical use of CSI lies in its dependence on real testbeds or computation-intensive simulations, which may complicate and delay research and development work. Recently, comprehensive CSI data set generation has emerged as one of the key solutions for such problems. It therefore enables researchers and practitioners to work with realistic models of wireless channels without any physical or computational complexities involving a testbed or large-scale simulation. Machine learning models are found promising in generating synthetic yet realistic CSI datasets. These algorithms leverage existing partial data and implement probabilistic models to simulate channel characteristics from real-world scenarios [36]. The machine learning Generative Adversarial Networks (GANs), as explored in this study, address these challenges by framing CSI generation as mini-max 2-player optimization problem.

This report aims to develop a framework for generating CSI datasets based on GANs which may overcome computational and logistical challenges. The main contributions are as follows: first, emphasize the impact of using noisy training datasets on the GAN performance; second, generating scalable synthetic CSI datasets capturing spatial and temporal variability in urban wireless environments.

The structure of this report is organized as follows: Section 2 provides an overview of the topic and research scope in addition to a discussion on the relevant literature. Section 3 outlines the methodology employed in the development of the GAN model. Section 4 presents an evaluation of the proposed model to testbed data, while Section 5 discusses the findings. Finally, Section 6 concludes the study, summarizing the key insights and implications of the research.

2 PRELIMINARIES

Advances in urban computing for sustainable urban and suburban development rely on resilient wireless infrastructure as a key enabler of various urban verticals [35]. Wireless communication supports the dynamic social, economic, cultural, technological, health, logistical, and environmental needs of modern communities. For instance, smart cities leverage wireless technologies for virtual social engagement (e.g., Meta and X), economic growth through

CS-5834 Intro to Urban Computing, Aug. 27–Dec. 17, 2024, Virginia Tech, VA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

digital trading and e-commerce (e.g., Amazon), and cultural enrichment via immersive virtual and augmented reality experiences, such as virtual heritage frameworks [7]. It also drives technological progress through Industry 4.0 and smart manufacturing [20], strengthens telemedicine during health crises [9, 26], and improves public safety via smart transportation systems. Moreover, applications like smart homes and environmental sensors for pollution and climate monitoring highlight wireless infrastructure's essential role in advancing urban planning and computing algorithms. These examples underscore wireless connectivity as a pivotal enabler of urban innovation and sustainable progress.

To meet the diverse demands for wireless services, particularly in urban areas, 3GPP[2] and the O-RAN Alliance[23] introduced RAN slicing, which ensures performance isolation among wireless services by preventing the degradation of one service from affecting others. This approach is especially relevant in urban settings, where diverse services require tailored wireless infrastructure to accommodate fluctuating demands over their lifetimes. RAN slicing creates on-demand virtual networks that dynamically allocate resources to adapt to changing urban demands and wireless network conditions.

The deployment of RAN slicing hinges on understanding fluctuations in two categories of network datasets: i) wireless network conditions and ii) urban use-case demands. Accurate forecasting of wireless conditions is critical, as resource demands are directly tied to channel quality and capacity, determined by the Signal-to-Interference-Noise Ratio (SINR). Improved SINR boosts channel capacity, reduces bit error rates, and lowers the computational complexity of error correction, enabling efficient resource allocation and effective wireless resource planning. Likewise, forecasting urban use-case traffic demands is crucial for predicting wireless traffic generated by diverse urban applications [11, 14, 37] and estimating the network resources required to adapt to traffic variations. A deeper understanding of these fluctuations would contribute to achieve accurate forecasting and enable the development of robust RAN slicing for urban services. This report focuses on forecasting wireless network conditions, while ongoing research addresses urban traffic demands.

Accurately capturing Channel State Information (CSI) is crucial for forecasting variations in wireless conditions and enabling robust urban wireless slice resource planning. CSI is a complex number comprising normalized In-phase (I) and Quadrature (Q) components of the received wireless signal. Mathematically, if a signal $x(t)$ is transmitted and the received signal is $y(t)$, the CSI is defined as $CSI = \frac{Y}{X}$, where Y and X are the frequency-domain impulse responses of the received and transmitted signals, respectively, calculated using the Fast Fourier Transform (FFT). Statistically, CSI is represented by the channel coefficient matrix $H = \frac{Y}{X}$, which captures the spatial and temporal properties of the wireless channel as it evolves with changes in context and network configurations.

Several factors influence the impulse response of wireless channels, including carrier frequency, user mobility, bandwidth, antenna configuration, propagation environment, the number of active wireless links, and user positions relative to base stations. For instance, urban areas are prone to scattering due to obstacles, leading to a low probability of Line-of-Sight (LoS), while rural areas often experience higher LoS probabilities but are more affected by weather conditions. These factors collectively govern the variability and complexity of wireless channels, emphasizing the importance of accurate CSI estimation and forecasting for efficient resource allocation and advanced signal processing in modern networks.

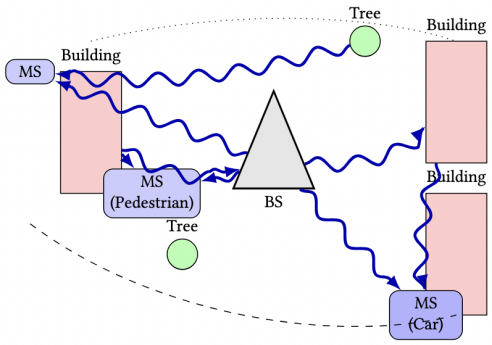

Figure 1: Capturing CSI within various wireless propagation

2.1 Research Scope

In this research project, we focus on methods for acquiring CSI datasets, which, as discussed earlier, are essential for robust urban wireless slice planning. There are two primary approaches to obtaining CSI datasets: i) real-time measurements and ii) synthetic generation methods. While real-time measurements provide ground-truth CSI values, they are highly resource- and labor-intensive; thus, neither scalable nor adaptive. Additionally, the deployment of RAN slicing is still in its early stages, with only a few trial implementations conducted in urban areas [12, 21]. These limitations make synthetic generation a more practical and widely adopted approach in the research and development of wireless network design.

Synthetic CSI datasets offer the flexibility to model diverse wireless environments, operational scenarios, and use-case demands that are difficult to capture comprehensively through real-world measurements. However, generating reliable synthetic datasets raises critical research questions, such as:

- *How can we ensure that synthetic datasets accurately reflect the variations in wireless channels under urban scenarios?*
- *What models and algorithms can effectively simulate the spatial and temporal dynamics of urban wireless environments?*
- *How can synthetic datasets be validated to ensure their relevance for RAN slicing design and resource planning?*

Addressing these questions is crucial for advancing RAN slicing research. Therefore, this report focuses on the synthetic generation of CSI datasets, framing our research objectives to explore methods for generating reliable datasets that reflect real-world wireless channel conditions in urban settings, with a particular emphasis on their application in RAN slicing design for urban use cases.

In this context, we consider a RAN slicing system where the wireless spectrum is dynamically allocated among various service

categories, referred to as slice service types (SSTs). Each SST is characterized by its slice bandwidth (B^s), and slice carrier frequency (f_c^s), both determined by spectrum slicing algorithms, such as the approach detailed in [14]. While the specific techniques for spectrum slicing and the computation of optimal radio numerology patterns for dynamic bandwidth allocation are beyond the scope of this report, we assume that these data are accessible to the system.

It is important to note that the spectrum slicing process is inherently dynamic. As a result, the bandwidth per slice (B^s) varies over time based on the slicing configuration, and the slice carrier frequency (f_c^s) is determined by the allocated bandwidth and the frequency offset between the main carrier frequency and the slice carrier frequency. For a detailed explanation of the spectrum slicing process, we refer to [14].

In this report, we focus on the temporal variability of B^s and f_c^s, and the antenna configuration per SST, which are treated as random variables that evolve over time according to the behavior of spectrum slicing optimization algorithms. This temporal variability plays a critical role in the generation of synthetic CSI datasets, as it reflects the dynamic nature of wireless resources in urban RAN slicing scenarios. Understanding these variations is crucial for accurately modeling urban wireless environments and tailoring network resources to meet fluctuating service demands.

2.2 Related Work

The synthetic generation of Channel State Information (CSI) datasets has been extensively explored in three primary research directions: i) stochastic geometry-based channel models, which utilize empirical data, geometric characteristics, and propagation environments to derive the statistical properties and variations of wireless channels, e.g.,[19, 25]; ii) stochastic correlated-based channel models, which capture spatial and temporal correlation in wireless channels, e.g.,[31, 32]; and iii) Machine Learning (ML)-based synthetic CSI generation methods, which leverage data-driven approaches to generate CSI datasets with high accuracy and adaptability to complex scenarios, e.g., [6, 8, 10, 13, 15–18, 22, 24, 27–29, 34].

2.2.1 ***Stochastic geometry methods.*** *Stochastic geometry-based channel models* generate CSI by simulating wireless channels using empirical data and geometric characteristics of the propagation environment to provide statistical properties of wireless channels and their variations. The WINNER II model [19] offers comprehensive modeling for diverse propagation scenarios but is computationally intensive for large-scale wireless networks with diverse use-cases.

To address this limitation, the Clustered Delay Line (CDL) model simplifies CSI generation by capturing temporal and spatial correlations in multiple wireless links. Pessoa et al. [25] introduced a CDL-based model integrating measurements with 3GPP's CDL profiles to enable fast simulations for link- and system-level analyses in rural settings.

While CDL models are simpler and more computationally efficient, particularly for use cases like 5G multiple-in multiple-out (MIMO) in remote areas, they lack the fine-grained adaptability and detailed propagation modeling of WINNER II. The two models reflect a trade-off between simulation speed and detailed scenario adaptability, serving distinct roles based on wireless network complexity and scale.

2.2.2 ***Statistical correlation methods.*** *Stochastic correlated-based channel models* synthesize CSI datasets by capturing statistical properties of the wireless channel, providing a tractable framework for analyzing complex wireless environments. The Kronecker-based stochastic model (KBSM) synthesizes spatial correlations by introducing two separate transmitter and receiver correlation matrices, but its oversimplification, such as neglecting scatterer evolution, limits its realism, particularly in massive MIMO scenarios. Wu et al. [32] addressed this limitation by introducing an enhanced KBSM incorporating a birth-death process to model scatterer evolution along large antenna arrays, improving spatial correlation accuracy.

On the other hand, pervasive correlated channel models (PCCMs) offer a more flexible framework, representing MIMO channels with high statistical accuracy. Wang et al. [31] introduced PCCM, which models joint correlations in both the magnitude and phase of channel coefficients using independent and nonidentically distributed random variables. PCCM achieves compatibility with existing stochastic models like KBSM while providing greater fidelity in complex propagation environments.

While KBSM prioritizes simplicity and computational efficiency, PCCM offers enhanced realism at the cost of increased complexity. However, these models primarily capture the high-level statistical properties of wireless channels and have limitations in reflecting the impact of antenna geometries and terrain variations in wireless environments. They provide valuable statistical insights but lack the granularity needed to capture dynamic CSI variations across diverse propagation scenarios.

2.2.3 ***Machine learning methods.*** A variety of ML algorithms have been developed to synthesize near-realistic CSI datasets for wireless network planning. In the following, we highlight recent ML-based approaches for CSI dataset generation in different contexts.

Probabilistic and Dimensionality Reduction Models such as Principal Component Analysis (PCA) combined with probabilistic modeling have been adopted to generate CSI datasets for millimeter-wave (mmWave) channels. These methods maintain computational efficiency while capturing spatial and temporal variations in mmWave wireless channels, providing near-realistic and scalable datasets for developing resource management algorithms [27]. Nevertheless, while computationally efficient, these models are constrained in their ability to capture complex, dynamic variations in diverse propagation environments, limiting their adaptability to urban wireless slices.

Generative Adversarial Networks (GANs) models replicate realistic channel conditions by learning from limited labeled CSI data, enabling the generation of large, high-fidelity synthetic datasets. These models support resource allocation interventions by addressing data scarcity and enhancing robustness to channel variability [15]. However, GANs require extensive tuning and are prone to instability during training, particularly when modeling highly dynamic urban wireless environments. Ensuring the accuracy of generated datasets across diverse use cases also remains a challenge.

Hybrid Neural Network models combining Convolutional Neural Networks (CNNs) and attention mechanisms effectively compress and reconstruct CSI matrices. These approaches reduce feedback overhead while maintaining accuracy in channel representation, supporting scalable 5G applications in dynamic urban systems [24]. Unfortunately, these models often require large computational resources for training and are sensitive to architectural choices, which can limit their generalizability across varying propagation environments. Most importantly, large volumes of realistic datasets are necessary for constructing reliable CSI datasets.

Digital Twins frameworks integrate untrained neural networks with conditional GANs to reconstruct CSI using prior spatial and temporal channel knowledge. These models offer low-overhead solutions suitable for high-mobility and complex urban settings [6]. Despite their advantages, their reliance on prior knowledge restricts their applicability in scenarios lacking accurate initial models, reducing scalability in diverse RAN slicing contexts.

Deep Learning-Based Prediction Frameworks, such as 3D CNNs, leverage historical data to predict future CSI by capturing spatial and temporal correlations. These frameworks leverage the synthetic CSI data to optimize resource allocation in massive MIMO systems, enabling proactive management with reduced signaling overhead [8]. Nonetheless, their performance depends heavily on the availability and quality of historical data, which may not adequately represent highly dynamic urban wireless scenarios.

Combined Probabilistic and Neural Network Frameworks synthesize CSI using PCA and multivariate normal distributions, providing realistic data for urban environments [28]. Neural networks trained on realistic 5G operational scenarios further enhance the accuracy of predicted CSI data [29]. However, these approaches are domain-specific, and they can not be generalized across diverse deployment scenarios and require significant computational resources.

Despite their significant advancements, current methods encounter limitations in addressing the complexities of dynamic, heterogeneous wireless environments: i) most models struggle to generalize across diverse urban and sub-urban propagation scenarios and the dynamic operations of urban wireless slices with high mobility and interference; ii) computational efficiency and scalability remain challenging for hybrid and deep learning models in large-scale deployments; iii) existing frameworks lack a unified approach to integrate probabilistic, generative, and neural network-based methods, which could better capture the multifaceted characteristics of wireless channels. Therefore, further research is needed to develop synthetic CSI generation methods that combine high fidelity, scalability, and computational efficiency while adapting to the dynamic variations of urban wireless slices.

2.3 Problem Statement

This report addresses the problem of developing a synthetic CSI dataset tailored for urban wireless slice design. The primary challenge lies in capturing the rapid, high-entropy fluctuations of wireless channels, which exhibit temporal independence—each CSI instance is statistically independent of its predecessor—while remaining context-dependent, influenced by the wireless environment, antenna system configurations, and the users' mobile activities. The objective is to create a CSI generation framework that models the uncertain yet context-specific dynamics of wireless channels to support robust urban wireless slice resource planning.

3 METHODOLOGY

In this report, we propose methods for generating synthetic CSI datasets based on the defined problem, where the goal is to ensure that the joint distribution of the generated synthetic data closely approximates that of the realistic CSI dataset, expressed as $\mathbb{P}(\mathbf{x_{synth}}) \approx \mathbb{P}(\mathbf{x_{real}})$. Here, $\mathbb{P}(\mathbf{x})$ represents the joint distribution of the matrix $\mathbf{x}$, and $\mathbf{x_{synth}}$ and $\mathbf{x_{real}}$ denote to the synthetic and realistic CSI datasets, respectively.

As discussed in Section **??**, obtaining realistic CSI datasets for all wireless operational scenarios remains challenging. To address this limitation, we outline a methodology for simulating realistic propagation scenarios to generate CSI datasets for selected operational conditions. These simulated datasets serve as essential training data for generative models.

In the following, we first examine the key statistical properties of CSI datasets. Next, we describe the process of generating training data in situations where only limited realistic CSI datasets are available for a specific wireless operational scenario. Then, we present methods for generating synthetic CSI datasets that can generalize to diverse operational scenarios, with a particular focus on their applicability to RAN slicing operations. Following that, we underline the metrics used for performance evaluation.

3.1 CSI Data Statistical Properties

CSI datasets have two unique properties: *i) Temporal Independence*; and *ii) Conditional Dependence Within a Data Row*. On one hand, each CSI row of the CSI dataset, corresponding to a single time instance, is independent of the preceding and successive rows, and thus, there is no temporal dependency between rows. On the other hand, the joint probability distribution of the attributes within a row (e.g., antenna configuration, user mobility, network slice configuration, SINR, Real, and Imaginary components) exhibits conditional dependence, which prohibits factorization into independent distributions.

Consider that the CSI dataset consists of n rows, where each row $\mathbf{x}_i$ represents a snapshot of the CSI data at epoch t_i with m attributes such that $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}], \quad i \in \{1, 2, \ldots, n\}$, where $x_{i,j}$ is the j-th attribute (e.g., antenna geometry, user mobility, SINR, real part, imaginary part) in the i-th row.

The CSI data rows are highly time-varying that are generated at each epoch t such that each row $\mathbf{x}_i$ is sampled from underlying random distribution $P(\mathbf{x})$ and the rows $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ are independent of each other [31]. Therefore, $P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} P(\mathbf{x}_i)$. In this sense, each row $\mathbf{x}_i$ is independent of all other rows, and therefore:

$$P(\mathbf{x}_i \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}) = P(\mathbf{x}_i) \tag{1}$$

Thus, the rows of the CSI dataset are temporally independent by definition.

Within each row $\mathbf{x}_i$, the attributes $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ exhibit *conditional dependence* under fixed conditions C (e.g., antenna geometry, propagation scenario). Specifically, $P(\mathbf{x}_i \mid C) \neq \prod_{j=1}^{m} P(x_{i,j} \mid C)$. Consider that $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,m}]$ is the set of attributes in the

i-th row that C represents a set of fixed conditions (e.g., antenna geometry, propagation scenario). Then the joint probability distribution of the attributes in a row, conditioned on C can be written as follows

$$P(\mathbf{x}_i \mid C) = P(x_{i,1}, x_{i,2}, \ldots, x_{i,m} \mid C). \tag{2}$$

To show that, consider the attributes $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ include antenna geometry, user mobility, SINR, Real component of CSI (In-phase I), Imaginary component of CSI (Quadrature Q). These attributes are ***interdependent*** because the real and imaginary parts are projections of the same complex-valued channel coefficient CSI. Furthermore, SINR, for example, is another measure that reflects the channel quality, which itself is strongly correlated with the CSI [31]. Likewise, the CSI is dependent on the antenna geometry and user mobility. Thus, under the fixed condition C, the joint distribution must capture these inter-dependencies. For example, $P(SINR, I, Q \mid C) \neq P(SINR \mid C) \cdot P(I \mid C) \cdot P(Q \mid C)$. However, the joint distribution of this subset of attributes can be written as follows.

$$P(SINR, I, Q \mid C) = P(SINR \mid I, Q, C) \cdot P(I, Q \mid C), \tag{3}$$

where $P(I, Q \mid C)$ is the joint distribution of the CSI's real and imaginary components, and the $P(SINR \mid I, Q, C)$ is the conditional distribution of SINR given the real and imaginary components. In this sense, Equation 3 captures the mutual dependencies among the attributes. Therefore, the joint distribution cannot be factorized into independent distributions of the individual attributes such that $P(\mathbf{x}_i \mid C) \neq \prod_{j=1}^{m} P(x_{i,j} \mid C)$.

3.2 Baseline

In this project, we construct our baseline scenario for synthetic CSI data generation by using the stochastic-geometry channel model (SCM) [19], also known as WINNER II. In SCM, the joint distribution of CSI data $\mathbb{P}(\mathbf{x})$ can capture spatial correlations across the feature-dependent CSI target classes in the datasets. In this model, the wireless channel is represented as a superposition of clusters and radio paths (i.e., paired transmitter-receiver radio links between antenna elements), which can be expressed as follows.

$$\mathbf{H}(t, fs^k) = \sum_{c=1}^{C} \sum_{l=1}^{L_c} \mathbf{H}_{c,l}(t, f^{s_k}), \tag{4}$$

where $\mathbf{H}(t, fs^k)$ represents the channel matrix as a function of time t and slice carrier frequency fs^k, C denotes the number of radio clusters, where each cluster comprises multiple radio links that are scattered or reflected due to propagation in a non-line-of-sight (NLOS) wireless environment. L_c represents the number of scattered and reflected radio links in the c-th cluster, and $\mathbf{H}_{c,l}(t, fs^k)$ denotes the contribution of the l-th radio link within the c-th cluster. Each channel component $\mathbf{H}_{c,l}(t, fs^k)$ in equation **??** can be written as follows.

$$\mathbf{H}_{c,l}(t, f^{s_k}) = \mathbf{A}_{c,l} e^{j\phi_{c,l}} e^{j2\pi f_d^s t} e^{-j2\pi f^{s_k} \tau_{c,l}}, \tag{5}$$

where $\mathbf{A}_{c,l}$ is the amplitude matrix, which includes antenna gains and path loss of l-th link in c-th cluster, $\phi_{c,l}$ random phase offset, which uniformly distributed in $[-\pi, \pi]$, f_d^s is the doppler frequency due to s-th slice's user mobility, $\tau_{c,l}$ is the time delay of the l-th radio link in the c-th cluster.

Therefore, the joint probability distribution of the channel coefficients (i.e., the CSI dataset) can be expressed as follows.

$$\mathbb{P}(\mathbf{x}) = \mathbb{P}(\mathbf{H}) = \prod_{c=1}^{C} \prod_{l=1}^{L_c} \mathbb{P}(\mathbf{A}_{c,l}, \phi_{c,l}, \tau_{c,l}, \theta_{AoA}, \theta_{AoD}), \tag{6}$$

where $\mathbf{A}_{c,l}, \phi_{c,l}, \tau_{c,l}, \theta_{AoA}, \theta_{AoD}$ are independent random variables. Thus:

$$\mathbb{P}(\mathbf{A}_{c,l}, \phi_{c,l}, \tau_{c,l}, \theta_{AoA}, \theta_{AoD}) = \mathbb{P}(\mathbf{A}_{c,l})\mathbb{P}(\phi_{c,l})\mathbb{P}(\tau_{c,l})\mathbb{P}(\theta_{AoA})\mathbb{P}(\theta_{AoD}). \tag{7}$$

The $\mathbb{P}(\mathbf{A}_{c,l})$ models the generalized Gamma Complex Gaussian random distribution of the amplitude matrix, which statistically captures the variations in the CSI amplitude as described in [31]. Furthermore:

$$\mathbb{P}(\phi_{c,l}) = \frac{1}{2\pi}, \quad \phi_{c,l} \in [-\pi, \pi], \tag{8}$$

$$\mathbb{P}(\phi_{c,l}) = \lambda e^{-\lambda \tau_{c,l}}, \quad \tau_{c,l} \geq 0, \tag{9}$$

where λ is the rate parameter of the exponential distribution. The $\mathbb{P}(\theta_{AoA}, \theta_{AoD})$ is the angular random distribution of arrival and departure of radio beams modeled as Wrapped Gaussian random variables [19]. Accordingly, equation 6 defines the joint probability distribution of CSI data based on empirical data for each propagation scenario and wireless network configuration.

Based on the SCM analytical framework, we leverage the capabilities of WINNER II to develop a simulated CSI generation method, serving as a baseline scenario for our work. In Algorithm 1, the CSI dataset is generated using the WINNER II channel model for a RAN slice, which is assumed to be pre-configured to share the wireless spectrum through spectrum slicing techniques described in [14]. The input parameters include antenna configurations, slice carrier frequency, number of frames, and user distances.

It is important to highlight that the Signal-to-Interference-plus-Noise Ratio (SINR) is modeled by assuming a typical value for transmitted signal power strength and Additive White Gaussian Noise (AWGN), which follows a normal distribution. These values are pre-processed before the algorithm executes.

The process begins by initializing the base station (BS) and user equipment (UE) antenna arrays. The positions are translated to non-negative coordinates, and the distance vector and azimuth angles are computed. For each BS antenna, a complex Gaussian transmit signal is generated, scaled, and passed through the WINNER II channel model to simulate the received signal. The CSI response is extracted for each UE antenna and stored in a matrix.

In the post-processing stage, the real and imaginary components of CSI are extracted at each epoch for all transmit and receive antennas. Metadata, including the numerology pattern, antenna geometry, distance, and azimuth, is appended to the CSI data. The final dataset is saved in CSV format, providing a structured output for Massive MIMO experiments.

Algorithm 1 SCM-based CSI Generation

Input: N, d, N_u, d_u, f_c, $frameLen$, $range$, $UserDistance$
Output: CSI matrix **CSI**

Initialize parameters:
  Antenna arrays: $AA_{BS} \leftarrow 3N$, $AA_{UE} \leftarrow N_u$
  Distance to base station $D_{BS} \leftarrow$ UserDistance
  Carrier frequency f_c and frame length $frameLen$
Define WINNER II layout and channel model
Translate positions for BS and UE to non-negative coordinates
Calculate distance vector $\mathbf{d}_{vec}$ and azimuth angle θ
Create CSI Matrix: CSI $\leftarrow \mathbf{0}_{frameLen \times N_u \times 3N}$
for each base station antenna $bs\_ant \in [1, 3N]$ **do**
  Generate transmit signal **txSig**$[bs\_ant]$ with:

$$\textbf{txSig} \leftarrow \mathcal{CN}(0,1) \text{ scaled to } tx\_power$$

  Pass **txSig** through WINNER II channel model to obtain **rx_temp**
  for each mobile station antenna $ms\_ant \in [1, N_u]$ **do**
    Extract CSI response:

$$\textbf{CSI}(:, ms\_ant, bs\_ant) \leftarrow \textbf{rx\_temp}(:, ms\_ant)$$

  end for
end for
Postprocessing:
Define static parameters: numerology pattern, antenna geometry, and scenario
for each time index $t \in [1, frameLen]$ **do**
  for each transmit antenna $tx \in [1, N_u]$ **do**
    for each receive antenna $rx \in [1, 3N]$ **do**
      Extract real and imaginary components of CSI:

$$\text{real} \leftarrow \Re(\textbf{CSI}[t, tx, rx]), \quad \text{imag} \leftarrow \Im(\textbf{CSI}[t, tx, rx])$$

      Store CSI data with metadata:

numerology, geometry, distance, azimuth, real, imag

    end for
  end for
end for

3.3 Generative Adversarial Network (GAN)

The aforementioned SCM model performs well when the network system configurations are pre-determined and the network has already been deployed. However, it relies heavily on detailed information about the wireless environment and user density, which limits its practicality for generating CSI data across diverse wireless operational scenarios, particularly in urban slice settings. Moreover, SCM-based simulated datasets inherently contain noise due to assumptions about wireless configurations that may not universally hold. Additionally, the conditional dependence and temporal independence statistical properties of CSI datasets necessitate rethinking an efficient technique for developing a robust CSI generation model.

Fortunately, Generative Adversarial Networks (GANs) based methods [38], have shown significant contributions in various research domains due to their ability to capture a variety of statistical properties of datasets, even when trained on noisy or biased data. This aligns well with our objective of generating synthetic CSI data that accurately reflects the complexities and dynamics of urban wireless environments.

GANs have the capability to generate synthetic datasets that approximate the statistical properties of real-world ones. They leverage the powerful features of neural networks (NN) by adopting the concept of adversarial training, where two neural networks are involved: i) generator (G) that aims to generate synthetic data with a joint distribution similar to that of realistic data ; and ii) discriminator (D) that classifies real data samples and synthetic ones generated by the generator. Equation 10 demonstrates the formulation of the GAN approach as a two-players minimax optimization, rooted in the context of game theory [5], i.e., G is player 1 and D player 2.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_z}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right]. \tag{10}$$

where $V(D,G)$ is the value function that we aim to maximize to accurately classify real data and fake data (generated), $\mathbb{E}_{\mathbf{x}\sim p_{data}}\left[\log D(\mathbf{x})\right]$ models the expected value of the discriminator output given real data samples $\mathbf{x}$, $D(\mathbf{x})$ is the probability that $\mathbf{x}$ is real data, and $\mathbb{E}_{\mathbf{z}\sim p_z}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right]$ is the expected value of the discriminator output when given fake data generated by G using $\mathbf{z}$ random noise. Here, $1 - D(G(\mathbf{z}))$ is the probability that D classifies the generated sample $G(\mathbf{z})$ as fake. The Generator G aims to minimize its loss by turning around the discriminator D, while D seeks to maximize its accuracy to correctly classify real and fake (generated) data. In this context, the adversarial training process iteratively updates the discriminator neural network to enhance its ability to distinguish between real and synthetic data. Furthermore, the Generator enhances its capability to produce realistic samples that can deceive the discriminator. At equilibrium, the generator produces data that is closely correlated with real data, effectively approximating the true data distribution.

We adopt the GAN algorithm to generate synthetic CSI datasets. In Algorithm 2, G takes features (e.g., wireless network parameters) and generates synthetic CSI values representing the real and imaginary components of the CSI data. On the other hand, D compares the synthetic CSI with the realistic ones, which are SCB-simulated in our project, to determine their authenticity and provide feedback to the generator. The algorithm iteratively minimizes the adversarial loss using Binary Cross-Entropy Loss (BCELoss) while updating both the generator and discriminator with the Adam optimizer. The training process ensures that the synthetic CSI data captures the statistical properties of the simulated CSI datasets and approximates the joint distribution of CSI values. Therefore, the GAN algorithm involves two key phases: training and testing.

In the preprocessing stage, it reads the input CSI datasets that include features and targets (real and imaginary terms). The data is preprocessed by combining multiple CSV files into a single DataFrame. Features (X) and targets (Y) are separated, with categorical and numerical attributes. Then, the processed data is converted into tensors for training.

In the initialization and training, G takes input dimensions of data features and outputs two values representing the CSI's real and imaginary terms, while D distinguishes SCB-simulated data from synthetic data. Adam optimizer is used with learning rate η,

and the Binary Cross-Entropy Loss (BCELoss) function used for the adversarial training. The GAN trains over E epochs with batch size B. In each epoch, for a given batch of real data $\mathbf{X}_b, \mathbf{Y}_b$, where $b \in B$, G takes $\mathbf{X}_b$ features as input and outputs the deceived target values $\hat{\mathbf{Y}} = G(\mathbf{X}_b)$. On the other hand, D updates its parameters through NN back-propagation and computes its loss L_D as follows.

$$L_D = BCELoss(D([\mathbf{X}_b, \mathbf{Y}_b]), 1) + BCELoss(D([\mathbf{X}_b, \hat{\mathbf{Y}}]), 0), \quad (11)$$

G aims to deceive D by minimizing the loss $L_G = BCELoss(D([\mathbf{X}_b, \mathbf{Y}_b]), 1)$, which manifests the D's data classification. Similar to G, D updates its parameters through NN back-propagation.

In the post-processing phase (i.e., testing phase), (G) is used to forecast CSI data for unseen test inputs, which are preprocessed similarly to the training data. The generated CSI is represented as ($\hat{\mathbf{Y}}_{test} = G(\mathbf{X}_{test})$). The resulting CSI outputs are saved in CSV format, and the algorithm returns the predicted CSI matrix ($\hat{\mathbf{Y}}_{test}$).

We argue that GANs can approximate the joint distribution of CSI datasets through adversarial training, where the feature and target components exhibit statistical properties that the generator learns to replicate. However, the performance of GANs depends on the quality of the training datasets, the level of noise they contain, and whether they are balanced. This is particularly relevant when simulated datasets are used for training due to the limited availability of realistic CSI data. Despite these challenges, GANs offer a scalable approach for generating synthetic CSI data in wireless systems. We discuss our argument through our presentation of the results and discussion.

Algorithm 2 GAN-Based CSI Generation and Testing

Input: File paths $\mathbf{F}$, test file $\mathbf{F}_{test}$, model path $\mathbf{M}_{path}$, batch size B, epochs E, learning rate η
Output: Predicted CSI values $\mathbf{H}_{pred}$, test results in CSV format

Load and Preprocess Training Data:
  Combine CSV files $\mathbf{F}$ into one DataFrame
  Separate features $\mathbf{X}$ and targets $\mathbf{Y}$ (real, imag)
  One-hot encode categorical columns and normalize numerical columns
  Return tensors: $\mathbf{X}_{train}, \mathbf{Y}_{train}$
Initialize GAN Components:
  Generator G with input dimension d_{input} and output dimension 2
  Discriminator D with input dimension $d_{input} + 2$
  Optimizers: $O_G, O_D \leftarrow \text{Adam}(\eta)$
  Loss function: Binary Cross Entropy Loss (BCELoss)
if Model exists at $\mathbf{M}_{path}$ **then**
  Load pre-trained G, D, and optimizers O_G, O_D
else
  Train GAN:
  for epoch $e \in [1, E]$ **do**
    for each batch $(\mathbf{X}_b, \mathbf{Y}_b) \in$ Dataloader($\mathbf{X}_{train}, \mathbf{Y}_{train}, B$) **do**
      Generate Real and Fake Data:
      Real data: **real_data** = $[\mathbf{X}_b, \mathbf{Y}_b]$
      Fake target: $\hat{\mathbf{Y}} = G(\mathbf{X}_b)$
      Fake data: **fake_data** = $[\mathbf{X}_b, \hat{\mathbf{Y}}]$
      Train Discriminator:
      Compute loss: L_D = BCELoss(D(**real_data**), 1) + BCELoss(D(**fake_data**), 0)
      Update D: $O_D \leftarrow$ backprop(L_D)
      Train Generator:
      Compute loss: L_G = BCELoss(D(**fake_data**), 1)
      Update G: $O_G \leftarrow$ backprop(L_G)
    end for
    Print L_D, L_G for epoch e
  end for
  Save G, D, O_G, O_D to $\mathbf{M}_{path}$
end if
Test GAN Model:
  Preprocess test data $\mathbf{F}_{test}$ to tensors $\mathbf{X}_{test}, \mathbf{Y}_{test}$
  Generate predicted CSI: $\hat{\mathbf{Y}}_{test} = G(\mathbf{X}_{test})$
  Compute MSE: MSE = mean($(\hat{\mathbf{Y}}_{test} - \mathbf{Y}_{test})^2$)
Save Results:
  Combine test inputs $\mathbf{X}_{test}$ and predictions $\hat{\mathbf{Y}}_{test}$
  Save results to CSV with columns: inputs, predicted real, predicted imag
return $\hat{\mathbf{Y}}_{test}$

3.4 Evaluation Metric

To evaluate the GAN model, we use the Mean Squared Error (MSE) as a performance metric to measure the accuracy of the generated CSI data. The MSE is calculated as the mean of the squared differences between the predicted and actual CSI values, i.e., $MSE = (\hat{\mathbf{Y}}_{test} - \mathbf{Y}_{test})^2$.

3.5 Input and Output Data Descriptions

Table 1: WINNER II Model Input Configuration and Output Data

Parameter	Description	Type/Output
N	Antennas per sector (7 radios × 2)	Integer
d	Spacing between base station antennas	Float (m)
Nu	Antennas per User Equipment (UE)	Integer
du	User antenna spacing	Float (m)
f_c	Carrier frequency	Float (Hz)
range	Maximum coverage range	Float (m)
UserDistance	Distance between BS and UE	Float (m)
azimuth	Azimuth angle	Float (deg)
frameLen	Number of time samples	Integer
CSI Matrix	Simulated CSI data	Complex matrix
numerology_pattern	Numerology pattern	Integer
antenna_geometry	Antenna array configuration	String (e.g., ULA)
base_station_antennas	Total number of BS antennas	Integer
user_equipment_antennas	Total number of UE antennas	Integer
carrier_frequency	Carrier frequency	Float (Hz)
antenna_spacing	Antenna spacing	Float (m)
scenario	Propagation scenario (e.g., D1)	String
Output: CSV file	Simulated CSI data saved as CSV	File

Table 2: GAN Model Input Data and Output Description

Input Parameter	Type	Range	Description
Input Data	Float/Array	Normalized (0-1)	CSI data
Noise Vector	Float/Vector	[-1, 1]	Random noise for GAN input
Dimensions	Integer	Varying	Shape of CSI data
Output: GAN Data	Float/Array	Synthesized CSI	Generated CSI

4 EVALUATION

In this section, we outline the procedures implemented to apply the methods discussed in Section 3 for CSI dataset generation. We begin with the collection of realistic CSI datasets and leverage the SCM model to generate simulated CSI datasets for various propagation scenarios. Following that, we employ a Vanilla GAN model, tune it to match our research scope, and train it using the simulated data. Finally, we evaluate the model's performance.

4.1 Setup

Collection of realistic datasets: Realistic CSI datasets are not widely available from commercial vendors due to Non-Disclosure Agreements (NDAs). However, the National Science Foundation (NSF) has launched the PAWR program, which deploys wireless testbeds across various states in the United States. Each testbed covers a specific region and focuses on use cases aligned with the available spectrum, ensuring compliance with spectrum regulations. For this report, we collected CSI data from the ARA testbed located in Ames, Iowa. With assistance from the ARA team [1], we followed the instructions outlined in their user manual [30] to perform CSI measurements. We obtained three datasets describing wireless communication in agricultural fields, where sensors collect data and transmit it to a base station connected to a central office for processing and analysis.

The CSI dataset is stored in a file named $ue_{uplink} * * * .data$, which contains five subkeys: *bin*, *ch*, *data*, *frame*, and *hdl*. The *hdl* key represents the handle, indicating the user equipment (UE) index. The *ch* key refers to the wireless channel assigned to each UE, with two channels per user (indexed as 0 and 1). The *frame* key denotes the frame number, while *bin* specifies the frequency offset from the center frequency. The *data* subkey contains the actual normalized *IQ* values for each UE, channel, and frame.

The ARA base station (BS) is equipped with 42 operational antennas. Uplink pilot signals from each UE are received by all BS antennas. Of the 512 values in the CSI data, the first 4 correspond to two reference antennas. Each reference antenna has both In-phase (I) and Quadrature (Q) components. The next 84 values represent the I and Q components for the 42 BS antennas (42 antennas × 2 channels). The remaining values, from rows 85 to 512, are zeros, as the system supports up to 256 antennas, but only 42 are operational. This results in zeros for the unused 214 antenna entries (214 antennas × 2 channels = 428 entries).

All UEs in the collected CSI data are stationary sensors with a Line-of-Sight (LoS) connection to one sector of the three-sector base station antenna system. The datasets thus provide a valuable resource for studying CSI in rural agricultural environments with stationary UEs.

Simulated CSI Dataset: Due to the lack of datasets that comprehensively represent diverse propagation scenarios, we employ the WINNER II Toolbox, which utilizes the Stochastic Channel Model (SCM) for generating channel coefficient matrices (i.e., CSI). The toolbox is used to generate multiple datasets corresponding to various propagation scenarios by configuring the following components:

- The antenna patterns are configured with azimuth angles set to 45 and −45, respectively. Antenna geometries are selected from among Uniform Linear Arrays (ULA), Uniform Planar Arrays (UPA), and Uniform Circular Arrays (UCA). A dipole field pattern is specified for each antenna configuration.
- The system is configured by defining key parameters such as the number of users, their mobility patterns, the number of radio links, and the propagation scenarios. Additional parameters included the center frequency, deployment scenarios (e.g., urban, dense urban, or rural), and the Line-of-Sight (LoS) or Non-Line-of-Sight (NLoS) conditions.
- WINNER II parameters are meticulously set to simulate the wireless channel for each scenario, including large-scale fading, spatial correlations, and multipath components.
- The transmitted signal is defined to traverse through the modeled WINNER II wireless channel with Additive White Gaussian Noise (AWGN) model to model the noise figure.

4.2 SCM model validation

Figure 2 illustrates the performance of the WINNER II model in generating synthetic CSI data that simulates the CSI datasets obtained from the ARA testbed. The performance is measured versus the ARA testbed datasets. The resulting Mean Squared Error (MSE) is 2.46 for both components of the CSI datasets.

4.3 GAN model validation

Although simulated datasets are used for GAN training, we aim to validate the GAN against realistic datasets rather than the simulated ones to assess its effectiveness in generating synthetic data.

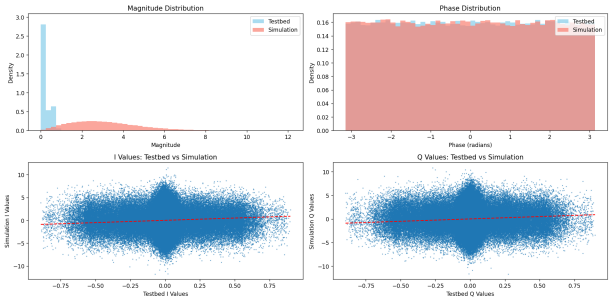

Figure 2: WINNER II model validation

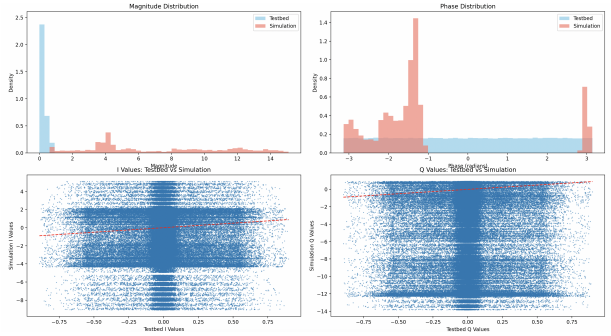

Figure 3: GAN model validation

Figure 3 illustrates the performance of the GAN model in generating synthetic CSI data that using the simulated CSI datasets obtained from the WINNER II model. The performance is measured versus the ARA testbed datasets. The resulting Mean Squared Error (MSE) are 3.1301 and 7.5083, for CSI's real and imaginary terms, respectively.

5 DISCUSSION

The obtained preliminary results highlight the influence of training data on GAN performance. As shown in Figure 2, while the MSE is reasonably acceptable compared to the realistic testbed data, the impact on GAN performance remains significant. Additionally, it is important to note that our results are based on data collected from a single scenario. Incorporating data from multiple simulated scenarios could provide deeper insights into the GAN's performance.

Our plan is to leverage the findings to support policymakers and urban planners in allocating wireless infrastructure, determining base station placements, and optimizing network configurations to advance urban development in these regions. However, the current preliminary results are insufficient to provide insights for policymakers into using the GAN model for wireless planning in urban settings. These preliminary results underscore the importance of utilizing balanced training data, but, at this stage of research, we can not validate the use of the GAN model for these urban planning applications.

6 CONCLUSION

In this report, we address the challenge of generating synthetic Channel State Information (CSI) for network slicing resource planning in urban settings, given the lack of realistic CSI datasets. We explore three major categories of synthetic CSI generation methods: stochastic geometry-based models, statistically correlated models, and machine learning-based generative methods. Among these, we adopt Generative Adversarial Networks (GANs) since it has demonstrated relatively high accuracy for this purpose. For this preliminary study, we start with adopting the baseline version of the GAN model, i.e., Vanilla GAN, using the stochastic channel model WINNER II to generate training datasets. The results highlight the impact of using simulated datasets to train the GAN, with the SCM-simulated data achieving better Mean Squared Error (MSE) performance than the GAN-generated data during validation.

Implications and learned lessons: Our takeaways from this project lie in the following:

(1) Training GANs solely with simulated datasets is **NOT** recommended, as it may not yield accurate or reliable results for practical applications;
(2) Employing a diverse range of generative models is essential to improve robustness and adaptability to different scenarios;
(3) Incorporating a variety of performance metrics is crucial for a comprehensive evaluation of model effectiveness and accuracy;
(4) Given the availability of a testbed, conducting experiments under operational scenarios is strongly encouraged to validate and refine the model's performance in real-world conditions

Limitations:

(1) This study focused solely on the baseline GAN model (Vanilla GAN) without exploring other advanced GAN variants discussed in the literature;
(2) Logistical challenges and a lack of manpower hindered the ability to conduct in-field experiments, particularly under mobility scenarios, to obtain more realistic datasets;
(3) The performance metrics used were insufficient to comprehensively evaluate the effectiveness of the proposed model;
(4) The coarse-grained SCM-simulated data generation revealed discrepancies in distribution compared to real-field configurations, highlighting the need to explore correlated statistical methods for generating more representative training data since we study the statistical properties of the CSI datasets that would facilitate using statistically correlated model;

Future work: Our ongoing work focuses on refining the SCM to minimize noise in SCM-based simulated datasets used for training the GAN. We also plan to explore additional machine learning models and experiment with various GAN architectures to conduct comparative analyses.

ACKNOWLEDGMENTS

Acknowledgements. Here, we acknowledge external collaborators and people who provided you access to data or contributed domain expertise (ARA testbed team) Islam Taimor, Sarah, Mohamed Solieman, and Their Director Dr. Hongwei Zhang.

AUTHOR CONTRIBUTIONS

The following contributions were made by the team members:

- **Management Tasks**
  - Brainstorming and participation in weekly team meetings and recording of meeting minutes.
  - Setting the work plan and organizing tasks.
- **Technical Tasks**
  - Conducted a literature review by Taeesh and Lamice.
  - Dataset collection and data development by Lamice and Emad.
  - Implemented the MATLAB WINNER II model by Ahmed and Emad.
  - Generated simulated data for analysis by Ahmed and Taeesh.
  - Developed the GAN model for contextual space by Ahmad, Lamice and Taeesh.
  - Implemented the stochastic contextual bandit algorithm by Ahmed and Emad.
  - Collected and visualized the results for evaluation by Ahmed and Taeesh.
- **Writing Tasks**
  - **Illustrative Figures and Tables:** Created background figures, literature review diagrams, system model illustrations, parameter tables, dataset components, and relevant visualizations by Emad and Taeesh.
  - **Introduction:**
    - Drafted by Emad.
    - Reviewed and polished by Ahmed, Taeesh and Lamice.
    - Proofread and improved by Emad, Ahmed, Taeesh and Lamice.
  - **Related Work:**
    - Written by Taeesh and Lamice.
  - **Methodology:**
    - Drafted by Emad.
  - **Results:**
    - Prepared by Ahmed, Taeesh, Emad and Lamice.
  - **Discussion:**
    - All team members provided insight into the results.
  - **Conclusion:**
    - Drafted by Ahmed and Emad.
    - Proofread and improved by Taeesh and Lamice.

The code and datasets used for synthetic CSI generation can be found in [3, 4].

REFERENCES

[1] [n. d.]. ARA User Manual — ARA User Manual 0.1 documentation. https://arawireless.readthedocs.io/en/latest/index.html

[2] 3GPP. 2024. *Open RAN.* https://www.3gpp.org/news-events/3gpp-news/open-ran Accessed: December 14, 2024.

[3] Ahmed Aredah. [n. d.]. Synthetic CSI Data Repository. https://github.com/AhmedAredah/syntheticCSIData. Accessed: 2024-12-17.

[4] Taeesh Azal Assadi. [n. d.]. Synthetic CSI Data Generation Source Code (MATLAB). https://github.com/taeesh1309/syntheticCSIData/blob/main/src/generation_updated.m. Accessed: 2024-12-17.

[5] Emmanual N Barron. 2024. *Game theory: an introduction.* John Wiley & Sons.

[6] Brenda Vilas Boas, W Zirwas, and M Haardt. 2022. Machine Learning for CSI Recreation in the Digital Twin Based on Prior Knowledge. *IEEE Open Journal of the Communications Society* 3 (2022), 1578–1591. https://doi.org/10.1109/OJCOMS.2022.3208323

[7] Ann Borda and Jonathan P. Bowen. 2017. Smart Cities and Cultural Heritage – A Review of Developments and Future Opportunities. BCS Learning & Development. https://doi.org/10.14236/ewic/eva2017.2

[8] Daoud Burghal, Yang Li, Pranav Madadi, Yeqing Hu, Jeon-Hoon Jeon, Joonyoung Cho, A Molisch, and Jianzhong Zhang. 2023. Enhanced AI-Based CSI Prediction Solutions for Massive MIMO in 5G and 6G Systems. *IEEE Access* 11 (2023), 117810–117825. https://doi.org/10.1109/ACCESS.2023.3324399

[9] Sheryl Chang, Bryan Y. Lim, Xuefeng Jiang, Emilio Zagheni, and Samuel V. Scarpino. 2021. Supporting COVID-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).* https://www.medrxiv.org/content/10.1101/2021.03.20.21254022v1.full.pdf Accessed: 2024-12-02.

[10] Ben Earle, A Al-Habashna, Gabriel A Wainer, Xingliang Li, and Guoqiang Xue. 2021. Prediction of 5G New Radio Wireless Channel Path Gains and Delays Using Machine Learning and CSI Feedback. *2021 Annual Modeling and Simulation Conference (ANNSIM)* (2021), 1–11. https://doi.org/10.23919/ANNSIM52504.2021.9552072

[11] Ericsson. [n. d.]. *Ericsson Mobility Report Business Review 2024.* Technical Report.

[12] Rajeev Gangula Sakthivel Velumani Davide Villa Leonardo Bonati Michele Polese Gabriel Arrobo Christian Maciocco Tommaso Melodia Hai Cheng, Salvatore D'Oro. 2024. ORANSlice: An Open-Source 5G Network Slicing Platform for O-RAN. *arXiv preprint* 2410.12978v1 (2024). https://arxiv.org/html/2410.12978v1 Accessed: December 14, 2024.

[13] Ravi Hosamani and Yerriswamy T. 2022. Deep Learning-Based CSI Estimation Using Synthetic Dataset. *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)* (2022), 1–5. https://doi.org/10.1109/icdcece53908.2022.9792900

[14] Abdullah Hossain and Nirwan Ansari. 2023. 5G Multi-Band Numerology-Based TDD RAN Slicing for Throughput and Latency Sensitive Services. *IEEE Transactions on Mobile Computing* 22, 3 (3 2023), 1263–1274. https://doi.org/10.1109/TMC.2021.3106323

[15] Ben Hughes, Shruti Bothe, H Farooq, and A Imran. 2019. Generative Adversarial Learning for Machine Learning empowered Self Organizing 5G Networks. *2019 International Conference on Computing, Networking and Communications (ICNC)* (2019), 282–286. https://doi.org/10.1109/ICCNC.2019.8685527

[16] Mayuko Inoue and Tomoaki Ohtsuki. 2024. Source CSI Dataset for Multi-Task CSI Feedback. In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC).* IEEE, 1042–1043.

[17] Mayuko Inoue, Tomoaki Ohtsuki, Kohei Yamamoto, and Guan Gui. 2023. Evaluation of source data selection for DTL based CSI feedback method in FDD massive MIMO systems. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC).* IEEE, 182–187.

[18] Muhammad Nur Aqmal Khatiman, Asma' Abu-Samah, Muhammad Amin Azman, R Nordin, and N Abdullah. 2023. Generation of Synthetic 5G Network Dataset Using Generative Adversarial Network (GAN). *2023 IEEE 16th Malaysia International Conference on Communication (MICC)* (2023), 141–145. https://doi.org/10.1109/MICC59384.2023.10419563

[19] Pekka Kyösti and Xiongwen Zhao. 2008. *WINNER II channel models.* Technical Report. https://www.researchgate.net/publication/234055761

[20] Michal Lom, Ondrej Pribyl, and Miroslav Svitek. 2016. Industry 4.0 as a part of smart cities. In *2016 Smart Cities Symposium Prague (SCSP).* 1–6.

[21] Yuanqiu Luo, Ming Jiang, Dezhi Zhang, and Frank Effenberger. 2023. Field Trial of Network Slicing in 5G and PON-Enabled Industrial Networks. *Wireless Commun.* 30, 1 (Feb. 2023), 78–85. https://doi.org/10.1109/MWC.002.2200215

[22] R Uma Mageswari, Gousebaigmohammad, Devee siva prasad Dulam, S Shitharth, G Surya Narayana, A Suresh, JaikumarR, Leena Bojaraj, S Chandragandhi, and Amsalu Gosuadigo. 2022. Machine Learning Empowered Accurate CSI Prediction for Large-Scale 5G Networks. *Wireless Communications and Mobile Computing* (2022). https://doi.org/10.1155/2022/7085731

[23] O-RAN Alliance. 2024. *O-RAN Alliance: Towards Open and Intelligent RAN.* https://www.o-ran.org/ Accessed: December 14, 2024.

[24] Hanli Peng. 2024. Attention-based deep learning approach for CSI feedback under 5G TDL channel. *Journal of Physics: Conference Series* 2711 (2024). https://doi.org/10.1088/1742-6596/2711/1/012001

[25] Alexandre Matos Pessoa, Bruno Sokal, Carlos FM E Silva, Tarcisio Ferreira Maciel, Andre LF De Almeida, and Francisco Rodrigo Porto Cavalcanti. 2020. A CDL-based channel model with dual-polarized antennas for 5G MIMO systems in rural remote areas. *IEEE Access* 8 (2020), 163366–163379.

[26] Md Ileas Pramanik, Raymond Y.K. Lau, Haluk Demirkan, and Md. Abul Kalam Azad. 2017. Smart health: Big data enabled health paradigm within smart cities. *Expert Systems with Applications* 87 (2017), 370–383. https://doi.org/10.1016/j.eswa.2017.06.027

[27] U F Siddiqi, S M Sait, and K Al-Utaibi. 2021. A Machine Learning Method to Synthesize Channel State Information Data in Millimeter Wave Networks. *IEEE Access* 9 (2021), 83441–83452. https://doi.org/10.1109/ACCESS.2021.3087630

[28] K C Sriharipriya, J C Clement, Gerardine Immaculate Mary, Chandrasekharan Natraj, R Tharun Kumar, and R Gokul. 2024. Enhanced synthetic generation of channel state information for millimeter-wave networks in 5G communication

systems. *Internet Technology Letters* (2024). https://doi.org/10.1002/itl2.577

[29] Alexander V Stenin and A Kalachikov. 2022. Numerical Evaluation of the Channel Estimation in 5G NR Based on Machine Learning. *2022 IEEE 23rd International Conference of Young Professionals in Electron Devices and Materials (EDM)* (2022), 285–288. https://doi.org/10.1109/EDM55285.2022.9855055

[30] Ara Wireless Team. 2024. *Ara Wireless User Manual: AraRan Experiment - Skylark CSI.* https://arawireless.readthedocs.io/en/latest/ara_experiments/araran_experiments/AraMIMO_CSI.html#araran-experiment-skylark-csi Accessed: 2024-12-14.

[31] Min Wang, Yaping He, Haiming Wang, Cheng Xiang Wang, and Xiaohu You. 2024. A Pervasively Correlated Channel Model for Massive MIMO Transmission. *IEEE Transactions on Communications* 72, 4 (4 2024), 2441–2456. https://doi.org/10.1109/TCOMM.2023.3343386

[32] Shangbin Wu, Cheng-Xiang Wang, el-Hadi M Aggoune, and Mohammed M Alwakeel. [n. d.]. *A Novel Kronecker-Based Stochastic Model for Massive MIMO Channels.*

[33] F Xiao, Xiaohui Xie, Zhetao Li, Qingyong Deng, Anfeng Liu, and Lijuan Sun. 2018. Wireless Network Optimization via Physical Layer Information for Smart Cities. *IEEE Network* 32 (2018), 88–93. https://doi.org/10.1109/MNET.2018.1700281

[34] Han Xiao, Wenqiang Tian, Wendong Liu, and Jia Shen. 2022. ChannelGAN: Deep learning-based channel modeling and generating. *IEEE Wireless Communications Letters* 11, 3 (2022), 650–654.

[35] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology* 5, 3 (9 2014). https://doi.org/10.1145/2629592

[36] Pan Zhou, Jie Xu, Wei Wang, Changkun Jiang, Kehao Wang, and Jia Hu. 2020. Human-Behavior and QoE-Aware Dynamic Channel Allocation for 5G Networks: A Latent Contextual Bandit Learning Approach. *IEEE Transactions on Cognitive Communications and Networking* 6 (2020), 436–451. https://doi.org/10.1109/TCCN.2020.2969631

[37] Junior Momo Ziazet, Brigitte Jaumard, H. Duong, P. Khoshabi, and Emil Janulewicz. 2022. A Dynamic Traffic Generator for Elastic 5G Network Slicing. In *2022 IEEE International Symposium on Measurements and Networking, M and N 2022 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MN55117.2022.9887734

[38] Cong Zou, Fang Yang, Jian Song, and Zhu Han. 2024. Generative Adversarial Network for Wireless Communication: Principle, Application, and Trends. *IEEE Communications Magazine* 62, 5 (2024), 58–64.

Received 17 December 2024; revised 01 February 2025; accepted xx YYYY 2025