Toward Addressing Training Data Scarcity Challenge in Emerging Radio Access Networks: A Survey and Framework

Haneya Naeem Qureshi, Usama Masood, *Graduate Student Member, IEEE*, Marvin Manalastas, *Graduate Student Member, IEEE*, Syed Muhammad Asad Zaidi, Hasan Farooq, Julien Forgeat, Maxime Bouton, Shruti Bothe, Per Karlsson, Ali Rizwan, and Ali Imran, *Senior Member, IEEE*

***Abstract*—The future of cellular networks is contingent on artificial intelligence (AI) based automation, particularly for radio access network (RAN) operation, optimization, and troubleshooting. To achieve such zero-touch automation, a myriad of AI-based solutions are being proposed in literature to leverage AI for modeling and optimizing network behavior to achieve the zero-touch automation goal. However, to work reliably, AI based automation, requires a deluge of training data. Consequently, the success of the proposed AI solutions is limited by a fundamental challenge faced by cellular network research community: scarcity of the training data. In this paper, we present an extensive review of classic and emerging techniques to address this challenge. We first identify the common data types in RAN and their known use-cases. We then present a taxonomized survey of techniques used in literature to address training data scarcity for various data types. This is followed by a framework to address the training data scarcity. The proposed framework builds on available information and combination of techniques including interpolation, domain-knowledge based, generative adversarial neural networks, transfer learning, autoencoders, few-shot learning, simulators and testbeds. Potential new techniques to enrich scarce data in cellular networks are also proposed, such as by matrix completion theory, and domain knowledge-based techniques leveraging different types of network geometries and network parameters. In addition, an overview of state-of-the art simulators and testbeds is also presented to make readers aware of current and emerging platforms to access real data in order to overcome the data scarcity challenge. The extensive survey of training data scarcity addressing techniques combined with proposed framework to select a suitable technique for given type of data, can assist researchers and network operators in choosing the appropriate methods to overcome the data scarcity challenge in leveraging AI to radio access network automation.**

***Index Terms*—Scarce data, training data, big data, emerging cellular networks, RAN, machine learning, synthetic data generation, interpolation, simulators, testbeds.**

Manuscript received 24 December 2021; revised 15 August 2022 and 4 February 2023; accepted 22 March 2023. Date of publication 1 May 2023; date of current version 23 August 2023. This work was supported in part by the National Science Foundation under Grant 1923669 and Grant 1730650; in part by the Qatar National Research Fund (QNRF) under Grant NPRP12-S 0311-190302; and in part by the Unrestricted Award from Ericsson Research, CA, USA. *(Corresponding author: Haneya Naeem Qureshi.)*

Haneya Naeem Qureshi, Usama Masood, Marvin Manalastas, and Syed Muhammad Asad Zaidi are with the AI4Networks Research Center, School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK 74135 USA (e-mail: haneya@ou.edu).

Hasan Farooq, Julien Forgeat, Maxime Bouton, Shruti Bothe, and Per Karlsson are with the Ericsson Research Department, Ericsson, Santa Clara, CA 95054 USA.

Ali Rizwan is with the Department of Electrical Engineering, Qatar University, Doha, Qatar.

Ali Imran is with the AI4Networks Research Center, School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK 74135 USA, and also with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K.

Digital Object Identifier 10.1109/COMST.2023.3271419

I. Introduction

FUTURE cellular networks are envisioned to have big data enabled network automation capabilities [1]. This includes functionalities of self-optimization, self-healing and self-configuration [2], [3] that are essential to ensure the viability and sustainability of future cellular networks amid challenges, such as amalgam of new technologies, growing complexity, resource inefficiency and shrinking profit margins. In order to enable these automation capabilities in next generation cellular networks, the process of heterogeneous base station (BS) deployment, implementing existing and newly proposed network features and tuning the associated network parameters has to be meticulous. This is because the process of selecting an optimal network configuration that can maximize the vital key performance indicators, like coverage, capacity, reliability or energy efficiency is a rather challenging task. Identifying the optimal network configuration is necessary for network operators to fulfill the promises made by much anticipated 5G and beyond networks and to realize the efficacy of several new use cases.

Research community heavily rely on mathematical yet tractable analytical models [4], [5], [6], [7], [8], [9] to propose planning, operation and optimization of different aspects of network. They, however, are based on restrictive assumptions and simplifications with respect to transceiver architecture, base station and user distributions and propagation characteristics, to name a few. Furthermore, stochastic geometry-based models are unable to capture the network dynamics which include mobility management and transmission latency. Therefore, several machine learning (ML) based techniques are proposed in current literature that leverage training and tuning of ML based models to determine the

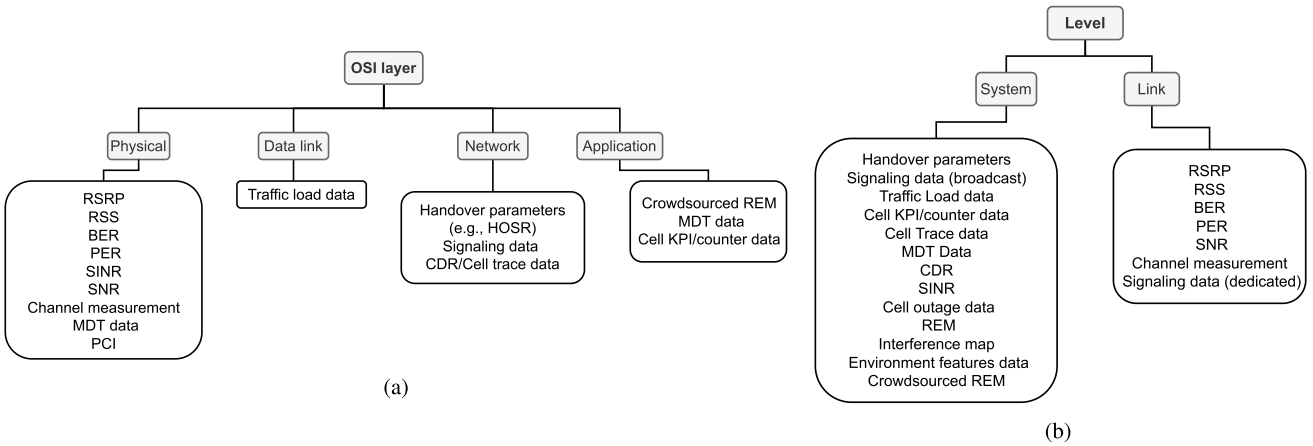


Fig. 1. Types of data on which techniques to address data scarcity have been applied in literature according to (a) OSI-layer based categories (b) system/link level categories (figure is based on Table IV).

behavior of different configuration and optimization parameters (COPs), such as antenna tilt, transmit power, cell load in relation to different key performance indicators (KPIs), like coverage, capacity or energy efficiency [10], [11], [12]. These COP-KPI relationships can then be used for COP-KPI optimization. Moreover, in cellular networks context, awareness about radio environment in a wireless system is crucial given that the radio spectrum is a limited resource [13]. Ample data is required for constructing radio environment maps (REMs) which can be used for operations such as spectrum management, to construct interference maps, to make decisions about spectrum availability for enabling dynamic spectrum access, for assessing/monitoring network health, minimizing signalling, interference management, optimization of radio resources allocation, dynamic spectrum allocation, identify bad-signal areas, automatic neighbor relation, minimize drive tests, handovers optimization and coexistence of various technologies [14], [15]. However, all such techniques face a common key challenge that undermine their utility: scarcity/sparsity of the training data. This fundamental problem has two facets: (i) Data scarcity: Obtaining large amounts of pertinent training data from the operators is not a trivial task. Furthermore, as most of the data remain trapped in silos, even if willing, a single operator may not be able to provide the deluge of real data needed for developing models, e.g., user (traffic, mobility pattern, QoE expectations) and network behavior (spatio-temporally robust COP-KPI) models. (ii) Data sparsity: Network operators only try a limited range of COPs in live networks due to high probability of significant network performance impairment of live mobile network during the trial phase. Therefore, only a limited range of COP-KPI data can be obtained. Given that operators only try a limited range of COPs in live networks, despite sourcing from multiple operators, even when not scarce, the real data are expected to be sparse or unevenly distributed. In other words, term scarcity refers to problem when data is too little to train a model. Sparsity on the other hand refers to problem when there is some data, but it is thinly or unevenly distributed making reliable training of AI difficult. For sake of clarity, in rest of the manuscript we use only one term, scarcity to represent this problem irrespective of the reason behind data being not enough to train AI.

To illustrate the type of data in cellular networks which is scarce, Fig. 1 shows the data on which data augmentation techniques have been applied in literature according to OSI layers and system/link level categorization. Link level data corresponds to the point-to-point communication link, for example RSRP, and system level data takes the notion of data involving a large number of network elements including several links, for example REM. The use cases of these data are elaborate in later sections and are summarized in last two columns of Table IV.

To address the data scarcity challenge, one solution can be to obtain data from field trials. However, conducting independent field trials on a large scale is costly and time-consuming, especially in dynamic scenarios, where the number and locations of measurements change, and it is infeasible to measure the radio frequency field strength values at every point of interest. Another way to obtain data is through mathematical models. However, they are based on too many assumptions and simplifications, that fail to depict real world scenarios. Moreover, in ultra-dense deployments, small cells contain far fewer users compared to macro cells. This makes user measurements at the base station of small cells scarce, which particularly poses a problem for automation solutions that leverage minimization of drive test (MDT) [16], [17], [18]. This problem is further aggravated if smaller bin size is used to reduce quantization error, attributing to the fact that many bins might not be visited by even a single user during the reporting period [18].

Deploying the new 5G and beyond network functionalities in a real world cannot be done arbitrarily. If the training data is poorly distributed or scarce, it might not represent the actual network scenario very well, which could lead to overfitting during the model training stage. In order to develop accurate models, machine learning algorithms require large amounts of true training data since a model based on scarce data would rely on assumptions and weak correlations [19]. In turn, unscrupulous network design and sub-optimal parameter configuration will hamper not only the capability of future networks that will impact the user experience negatively but

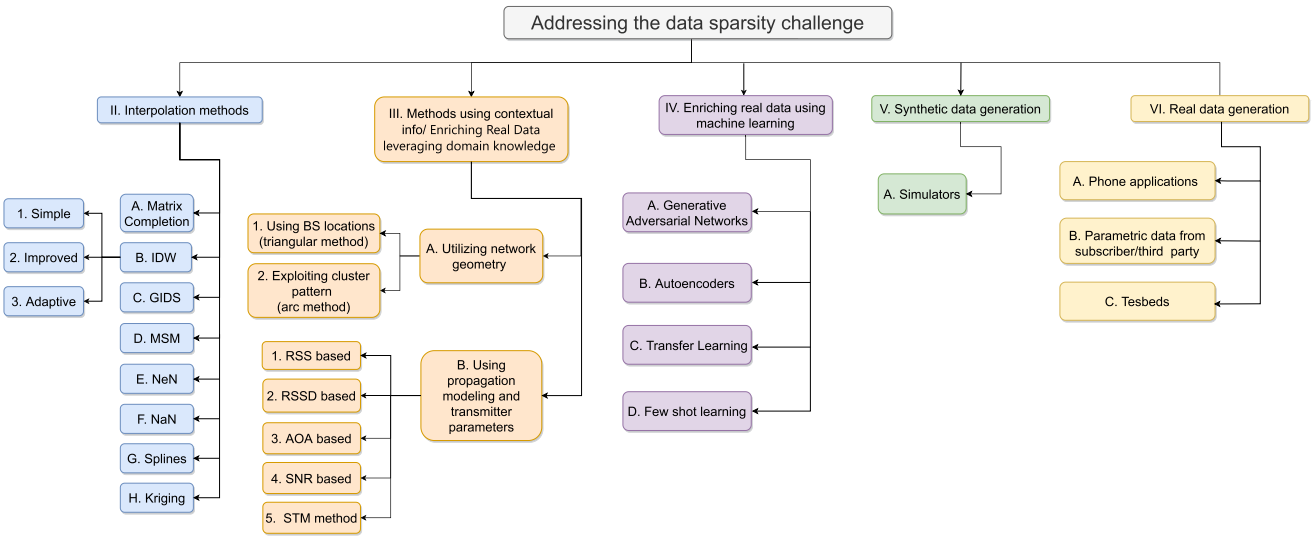


Fig. 2. This figure presents one possible taxonomy for classifying the techniques to address data scarcity in RAN.

will also increment the capital and operational expenditure (CAPEX/OPEX) of mobile operators [20].

A. Related Work

Data scarcity challenge has been addressed in the domain of environment sciences field, such as ecology, marine, agriculture, soil science, elevation, precipitation, and chemical concentrations, through review papers in [21], [22], [23], [24]. However, to the best of authors' knowledge, a survey paper on addressing the training data scarcity challenge in cellular networks is not present.

In cellular networks context, the closest survey papers to this work are [25], [26] and [27]. Authors in [25] focus on the task of radio environment map (REM) construction techniques. Advantages, disadvantages, and asymptotic complexity comparison of seven interpolation techniques (inverse distance weighted, nearest neighbor, spline, natural neighbor, modified Shepard's method, gradient plus inverse distance squared method and Kriging). They also discuss some indirect construction methods that combine interpolation with transmitter parameter information. However, since work in [25] is from 2014, many indirect methods developed after 2014 are not covered in it. Moreover, [25] is limited to the task of REM construction only. Several methods that have gained popularity in past recent years to enrich scarce data, like advanced machine learning techniques and synthetic data generation, that are a part of this survey, are also not included in [25].

The other relevant study to this work is the study in [26], where authors survey the use of interference maps. However, the study in [26] focuses on spectrum occupancy measurement data only while reviewing studies till 2016. In contrast, in this survey, we cover variety of RAN data. Like [25], popular methods in recent years to augment scarce data, like advanced machine learning techniques and synthetic data generation are also not included in [26] as addressing data scarcity problem is not the focus of the work in [26].

Simulators are another promising way to address the data scarcity challenge. Two existing surveys on simulators include [28] and [29]. Authors in [28] compare 4G and 5G simulators and authors in [29] provide a summary of the most significant 5G simulators. However, these works are restricted only to simulators as a tool for generating data.

Testbeds can also be used to generate real data to augment available scarce data. The work in [30] compares key testbeds around the world in terms of location, scale of deployment, type of access, key features, and supported experiments. However, these works are restricted to testbeds only, whereas this survey aims to address data sparsity challenge by considering additional techniques as identified in Fig. 2.

A more recent study from 2019 [27] surveyed the applications of deep learning-based techniques, like transfer learning, autoencoders, generative adversarial networks techniques for wireless networks. The authors introduce the basics of deep learning and then identify wireless applications where those techniques can be used, for instance, mobile data analysis, mobility analysis, wireless sensor network, network control, network security, signal processing, and other emerging wireless applications. While some of the techniques discussed in [27] can also be exploited to address data scarcity challenge in RANs to some extent for limited data types, the work in [27] is not focused on addressing the training data scarcity challenge in RAN. In contrast, this survey not only provides a comprehensive review of techniques that can address training data sparsity for a variety of RAN data but also it provides the first of its kind systematic framework to select the most suitable techniques for given data types.

To the best of authors' knowledge, there is no existing work that presents a consolidated survey and framework that aims to solve the training data scarcity challenge in cellular networks. This article presents the techniques in literature to address the training data scarcity problem over the period of 1991 to 2021 as they apply to radio access networks in wireless communications.

B. Contributions and Organization

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

- To address the training data scarcity challenge, we present an overview of existing techniques, and potential new and emerging techniques, such as matrix completion theory (Section II-A) leveraging different types of network geometries (Section III-A), and advanced machine learning techniques such as the use of generative adversarial networks (GANs) (Section IV-A), autoencoders (Section IV-B), transfer learning (Section IV-C) and few-shot learning (Section IV-D) to enrich scarce data in cellular networks. We also highlight the pros and cons of these approaches analyzed in context of different RAN focused use cases. A taxonomy of training data enrichment techniques is developed by grouping these techniques into various categories as shown in Fig. 2.
- We present a comparison of existing and emerging simulators (Section V-A) as tools for generating synthetic data to overcome the data scarcity issue which can greatly benefit researchers as the characterization and comparison among features of different simulators will enable them to identify publicly accessible simulators and use them for their specific problems.
- An overview of state-of-the-art current and emerging testbeds for next generation cellular networks is presented in Section VI-B that will make readers aware of current and emerging platforms to access real data in order to overcome data scarcity challenge. Most of these testbeds are available to external experiments, which will foster collaboration among different academic institutions as well as with industry. This will in turn enable the utilization of these existing facilities to the fullest and accelerate quality research in the field of cellular networks.
- We propose a decision tree diagram, that will enable researchers and operators to choose appropriate methods to solve the training data scarcity challenge, based on the available information and network scenario.

It should be noted that measured data can be scarce and still be representative. On the other hand, data can be big but not representative. We begin by presenting an overview of techniques that will work best in the first case. In the case when data is scarce and representative, but the only information known are the measured data points and their location, interpolation methods in Section II are likely to perform best.

Moving forward, when some additional information beyond the data points and their locations is known, we can utilize the methods using contextual information or domain knowledge in Section III. Several machine learning techniques can also be leveraged to address the data scarcity challenge. These include generative adversarial networks, autoencoders, transfer learning and few-shot learning techniques (Section IV).

On the contrary, when the available data is big and non-representative or scarce and non-representative, the solution lies in either resorting to generate synthetic data (Section V) or get real data (Section VI). In addition, for scenarios with no starting real data, for example, for new or anticipated scenarios which are not yet deployed in a real network, simulators, and testbeds to generate real data are most likely going to be the best option for wireless communications community.

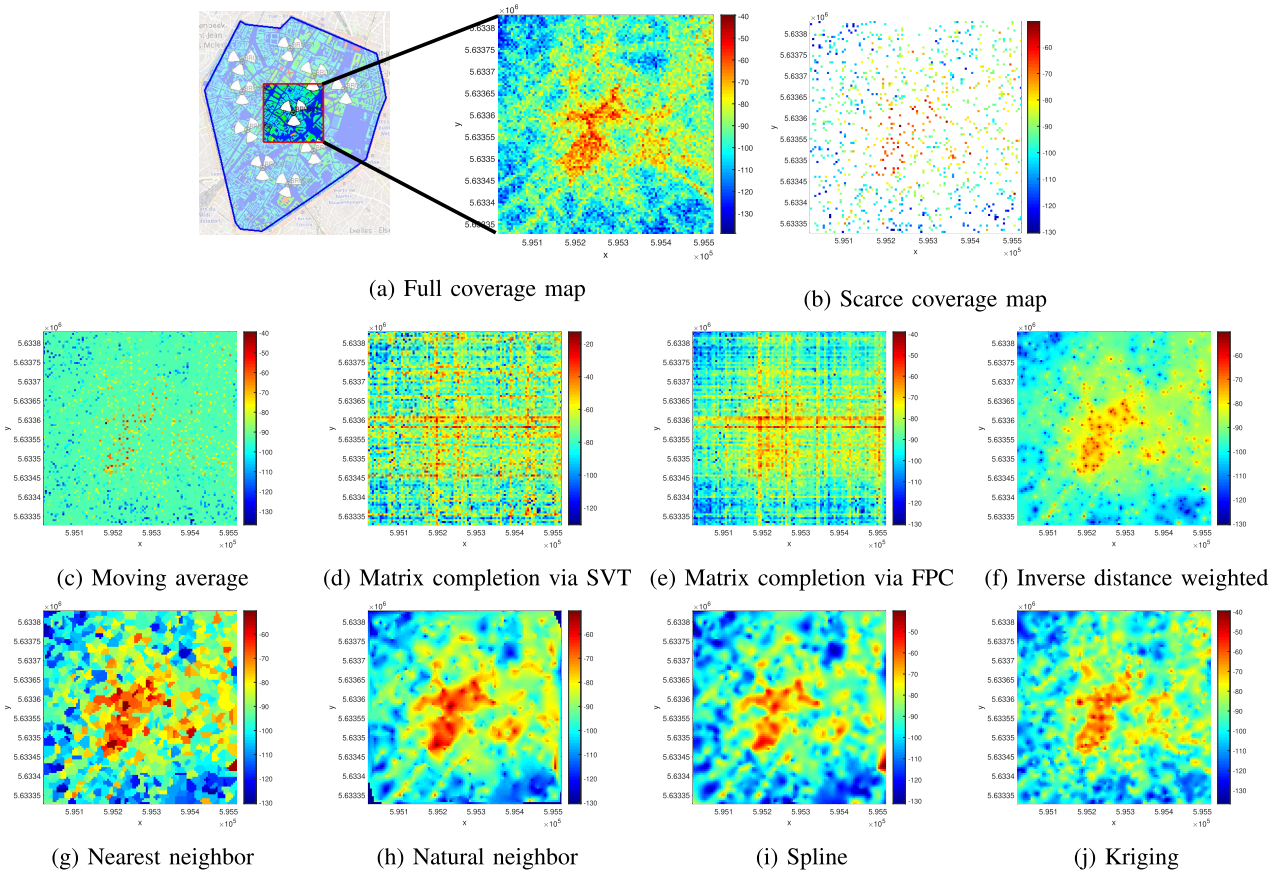
Other classifications of data augmentation techniques, such as those based on OSI layer based, or system and link level grouping of the data streams are also possible. However, many data scarcity techniques can be applied to the data corresponding to multiple layers and levels. Therefore, the rest of the paper is structured by organizing the techniques based on their technical grouping as shown in this tree diagram. i.e., each branch represents a section, and each leaf represents a subsection of the paper. Moreover, while it is intuitive to assume that data from different layers may require different generation techniques, but the suitability of a technique depends mainly on the characteristics of the data, e.g., availability of latent distribution, completeness, representativeness, temporal or spatial nature and context and so on. For example, traffic variation at base station data at the application level can be modelled as time series data, and same can be done for the packet error data at link level, and bit error data at physical layer. Similarly, data on traffic variation in space (system level data) bears similarity with, for instance, RSRP/SINR-based REM data (physical layer) and thus same techniques such as kriging, inverse distance weighted, nearest neighbor interpolation can be used. While in most cases, the characteristics and contexts of the data may suffice to choose the best technique, in some cases, additional knowledge that can be extracted from knowing which layer the data belongs to may be helpful in improving the data augmentation. However, so far in literature there does not exist examples of where knowledge of layer level mapping is exploited for data augmentation.

II. Interpolation Methods

When the only information required from cellular network are the measurement values (location-value pair) in order to recover the missing values, we classify such methods as 'interpolation methods', which assume that the data are spatially dependent and continuous over space [31], [32], [33].

Interpolation methods are widely used in literature for radio environment map (REM) augmentation. REM for a coverage area consists of radio information, such as signal strength, signal quality or interference [25]. Constructing REMs is done through manual drive tests, which leads to collection of data from scarce locations due to time and cost constraints. REM supports a variety of use cases, such as spectrum access management, identification of poor signal areas, automatic neighbor relation, power management, interference mitigation and management, optimization of radio resources allocation, radio resource management, dynamic spectrum allocation, handovers optimization, automated networks planning, maintenance and optimization of network parameters [25]. Therefore, complete REMs from the available scarce REMs are required to support these use cases.

Another type of widely used data on which interpolation techniques are applied is the minimization of drive test (MDT) data [34]. 3GPP has standardized MDT that allows network performance estimation at a base station by leveraging measurement reports gathered at the user equipment (UE) without the need for drive tests [35]. The MDT reports contain network coverage related performance indicators (such as



(a) Full coverage map (b) Scarce coverage map

(c) Moving average (d) Matrix completion via SVT (e) Matrix completion via FPC (f) Inverse distance weighted

(g) Nearest neighbor (h) Natural neighbor (i) Spline (j) Kriging

Fig. 3. Comparison of coverage map reconstruction techniques [34].

RSRP) measured at the UE. These reports are tagged with UEs' geographical location information and sent to their serving base stations [18]. MDT data can be scarce in areas of low user density, which will lead to inaccurate or sub-optimal coverage estimation models [34]. To address this problem, authors in [34] applied several interpolation algorithms, including the ones discussed in this section. Their results are illustrated in Fig. 3 and will be discussed further in the subsection pertaining to data enrichment technique used in each of the subfigures.

Different interpolation techniques can be applied in the cellular network context to address the data scarcity challenge. Each technique has its own set of advantages and disadvantages; we elaborate these techniques in this section.

A. Matrix Completion Theory

A recent work [34] applied matrix completion theory to cellular network data context. Assuming the coverage area is divided into bins, a coverage matrix $\boldsymbol{C}$ containing coverage indicator (such as RSRP measurements) is observed. A scheme that jointly exploits matrix factorization theory and convex optimization is used to recover the missing data in $\boldsymbol{C}$ [34].

This leads to the following optimization problem in order to find the missing values in matrix $\boldsymbol{C}$:

$$\begin{aligned} &\text{minimize} \quad \text{rank}\{\boldsymbol{P}\} \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{1}$$

where $\boldsymbol{P}$ is the decision variable in the optimization problem, the pair (i, j) denotes the i-th row and j-th column of the matrices $\boldsymbol{C}$ and $\boldsymbol{p}$ and Ψ is the set of locations corresponding to the observed entries ($(i,j) \in \Psi$ if C_{ij} is observed). However, the problem in (1) is known to be not only NP-hard, but also all known algorithms that provide exact solutions require time doubly exponential in the dimension n in both theory and practice [36]. However, the analysis presented in [36] proves that the coverage values in vacant bins can be obtained with high accuracy by solving the following alternate convex optimization problem:

$$\begin{aligned} &\text{minimize} \quad ||\boldsymbol{P}||_* \\ &\text{subject to} \quad P_{ij} = C_{ij} \quad (i,j) \in \Psi \end{aligned} \tag{2}$$

where $||\boldsymbol{P}||_*$ is the nuclear norm and is given as:

$$||\boldsymbol{P}||_* = \sum_{k=1}^{n} \sigma_k(\boldsymbol{P}) \tag{3}$$

In (3), $\sigma_k(\boldsymbol{P})$ denotes the kth largest singular value of $\boldsymbol{P}$. Equation (2) therefore aims to determine the matrix with minimum nuclear norm that fits the data.

The problem in (2) can be solved with the singular value-based threshold (SVT) algorithm presented in [37]. The SVT algorithm solves the following problem:

$$\begin{aligned} &\text{minimize} \quad \eta||\boldsymbol{P}||_* + \frac{1}{2}||\boldsymbol{P}||_F^2 \\ &\text{subject to} \quad \mathcal{O}_\Psi(\boldsymbol{P}) = \mathcal{O}_\Psi(\boldsymbol{C}) \end{aligned} \tag{4}$$

where $\mathcal{O}_\Psi$ is the orthogonal projector onto the span of matrices vanishing outside of Ψ so that the (i, j)th component of $\mathcal{O}_\Psi(\boldsymbol{P})$ is equal to P_{ij} if $(i,j) \in \Psi$ and zero otherwise. It is shown in [37] that the solution of the problem of (4) converges to that of (2) as $\eta \to \infty$. The SVT algorithm is iterative and produces a sequence of matrices $\{\boldsymbol{p}, \boldsymbol{Q}\}$. At each step, a soft-thresholding operation is performed on the singular values of the matrix $\boldsymbol{Q}^t$. Thus, by selecting a large value of the parameter, η in (4), the sequence of iterates, $\{\boldsymbol{P}^t\}$ converges to a matrix which nearly minimizes (2). Starting with $\boldsymbol{Q}^0 = \boldsymbol{0} \in \mathbb{R}^{(n\times n)}$, the algorithm inductively defines

$$\boldsymbol{P}^t = \text{shrink}\Big(\boldsymbol{Q}^{t-1}, \eta\Big) \tag{5}$$

$$\boldsymbol{Q}^t = \boldsymbol{Q}^{t-1} + \Delta_i \mathcal{O}_\Psi\big(\boldsymbol{C} - \boldsymbol{P}^t\big) \tag{6}$$

where $\{\Delta_i\}, i \geq 1$ is a sequence of scalar step sizes, until a stopping criteria is reached. The shrink function in (5) applies a soft-thresholding rule at level η to the singular values of the input matrix. It is defined as

$$\text{shrink}\Big(\boldsymbol{Q}^{t-1}, \eta\Big) = \mathcal{S}_\eta\Big(\boldsymbol{Q}^{t-1}\Big) := \boldsymbol{U}\mathcal{S}_\eta(\boldsymbol{\Sigma})\,\boldsymbol{V}^* \tag{7}$$

$$\mathcal{S}_\eta(\boldsymbol{\Sigma}) = \text{diag}\big(\{(\sigma_k - \eta)_+\}\big) \tag{8}$$

where $f_+ = \max(0, f)$. Equivalently, this operator is the positive part of f and simply applies a soft-thresholding rule to the singular values of $\boldsymbol{P}$, shrinking them towards zero. $\boldsymbol{U}$, $\boldsymbol{V}$ are matrices with orthonormal columns and the singular values $\boldsymbol{\Sigma}$ are positive. $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are obtained from the singular value decomposition of matrix $\boldsymbol{P}$ of rank r:

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \quad \boldsymbol{\Sigma} = \text{diag}(\{\sigma_k\}), 1 \leq k \leq r \tag{9}$$

In case of the presence of random shadowing in the model, the stopping criteria of the algorithm can be modified as follows:

$$||\mathcal{O}_\Psi\big(\boldsymbol{P}^t - \boldsymbol{C}\big)||_F^2 \leq (1+\zeta)m\phi^2 \tag{10}$$

where ζ is a fixed tolerance. The SVT algorithm is stopped when $\boldsymbol{P}^r$ is consistent with the data and obeys (10). Therefore, the reconstruction matrix, $\hat{\boldsymbol{C}}$ is the first $\boldsymbol{P}^t$ obeying (10).

Another similar rank minimization based algorithm used to recover the matrix $\boldsymbol{C}$ is the fixed point continuation (FPC) algorithm [38]. While SVT is efficient for large matrix completion problems, it only works well for very low rank matrix completion problems. For problems where the matrices are not of very low rank, SVT is slow and not robust and therefore, often fails [38]. To solve this problem, FPC-based algorithm is proposed in [38]. FPC-based algorithm has some similarity with the SVT algorithm in that it makes use of matrix shrinkage as in (5)-(8). However, it solves (4) by leveraging operator splitting technique [39].

Authors in [34] use matrix completion for the task of interpolating missing RSRP values from MDT-based data. Fig. 3(e), (f) is an illustrative example of their result. Authors in [34] conclude that this scheme is more likely to work well in small cells environments since matrix $\boldsymbol{C}$ will naturally be low ranked in such scenarios. This observation stems from the fact that propagation conditions are mostly dominated by line of sight in small cells and the standard deviation of shadowing is generally small. Moreover, the shadowing phenomenon that heavily determines coverage values, particularly in a small cell environment, remains correlated over small distances that separate users in the same small cell. However, the network scenario they consider consists of macro cell environment, therefore, the application of matrix completion to small cell environments needs further investigation.

Algorithm 1: Singular Value Thresholding Algorithm for Finding Missing Coverage Values

Input : sampled set Ψ and sampled entries $\mathcal{O}_\Psi(\boldsymbol{C})$, tolerance ζ, parameter η, step size Δ, increment α, number of maximum iterations, I_M, shadowing standard deviation ϕ, and cardinality of Ψ, m

Output: $\boldsymbol{P}^{opt}$

Set $\boldsymbol{Q}^0 = i_0\Delta\mathcal{O}_\Psi(\boldsymbol{C})$
Set $\tau_0 = 0$
for $t = 1$ to I_M
  Set $h_t = \tau_{t-1} + 1$
  repeat
    Compute $[\boldsymbol{U}^{t-1}, \boldsymbol{\Sigma}^{r-1}, \boldsymbol{V}^{t-1}]_{h_t}$
    Set $t_t = h_t + \alpha$
  until $\sigma_{h_t-\alpha}^{t-1} \leq \eta$
  Set $\tau_r = \max\{j : \sigma_j^{t-1} > \eta\}$
  Set $\boldsymbol{P}^t = \sum_{j=1}^{\tau_r}(\sigma_j^{t-1} - \tau)\boldsymbol{u}_j^{t-1}\boldsymbol{v}_j^{t-1}$
  if $||\mathcal{O}_\Psi(\boldsymbol{P}^t - \boldsymbol{C})\,||_F^2 \leq (1+\zeta)m\phi^2$ **then break**
  Set $Q_{ij}^t = \begin{cases} 0 & \text{if} \quad (i,j) \notin \Psi \\ Y_{ij}^{t-1} + \Delta(C_{ij} - P_{ij}^t) & \text{if} \quad (i,j) \in \Psi \end{cases}$
end *for* t
Set $\boldsymbol{P}^{opt} = \boldsymbol{P}^t$

B. Inverse Distance Weighted

In this section, we first discuss the simplest form of inverse distance weighted (IDW) method, the simple IDW. Then we highlight several improvements in simple IDW interpolation and finally present an adaptive IDW method from literature.

1) Simple IDW: The simplest form of IDW method is also known as the Shepard's method. It is based on the assumption that the distribution of signal samples is strongly correlated with distance. To estimate the missing received signal strength value, $\hat{c}$ (at a particular bin location, D) in the matrix $\boldsymbol{C}$, weighted average of N known signal strength values, c_k from N adjacent bins are used, where $k = 1 \cdots N$. Each known received signal strength value is weighted with a weight that is equal to the inverse of distance, $d_k = d(D, D_k)$ between the location of the bin with missing RSRP value and location of the k-th bin and raised to the power p. Mathematically, the missing received signal strength value is calculated as:

$$\hat{c} = \begin{cases} \dfrac{\sum_{k=1}^{N} \frac{1}{d_k^p} c_k}{\sum_{k=1}^{N} \frac{1}{d_k^p}} & \text{if } d_k \neq 0 \\ c_k & \text{if } d_k = 0 \end{cases} \tag{11}$$

The choice of p is an important parameter in this method. For $p < 1$, $\hat{c}$ remains no longer differentiable. Therefore,

the exponent has to exceed 1 for the interpolation function to remain differentiable with respect to spatial coordinates (Cartesian coordinates x and y that are used in distance calculation) [40]. It is shown by empirical testing that higher exponents tend to make the surface flat near all data points and the gradients over small intervals between data points are very steep. On the other hand, lower exponents tend to produce a relatively flat surface with short blips to achieve appropriate values at data points [40]. When $p = 0$ in (11), the missing coverage value is set equal to the weighted arithmetic average of the neighboring coverage values and the recovery method is often termed as the 'moving average method'.

Simple IDW method's disadvantages are that it leads to the production of the "bull's-eyes" effect, it is sensitive to measurement outliers, it introduces significant errors in case of non-uniform distribution measurements or unevenly distributed data clusters, computational error becomes highly significant in the neighborhood of a data point, the calculation of missing value increases proportionally with the number of data points, leading to inefficiency of the method when the number of data points is large. Also, there is no way of pre-determining the optimal weighting power factor that will construct the most accurate RF-REM. The appropriate search radius also needs to be optimized. Another drawback is the lack of directionality, i.e., different configurations of co-linear points could yield the same results, attributing to the fact that only the distances from the missing location to the points with known locations are considered and not their direction [25], [40].

However, the advantages of simple IDW method include its efficiency and ease of comprehension since it is intuitive. This interpolation works best with evenly distributed points.

An illustrative example of IDW for REM interpolation using MDT-based RSRP measurements is shown in Fig. 3(f). It can be seen from the figure that although techniques like kriging in Fig. 3(j) outperform IDW in terms of accuracy of REM construction, IDW does outperform several techniques like moving average in Fig. 3(c) and is usually preferred for its reduced computational complexity. IDW has been widely used for REM construction in outdoor environments, such as in [34], where authors use RSRP data to complete scarce REM using IDW. Results in [41] also favor the adoption of IDW for REM construction in a device-to-device network crowd-sourcing scenario consisting of Nakagami-m and Nakagami-lognormal channels.

2) Improved IDW: In order to address the drawbacks of simple IDW method in the preceding subsection, several improvements have been suggested in literature.

The focus of the work in [42] is on the reliable estimation of radio interference field with small number of measurements. For this purpose, different variants of IDW spatial interpolation method are employed which have proven robustness when dealing with limited number of observations [42].

Authors in [40], [43] and [23] improve the weighting function by proposing a framework to intelligently select the nearby data points to be used in predicting the missing data point. This approach is developed keeping the overall density of the data points into consideration.

Authors in [40] incorporate a direction factor, in addition to the distance factor in defining the weights. This direction factor is based on the cosine of angle of $D_i D D_j$, where $i \neq j$ and $i, j = 1 \cdots K$. If other data points D_j are in approximately the same direction from D as D_i, then the angles, $1 - cos(D_i D D_j)$ are close to 0. On the other contrary, if other data points are in the opposite D from D_i, then the angles $1 - cos(D_i D D_j)$ are close to 2. The direction factor in the improved weighting function in [40] leverages this fact.

Other improvements to simple IDW involve reduction of computational complexity and errors and making features of the interpolation function desirable, i.e., ensuring non-zero gradients at every location to achieve the desired partial derivatives for the function to remain differentiable [40], [44].

Since simple IDW assumes that the distance decay is uniform throughout the entire study area, it does not perform well in case of clustered data or data that depicts spatial variability. To address this problem, authors in [45] suggested an improvement based on the weighted median of data in the neighborhood of missing data point. The weighting function in [45] is a function of inverse-distance weights and the de-clustered weights that include the effects of distance and clustering among spatially correlated data in the estimator.

In order to increase the accuracy of predictions through the IDW method, authors in [46] proposed the use of piecewise least-square polynomial regression estimators to increase the accuracy, after evaluating fifteen different estimators using an extensive evaluation data set.

For reducing the "bull-eye" effect in simple IDW method, a distribution-based distance weighting (DDW) technique is used [23]. Weight calculations in DDW method are based on appropriate distributions according to available data, such as Gaussian, Lorentzian and Laplacian distributions. Such a distribution-based calculated ensures that if data variations are very small, then the distribution will have a fairly sharp peak and will cause the weighting to be more sensitive to the distance. On the contrary, if data included in the interpolation are more spread out, a distribution with a larger variance would be a good choice and this would result in the distances having less impact on the weight calculations.

Authors in [23] and [47] propose another improvement to the IDW-based method, that incorporates temporal dimension in addition to spatial dimension. Although these approaches are evaluated in the context of environmental data, such an approach can also be applied to wireless network data. In the approach in [23], time is treated independently from the spatial distance dimension and weights are calculated in two steps: using the inverse of 2D-spatial distance, followed by the inverse of the 1D-temporal distance [23]. Authors in [47] assume second-order non-stationarity of both spatial and temporal distributions of the data, based on which they treat the space-time variables in their proposed method as a sum of independent spatial and temporal non-stationarity components. Heterogeneous covariance functions are constructed to obtain the best linear unbiased estimates in spatial and temporal dimensions [47].

The applications of improved IDW techniques for cellular network data are far less common than their application to the

TABLE I
IMPROVEMENTS TO IDW INTERPOLATION

Improvement	References
Intelligent selection of data in neighborhood	[43], [40], [23]
Addition of directionality	[40]
Reduction of computational complexity	[43], [40], [23], [44], [40]
Reduction of computational errors	[44], [40], [46]
Addition of desirable features	[44], [40]
Extension to clustered/non-uniformly distributed data	[45]
Addition of temporal dimension	[23], [47]
Reduction of "bulls-eyes" effect	[23] [25]

environmental modeling/geoscience domain [23], [46], [47]. In wireless networks context, the study in [42] used improved IDW accounting for the direction, the number and set of considered neighboring points and the slope of the interpolation function, for radio interference field estimation based on distributed spectrum use measurements. It concluded that as compared to classical IDW, improved IDW experiences lower variance of mean absolute error but had more outliers [42].

3) Adaptive IDW: The IDW method assumes that the distance-decay structure is uniform throughout the entire study area. However, recognizing the potential of varying distance-decay relationships over area, authors in [44] proposed a variation in the value of weighting parameter, p according to the spatial pattern of sampled points in the neighborhood using information derived from empirical data. Intuitively, when the unsampled location has highly clustered points around its neighborhood, a small p is appropriate so that the nearest sampled values will not have an overwhelming influence on the estimated value. On the contrary, a large p is desirable when data is spatially dispersed since the more reliable source for the estimate will likely be influenced from the closest location, therefore, if a small p value is used in this case, the contributions from local and more reliable sources will be small, resulting in less reliable estimates [44].

In order to adjust p according to the spatial pattern of known data, authors in [44] first quantify the spatial pattern of sample locations in the form of nearest neighbor statistic:

$$R = r_o/r_e, r_e = \frac{1}{2(M/A)^{0.5}} \tag{12}$$

where r_e and r_o are the expected and observed average nearest neighbor distances respectively and A is the area under consideration.

After normalizing R to get the normalized local nearest neighbor statistic, μ_R, in the adaptive IDW method, this neighbor statistic carries a fuzzy membership that belongs to certain categories of p. This membership function is depicted in Fig. 4. As an example, μ_R corresponding to R of 0.8 will be 0.35, yielding two points in the membership degree (0.3 for category C and 0.7 for category B). The final p would then be a weighted sum of these membership degrees and corresponding p values (0.5 for category B and 1 for category C). Consequently, the final p will be: $0.7 \times 0.5 + 0.3 \times 1 = 0.65$.

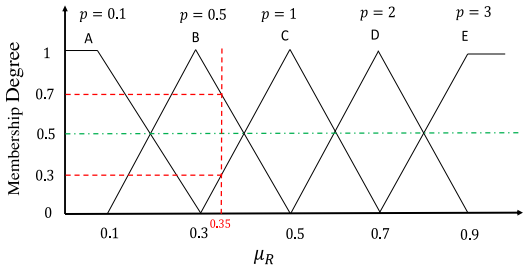


Fig. 4. Triangular membership function for different adaptive distance-decay parameters (modified from [44]).

Adaptive IDW (AIDW) method can outperform IDW and work well in situations where local variability is relatively large or spatial correlation structure of the data is not strong or data is too limited to support data intensive methods, such as kriging. It is shown to outperform ordinary Kriging, when the spatial structure of data was such that it could not be modeled accurately by a variogram function [44].

However, as compared to IDW, the AIDW method is computationally intensive as the distribution of p has to be formulated to find the optimal set of parameter values, which require significant level of heuristics [44].

C. Gradient Plus Inverse Distance Squared

Gradient plus Inverse Distance Squared interpolation (GIDS) combines multiple linear regression and inverse distance based weighted coefficients for the interpolating missing data. By assuming that the data of interest can be represented by a multivariate function, for the unsampled location, D, an ordinary least squared regression is done using N neighboring locations. This yields the coefficients which represent the location gradients. If the measurements are taken at different heights, GIDS method can incorporate the elevation dimension in interpolation too. Assuming $D = (x, y, z)$ with corresponding coefficients C_x, C_y, C_z, representing the x, y, z gradients respectively, the missing data point through GIDS can be estimated as [48]:

$$\hat{c} = \frac{\sum_{k=1}^{N}(c_k + C_x(x - x_k) + C_y(y - y_k) + C_z(z - z_k))/d_k^2}{\sum_{k=1}^{N} 1/d_k^2} \tag{13}$$

The advantage of GIDS method is its ability to account for signal level gradients and elevation of the terrain at the interpolated location and at locations of the measurements. However, this method is very sensitive to the selection of neighborhood points as a small neighborhood selection would leave out important measurements and a large neighborhood selection may introduce noise [25].

GIDS has been used for REM construction in [48], where authors conclude that when number available measurements are sufficient, then Kriging outperforms GIDS in terms of lower relative mean absolute error in most REM simulation scenarios. Note also that Kriging is highly sensitive to the performance metric used as it minimizes mean squared error

(MSE), so performs best when MSE is used as evaluation metric.

D. Modified Shepard's Method

The IDW based modified Shepard's method (MSM) is a local interpolation that makes the estimation based on a real multivariate function, *f*, whose local approximation is referred to as nodal functions. If Q_k is the output of the nodal function of the data point D_k (local approximation to f at x_k, y_k), then the missing value using the MSM method can be written as a weighted average of the nodal functions within some radius influence (about the missing data point), R_w in the following manner [42], [48]:

$$\hat{c} = \frac{\sum_{k=1}^{N} W_k Q_k}{\sum_{k=1}^{N} W_k} \tag{14}$$

First, the weights, W_k are calculated by the following formula:

$$W_k = \begin{cases} [R_w - d_k]/[R_w d_k]^p & \text{if } d_k < R_w \\ 0 & \text{if } d_k \geq R_w \end{cases} \tag{15}$$

Then, another radius, R_v around each known data point is considered and the weights are again calculated using (15), this time, replacing R_w with R_v.

This technique can be extended to multivariate case but is dependent upon optimization of R_w, R_q and p. It is also shown to perform poorly if measurements lie in a low-dimensional subspace [25]. However, this method can reduce the 'bull's eye' effect as compared to classical IDW methods.

An example of MSM application for the task of generating REM of total received signal power is illustrated in [48]. Authors in [48] use a wireless system simulator to simulate both indoor and outdoor scenarios with different levels of data scarcity. Among the considered methods of Kriging, MSM and GIDS, MSM generally performs somewhere in between the other two. For example, when the measurement points increase from 38 to around 695, the relative mean absolute error (RMAE) reduces from 7.5% to 1% for Kriging, 8% to 1.5% for MSM, and 9% to 2% for GIDS. They thus conclude that although Kriging performs best in terms of interpolation error, but due its high computational complexity and weak performance when observation points are low, MSM may be preferred as it is more flexible and robust.

E. Nearest Neighbor

The nearest neighbor (NeN) method is also known as proximal interpolation or point sampling. Let D_l be the nearest neighbor of the missing point, *D* and $d(D, D_l)$ denote the distance between D_l and D, then $\min\{d(D, D_k)\} = d(D, D_l)$, $k = 1 \cdots N$. In this case, the estimated value will be the same as the value in the nearest sampled location *l*. Mathematically, the weights, λ_k can be represented as [49]:

$$\lambda_k = \begin{cases} 1 \text{ if } k = l \\ 0 \text{ if } k \neq l \end{cases} \tag{16}$$

which leads to the missing point prediction as:

$$\hat{c} = \sum_{k=1}^{N} \lambda_k c_k = c_l \tag{17}$$

Nearest neighbor method is known for its low complexity. Among the considered techniques in [50] for the task of interference map interpolation, nearest neighbor interpolation is concluded to be the least complex method and natural neighbor, linear, cubic and quadratic interpolation techniques have shown to exhibit comparable performances.

Although nearest neighbor approach is of low complexity, it results in sharp transitions between the individual signal level zones and increases noise, especially at the boundary of a given area, since it does not consider the influence of the sample data points apart from the nearest neighboring data point [25], [51].

Fig. 3(g) illustrates an example of using nearest neighbor interpolation to interpolate scarce RSRP measurements for constructing coverage maps. It can be seen from the figure that compared to methods like kriging in Fig. 3(j) where the interpolated coverage map is smooth, nearest neighbor interpolation results in a representation that has more sharper transitions between adjacent values.

F. Natural Neighbor

The natural neighbor (NaN) interpolation is based on Voronoi decomposition (tessellation) of a set of given points in the plane. The received signal strength value at a particular location is found from a weighted average of *N* from all available measurements which fall within its 'natural neighborhood'.

The natural neighbors of any point are those associated with neighboring Voronoi polygons. If the 2-D point D_k is a natural neighbor of the 2-D point $\mathbf{D}$, the portion of Voronoi region, V_{D_k} stolen away by $\mathbf{D}$ is called the natural region of $\mathbf{D}$ with respect to $\mathbf{D}_k$. Initially, a Voronoi diagram is constructed of all the available coverage values. Then, a new Voronoi polygon is created around the interpolation point (missing coverage value). The proportion of overlap between this new polygon and the initial polygons is then used as weights. If we denote the Lebesgue measure of this natural region by $l_{\mathbf{D}_k}$, the natural coordinate associated to $\mathbf{D}_k$ is used as weights [14]:

$$\lambda_{\mathbf{D}_k}(\mathbf{D}) = \frac{l_{\mathbf{D}_k}(\mathbf{D})}{\sum_k l_{\mathbf{D}_k}(\mathbf{D})} \tag{18}$$

The weights are thus the ratio of the area of overlap to the total area of the new polygon. Once the weights are obtained, interpolation to find the missing coverage value can be carried out by a weighted sum of known coverage values.

The natural neighbor interpolation method performs well with non-homogeneous distribution of measurements as well. However, its major drawback is that it can not find missing signal values that lie outside the convex hull of Voronoi polygons since it requires that the points to be interpolated be in the convex hull of the measurement locations as the Voronoi cells of outer data points are open-ended polygons with an infinite area [25].

Another scheme similar to natural neighbor using an area-wise multi-criteria triangulation-induced interpolation algorithm which utilizes the linear interpolation to estimate the key performance indicators of the QoS inside a triangle with the

known values of its three vertexes is proposed to reconstruct the coverage maps in [52].

Fig. 3(h) is an illustrative example of the result obtained by applying natural neighbor for the task of interpolating missing RSRP values from MDT-based data in [34]. An important observation is the interpolation at the corners of the coverage map in Fig. 3(h), that do not have any value due to the inability of natural neighbors to fill the missing values that lie outside the convex of Voronoi polygons as identified above.

G. Splines

The spline method is also referred to as the radius basis function and 'rubber sheeting' [25]. It estimates the missing value by a mathematical function or piecewise defined polynomials called splines that minimizes the total surface curvature. This results in a smooth surface that passes exactly through the sampled points. This interpolation method is useful for estimating above maximum and below minimum points and for creating a smooth surface effect. However, because of this smoothing effect, the discontinuity in data might not be well estimated. Since it uses slope calculations or change over distance to estimate the missing values, when the known data points are too close together or have extreme differences in values, this method does not work well.

There are different kinds of splines, such as linear, quadratic, cubic, biharmonic and thin-plate splines. For example, for thin-plate splines, the unknown value is estimated as [14]:

$$\hat{c} = \sum_{k=1}^{N} w_k ||D - D_k||^2 \ln(||D - D_k||) \tag{19}$$

where $||.||$ is the Euclidean norm. w_k can be obtained by solving $\mathbf{Ow} = \mathbf{i}$, where $\mathbf{i}$ and $\mathbf{w}$ are the column vectors of input data points and weights respectively, while $\mathbf{O}$ is the matrix of output of the basis function ($||D-D_k||^2 \ln(||D-D_k||)$ in this case) for all possible input values.

A visual example of splines in the case of REM construction of RSRP measurements is illustrated in Fig. 3(i). Authors in [34] conclude that Splines and Kriging have similar performance quantitatively in terms of relative recovery error (Frobenius norm of recovered interpolated matrix minus the ground truth matrix divided by Frobenius norm of ground truth matrix).

H. Kriging

Kriging, unlike the other methods discussed above, also takes into account the statistical relationships in additional to spatial relationships among the measured data points to estimate the missing values of data.

In Kriging, the weights are based not only on the distance between the measured points and the prediction location but also on the overall spatial arrangement of the measured points [53], [54]. The weight coefficients are calculated by minimizing the variance of the estimation error, σ_e^2:

$$\sigma_e^2 = \mathbb{V}\left[\hat{C}_m - C_m\right] \tag{20}$$

where $\mathbb{V}$ is the variance operator and C_m is the missing coverage value located at the 2-D point, $\mathbf{p}$.

The first step in kriging therefore involves creating a prediction surface map in order to uncover the dependency rules to make predictions. To achieve this, kriging first creates a semivariogram and covariance functions to estimate the statistical dependence values that depend on the model of autocorrelation. To solve the optimization problem in (20), semivariogram function, γ is used to characterize the spatial correlation.

The next step is to fit a model to the points forming the empirical semivariogram. A mathematical function is used to fit the empirical semivariogram as the theoretical semivariogram model to model spatial autocorrelation. There are many variants of kriging based on advanced and robust semivariogram models, such as simple kriging, block kriging, factorial kriging, kriging with a trend, dual kriging, universal cokriging, kriging with an external drift, indicator kriging, probability kriging, to name a few. A comparison of these variants is presented in [21], [24]. Kriging weights then come from the semivariogram that was developed by analyzing the spatial nature of the data. These weights are a result of minimizing the variance in (20), which yield the following solution [49]:

$$\begin{bmatrix} \boldsymbol{\lambda} \\ \delta \end{bmatrix} = \mathbf{X}^{-1}\mathbf{y} \tag{21}$$

where $\mathbf{X}$ and $\mathbf{y}$ are defined as:

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,N} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ X_{N,1} & \cdots & X_{N,N} & \vdots \\ 1 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 1 \end{bmatrix} \tag{22}$$

Each element of matrix, $\mathbf{X}$, $X_{i,j} = \gamma(||\mathbf{p}_i - \mathbf{p}_j||)$ and each element of the column vector $\mathbf{y}$, $y_i = \gamma(||\mathbf{p} - \mathbf{p}_i||)$. The extra element in the weight vector solution in (21), δ, is the result of fitting by assuming a mean trend component in the reconstructed coverage matrix.

Kriging is applied on RSRP measurements for REM construction in [55], [56]. A more practical implementation of Kriging based approach using real data from the University of Colorado, Boulder campus has been demonstrated in [15]. In [57], the authors propose a REM construction method by combining residual maximum likelihood-based radio propagation parameter estimation with Kriging-based transmission power prediction. They then benchmark the performance of their proposed algorithm with a path loss-based method and a Kriging-based method without prior fit of a path loss model, using the metric of root mean square error (RMSE). Another Kriging-based radio environment map construction method based on mobile crowd sensing is proposed in [58]. Authors in [58] compare Kriging with the nearest neighbor and the inverse distance weighting interpolation algorithms and conclude that Kriging performs the best for their crowdsourced RSRP dataset. Kriging is applied in the context of a REM-enabled spectrum sharing mechanism for performance analysis for mobile cellular networks in [59]. Authors in [60] propose an improved Kriging algorithm by combining the concept of affinity propagation clustering in ordinary Kriging algorithm for REM construction. Another improvement over ordinary

Kriging is the fixed-rank Kriging proposed in [61]. However, it tends to neglect the small-scale structured variations of the data, which may result in a loss of accuracy [62]. To overcome the limitations of ordinary and fixed-rank Kriging, authors in [62] propose covariance tapering based Kriging. Neural network techniques are also applied to improve Kriging algorithm in [63], [64], [65].

In the domain of cognitive radio networks, authors in [14] compare three interpolation methods, namely, natural neighbor, kriging and spline for constructing interference cartographs from a scarce set of data. They conclude that both kriging and natural neighbor interpolations perform similarly when the channel uncertainty is lower and that the average efficiency of all interpolation techniques improves with increased shadowing decorrelation [14]. Authors in [51] conclude that Kriging performs best among nearest neighbor and inverse distance weighted (IDW) methods. Results in [49] again demonstrate the superior performance of Kriging among nearest neighbors, IDW and triangular irregular network interpolation, but has demonstrated the robustness of IDW method overall.

Authors in [48] compare Kriging, Modified Shepard's method (MSM) and Gradient plus inverse distance squared (GIDS) and IDW for creating radio environment maps. It is concluded that Kriging and IDW are most flexible among these methods and offer trade-off between the computational cost and accuracy.

Kriging has also been used in indoor environments, such as in [66], where authors compare various interpolation techniques, including Kriging, splines, weighted moving average, Theissen polygons, trend surfaces, classification, in terms of accuracy, spatial distribution of measurements, measurement density and impact of a fixed location inaccuracy for the task of signal strength prediction in an indoor environment. The results in [66] indicate that Kriging is a fairly robust technique overall, across all considered scenarios. Kriging has also shown to be the method which is least sensitive to the deployment of the sensors as compared to nearest neighbor and inverse distance weighted in [67], where the authors analyzed the impact of the number of sensors on the REM quality in the context of military wireless networks. They used data from real field tests with 39 sensors in an area of 4 km^2.

Fig. 3(j) is an illustrative example of the result obtained by applying Kriging for the task of interpolating missing RSRP values from MDT-based data in [34]. Authors in [34] report that among the methods considered in Fig. 3, kriging method performs the best with the least quantitative relative recovery error (Frobenius norm of recovered interpolated matrix minus the ground truth matrix divided by Frobenius norm of ground truth matrix) of less than 0.15. This is because in contrast to other interpolation methods where the weights are only dependent on the distance, the weights in kriging are based on the overall spatial arrangement of the measured points too.

The major drawbacks of Kriging are that it requires a large number of measurement points in order to achieve high precision and it involves significant input from the user in order to select the best fit function for the semivariogram. Identifying the most appropriate theoretical variogram for the given data (especially if it exhibits large spatial heterogeneity) is critical in order for Kriging to perform well. Although Kriging has relatively high computational complexity, it is the most commonly applied technique in the literature [31], [53] due to its higher precision. As Kriging is geostatistical method, it also can estimate the variances of predicted values in the unsampled location.

I. Lessons Learned

Among the interpolation methods, Kriging has been most widely used in literature due to its high accuracy. However, it is computationally expensive. Simpler and less computationally demanding techniques, like IDW, are shown to work best for evenly distributed data points. Kriging, GIDS, MSM and Splines can be used in cases where extrapolation is required. However, when extrapolation is not required, IDW, natural neighbors and nearest neighbors are candidate choices. Among these, natural neighbors require all data points be inside the convex hull of location measurement. Another method, matrix completion, although has shown to be very promising in other domains, its applicability to small cell environments where it will most likely work best needs further investigation.

III. Methods Using Contextual Information

The preceding section discussed techniques that can be leveraged to address the data scarcity challenge when the only known information are the measured data and their locations. However, if some additional information other than the observed data is known, we can employ other techniques leveraging that additional information, or use it to enhance the interpolation methods.

This additional information can be knowledge of propagation model, such as path loss and other relevant parameters, transmitter parameters, such as transmit power or antenna patterns, transmitter location estimation, network geometry, or characteristics of the operating environment. It is then combined with observed scarce data to augment it. Based on the availability of known information, different indirect approaches can be employed. For example, authors in [68] estimate the transmitter power and location using received signal strength (RSS) measurements and empirical model to enrich REM. Similarly, authors in [69] calibrate propagation model using transmit power, antenna diagram, azimuth and tilt angles before generating more RSS data through it.

A. Utilizing Geometry of Network

1) Triangular Method (Interpolation Using Locations of Data Base Stations): One way to estimate measurements for bins with no user reports can be using the geometry of the base stations as shown in Fig. 5. This is particularly suitable in ultra-dense deployment scenarios [70], where the data base stations (DBSs) are very densely deployed (by virtue of switching OFF DBSs to keep energy consumption and interference low). These additional measurements, after appropriate transformation, can then be used to increase the accuracy of interpolation

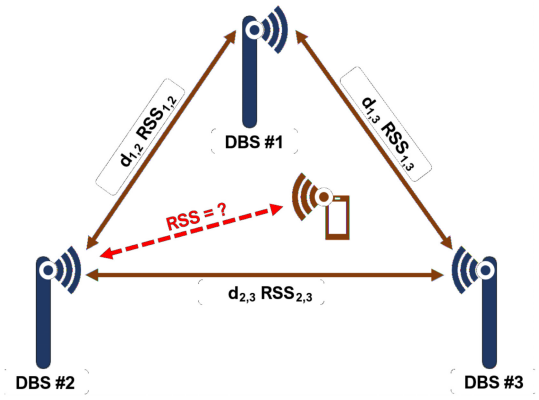


Fig. 5. Leveraging dense base station deployment to enrich scarce data.

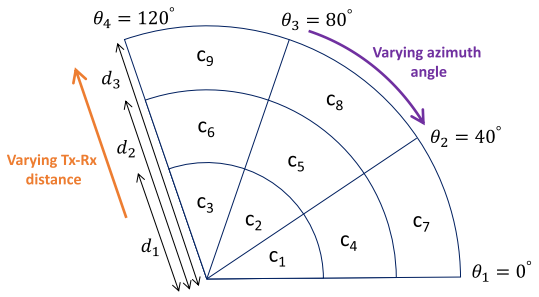


Fig. 6. Leveraging cluster geometry to enrich scarce data.

methods proposed above. However, this approach can complement only simple measurements such as received signal strength.

2) Arc Method (Exploiting Pattern Among Clusters in Polar Coordinates): Another way to enrich scarce data in a given network area can be by dividing the area into clusters into polar coordinates as shown in Fig. 6. Each cluster has a value that can show a given KPI, such as the average RSRP or SINR of the users in that cluster. To find the missing value in a particular cluster, geometric pattern among other clusters can be exploited, for example, if we travel along a particular circumference, we observe that the Tx-Rx distance remains constant on that circumference and the only variation is in azimuth angle (θ_1 to θ_4 in Fig. 6). Conversely, if we traverse a path radially outwards, we can notice that the azimuth angle remains the same but there is variation in Tx-Rx distance (d_1 to d_3 in Fig. 6 assuming base station is located at the center of the sector). If we model the received signal strength as a function of azimuth angle and Tx-Rx distance, this pattern can be exploited to find the unknown signal strength values.

Learning cluster values by exploiting this pattern using a supervised DNN has been proposed in [11]. However, authors in [11] has not used this approach to address the data scarcity challenge. In [11], correlations among their SINRs has been exploited to learn the locations of users at macrocells. However, we propose that such a model based on correlations among SINRs of known clusters can also be used to find the missing SINR in another cluster.

B. Through Propagation Modeling and Transmitter Parameter Estimation

1) Received Signal Strength (RSS) Based: The RSS based method to recover scarce data is based on a combination of analytical models with statistical evaluation through measurements [68]. The RSS at a particular receiver, i located at a distance, d can be represented as:

$$P_i(d) = P_t - L - 10p\log_{10}(d) + \phi \tag{23}$$

where P_t is the transmit power, L is the free space path loss and ϕ represents a lognormal random variable for shadowing. L, p and standard deviation of ϕ are environment dependent parameters.

After averaging out RSS measurements (in order to reduce random shadowing effect), and assuming the sample size of RSS measurements is large enough, the average RSS at a particular location can be estimated as:

$$P_i^{av}(d) \approx P_t - L - 10p\log_{10}(d), \text{ where } P_i^{av}(d) = \sum_{k=1}^{N} P_i^k(d)/N \tag{24}$$

After performing some algebraic manipulations, taking the anti-log of (24) and representing d is cartesian coordinates, (24) can be transformed into a regression problem which can be expressed as a system of linear equation as follows [13]:

$$\begin{bmatrix} 10^{\frac{-L-P_1^{av}(d)}{5p}} & 2x_1 & 2y_1 & -1 \\ 10^{\frac{-L-P_2^{av}(d)}{5p}} & 2x_2 & 2y_2 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 10^{\frac{-L-P_N^{av}(d)}{5p}} & 2x_N & 2y_N & -1 \end{bmatrix} \begin{bmatrix} 10^{\frac{P_t}{5p}} \\ x_t \\ y_t \\ x_t^2 + y_t^2 \end{bmatrix} = \begin{bmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ \vdots \\ x_N^2 + y_N^2 \end{bmatrix} \tag{25}$$

where x_t, y_t is the transmitter location and(x_i, y_i) is the i-th receiver location. Therefore, by solving (25) using least-squares methods, we get estimates for transmit power, P_t and the location of transmitter, (x_t, y_t). These estimates can then be used to evaluate estimated received power at the missing location, by first calculating the Tx-Rx distance at the missing location and then using it to find RSS. A similar method combining transmitter localization estimation with Kriging is proposed in [71].

Note that since path loss and shadowing parameters in the model are assumed to be known and are highly environment dependent, the quality of estimated is likely to be drastically affected if there is an error in estimation of propagation parameters, caused by, for example, high shadowing fading in the environment. However, this method is likely to improve if propagation conditions are not too drastic, for example, in rural areas and if the number of receivers with known measurements are large. It is also shown in [13] that unlike IDW and Kriging, RSS-based method is not affected by the minimum distance between receiver and transmitter and therefore, is more robust as compared to interpolation methods alone.

RSS algorithm was applied for the task of REM interference cartography generation in [68]. Results from [68] show that the transmitter location estimation error decreases in an exponential manner as the number of sensor measurements increases.

2) Received Signal Strength Difference (RSSD) Based: The RSSD method is based on the received signal strength difference (RSSD) between two base stations or transmitters. It is assumed that transmit power is known, transmitter location, (x_t, y_t) is estimated based on the idea that the ratio of the signal powers (or their differences expressed in dB) observed at two different receiver locations is related to the ratios of the transmitter to receiver distances. Specifically, the received power differences between any two receivers, located at (x_a, y_a) and (x_b, y_b) can be represented as [68]:

$$P_{ab} = 5p \log_{10}\left(\frac{(x_t - x_a)^2 + (y_t - y_a)^2}{(x_t - x_b)^2 + (y_t - y_b)^2}\right) \tag{26}$$

The transmitter location in (26) can then be estimated by solving a linear system of equations of the following form:

$$\begin{bmatrix} 1-\beta_{12} & -2(x_2-\beta_{12}x_1) & -2(y_2-\beta_{12}y_1) \\ 1-\beta_{13} & -2(x_3-\beta_{13}x_1) & -2(y_3-\beta_{13}y_1) \\ \vdots & \vdots & \vdots \\ 1-\beta_{1N} & -2(x_N-\beta_{1N}x_1) & -2(y_N\beta_{1N}y_1) \end{bmatrix} \begin{bmatrix} x_t^2+y_t^2 \\ x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \beta_{12}\left(x_1^2+y_1^2\right)-\left(x_2^2+y_2^2\right) \\ \beta_{13}\left(x_1^2+y_1^2\right)-\left(x_3^2+y_3^2\right) \\ \vdots \\ \beta_{1N}\left(x_1^2+y_1^2\right)-\left(x_N^2+y_N^2\right) \end{bmatrix} \tag{27}$$

where $\beta_{ab} = \frac{(x_t-x_a)^2+(y_t-y_a)^2}{(x_t-x_b)^2+(y_t-y_b)^2}$. Solution to (27) by ordinary least squares using available receiver locations yields estimates for $x_t, y_t, x_t^2+y_t^2$. Once the transmitter location has been estimated, the received signal level at any location can also be estimated by subtracting the path loss from transmitted signal power. As with RSS based method, this method is also dependent on selection of propagation parameters, such as path-loss exponent and shadowing spread.

Performance comparison between RSS and RSSD based methods for REM construction was done in [68]. Results in [68] show that the transmitter location estimation error decreases in an exponential manner as the number of sensor measurements increases. For example, as the number of measurements increase from 6 to 20, the transmitter location error decreases from to 75 m to around 23 m for RSSD based approach and it decreases from around 24 m to approximately 12 m for the RSS based method. As can be seen quantitatively, RSSD algorithm outperforms RSS based method for all measurement densities.

3) Angle of Arrival (AOA) Based: Using prior knowledge of transmit power and using measurements from N receivers with known locations, this method first estimates the angles of arrival at the locations of the measurements and combines them with the received signal powers to estimate the location of the transmitter. Once the location of the transmitter and its transmit power is available, any appropriate propagation model can be applied to estimate unknown data at different locations.

The signal model for received signal at i-th receiver is modeled as [72]:

$$\mathbf{R}_i = \sqrt{\alpha}(d_i)\mathbf{h}(\theta_i)s + \mathbf{n}_i \tag{28}$$

where s is the complex baseband transmitted signal with known transmit power, d_i is the unknown distance between the unknown transmitter and receiver, θ_i is the unknown angle by which the signal reached the i-th receiver and $\mathbf{n}_i$ is additive white Gaussian noise vector. The (θ_i, d_i) pair represents a unique position. The directional and attenuation characteristics of the channel $\mathbf{h}$ can be modeled by:

$$\mathbf{h}(\theta_i) = \begin{bmatrix} 1 \\ \exp\left(j\frac{\pi}{2}\sin(\theta_i)\right) \end{bmatrix}, \alpha(d_i) = \phi\left(\frac{c}{4\pi f}\right) d_i^{-p} \tag{29}$$

For the recovery of missing measurements, first, the angle of arrival based on the received signal strength is estimated at each receiver and then a fusion of these estimates is performed. For angle of arrival estimation, authors in [72] apply the multiple signal classification (MUSIC) algorithm and obtain estimated of the pair (θ_i, d_i), that translate into a location estimate for the i-th receiver:

$$\begin{bmatrix} \hat{x}_t^i \\ \hat{y}_t^i \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} \hat{d}_i \cos\left(\hat{\theta}_i\right) \\ \hat{d}_i \sin\left(\hat{\theta}_i\right) \end{bmatrix} \tag{30}$$

Next, these estimated locations are transferred to a central network that combines these estimates. One way to combine these estimates can be through simple averaging. Another fusion method proposed in [73] obtains the following over-conditioned system from the estimates:

$$\begin{bmatrix} -x_1 \sin\left(\hat{\theta}_1\right) + y_1 \cos\left(\hat{\theta}_1\right) \\ \vdots \\ -x_N \sin\left(\hat{\theta}_N\right) + y_N \cos\left(\hat{\theta}_N\right) \end{bmatrix} \approx \begin{bmatrix} -\sin\left(\hat{\theta}_1\right) & \cos\left(\hat{\theta}_1\right) \\ \vdots & \vdots \\ -\sin\left(\hat{\theta}_N\right) & \cos\left(\hat{\theta}_N\right) \end{bmatrix} \begin{bmatrix} \hat{x}_t \\ \hat{y}_t \end{bmatrix} \tag{31}$$

Solving this system of equations through least squares solutions yields the transmitter location, which can then be combined with known transmit power and a suitable propagation model to estimate signal strengths at unknown locations.

Authors in [72] use AOA based method for interference source localization to interpolate REMs. Authors in [72] compare the AOA based method with simple averaging method (where averaging of sensor estimates by all sensors is done) and SNR based method in Section III-B4, where sensor results are weighted by each sensor's SNRs. The AOA method outperforms the other two methods at low SINRs.

4) Signal to Noise Ratio (SNR) Based Method: The initial steps of this method are similar to AOA based method in which the estimation step at each receiver enables the estimation of the angle of arrival and the received signal power. However, in the later step, combination of the location

estimates is done through SNR-aided fusion. The basic idea of this approach is the observation that receivers far away from the transmitter yield worse location estimates. Hence the receiver results are weighted with their respective receiver's SNR, Γ_i as follows [72], [74]:

$$\begin{bmatrix}\hat{x}_t\\ \hat{y}_t\end{bmatrix} = \sum_{i=0}^{N} \frac{\Gamma_i}{\sum_{k=1}^{N}\Gamma_k}\begin{bmatrix}\hat{x}_t^i\\ \hat{y}_t^i\end{bmatrix} \tag{32}$$

where the received SNR at the *i*-th receiver is:

$$\Gamma_i(d) = E\left[\frac{\alpha(d_i)P_t}{N_o B}\right] \tag{33}$$

with N_o being the noise power density and B being the bandwidth of the receiver.

The SNR based method has been used for interference source localization for cognitive radio scenarios to interpolate REMs in [72]. Authors in [72] conclude that AOA based method using tens of sensor nodes with two antennas in an area of 2500 m $\times$ 2500 m can meet the location error requirement of FCC, which is $\pm$ 50 m and outperforms AOA based method at moderate to high SINR.

5) Self-Tuning Method: Another method utilizing propagation parameters but also taking the antenna pattern into account is the self-tuning method (STM) is proposed in [69]. In addition to leveraging characteristics of the operating environment, it performs estimation of the transmitter location, antenna parameters, transmit power and parameters of the propagation model such that the error between available measurements and predicted data is minimized.

Using the scarce data collected, the STM first estimates transmitter parameters and calibrates the propagation model. This is then used to predict missing data, such as signal levels. Among these transmission parameters, the location of transmitter is calculated using localization algorithms based on parameters such as angle of arrival or timing advance, time of arrival or time difference of arrival. Then, based on the transmitter location, distance from transmitter to receiver is calculated. This distance is then used in an appropriate propagation model. As an example, if the Okumura-Hata model is used, the received power at a particular location can be represented as:

$$\begin{aligned}P_r = {} & P_t - A_o - A_1\log_{10}(d) - A_2\log_{10}(H_e)\\ & - A_3\log_{10}(d)\log_{10}(H) + 3.2(\log_{10}(11.75H_m))^2\\ & - 44.49\log_{10}(f) + 4.78(\log_{10}(f))^2 - L_d - L_c + G\end{aligned} \tag{34}$$

where P_t is the transmit power, d is the transmitter-receiver distance, f is the operating frequency, L_d represents the diffraction loss, L_c is the loss through terrain clutter, H is the height of transmitter and A_o, A_1, A_2, A_3 are the constant coefficients. G represents the antenna gain and can be represented as [69]:

$$\begin{aligned}G = {} & G_{\max} - F_\theta + F_\theta\left|\cos^{p_1}\left(\frac{\theta_{azi}-\theta_u}{2}\right)\right|\\ & - F_\phi + F_\phi\left|\cos^{p_2}\left(\frac{\phi_{tilt}-\phi_u}{2}\right)\right|\end{aligned} \tag{35}$$

where ϕ_{tilt} is the tilt angle of the antenna, ϕ_u is the vertical angle from the reference axis (for tilt) to the user. θ_{azi} is the angle of orientation of the antenna with respect to horizontal reference axis, i.e., positive x-axis, θ_u is the angular distance of the user from the horizontal reference axis. G_{max} represents the maximum antenna gain and F_θ and F_ϕ are the front to back ratios in both directions, whereas the antenna form is approximated with the cosine functions to the power of p_1 and p_2

We suggest that another option for a more practical directional antenna model defined by 3GPP and utilized in [8] can be as follows:

$$\begin{aligned}G = {} & \lambda_\phi\left(G_{\max} - \min\left(12\left(\frac{\phi_u-\phi_{tilt}}{B_\phi}\right)^2, A_{\max}\right)\right)\\ & + \lambda_\theta\left(G_{\max} - \min\left(12\left(\frac{\theta_u-\theta_{azi}}{B_\theta}\right)^2, A_{\max}\right)\right)\end{aligned} \tag{36}$$

The additional antenna parameters in this model are the half power vertical and horizontal beamwidths, B_ϕ and B_θ respectively and the side and back lobe attenuation, $A_{\max}$.

Having defined a suitable propagation and antenna model, the optimal antenna, transmitter and propagation environment parameters can then be obtained by minimizing the mean squared error between the measured and estimated signal strengths. Authors in [69] solved this optimization problem in a non-least squared sense, using prior knowledge of the bounds for the parameters to be optimized.

After solving the optimization problem by a suitable algorithm, the optimized parameters are applied in the calculation of signal levels at unknown location to augment the existing data.

Note that L_d and L_c require knowledge of the propagation environment, such as access to clutter database of a mobile operator or knowledge of the digital elevation model [69]. Also, antenna parameters knowledge through antenna datasheets or antenna diagrams is required in this method.

STM has been applied for constructing the radio frequency layer of REM in [69]. When 1000 measurements are used, STM method obtains the lowest RMSE of 5, followed by Kriging with RMSE of 17.5, while IDW attains the highest RMSE of 22.5 [69].

C. Lessons Learned

The methods discussed in this section can be used in cases where some additional contextual information is known. Based on the network geometry, triangular method can be used in the case when transmitter locations are known, and arc method can be used in cases where transmitter locations are not known. When the propagation environment parameters are known, along with the transmit power and receivers' SNR, the SNR-based method can be used. However, if SNR is not known, but antenna characteristics are known, the STM method can be a potential candidate solution. There are also methods such as AOA based method, RSS, RSSD based method that do not require antenna or SNR information, but instead make use of mathematical equations/models after estimating or using prior knowledge of the transmit power and location. However, since

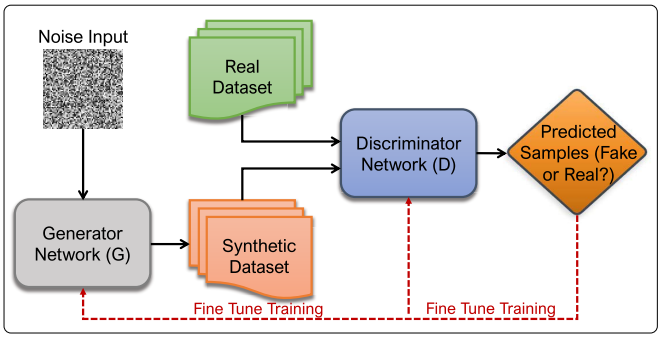


Fig. 7. Conventional GAN architecture.

these methods are mostly based on analytical models, they inherit some assumptions.

IV. Machine Learning Methods

Several machine learning techniques such as generative adversarial networks (GANs), autoencoders, transfer learning and few-shot learning techniques can be leveraged to address the training data scarcity challenge in radio access networks. In certain RAN use-cases involving higher dimensional datasets, these neural network based techniques can be trained with much less training data (or with higher performance for the same amount of data) due to their efficient learning ability for higher-dimensional datasets as compared to previously mentioned interpolation and contextual information based methods [75]. Examples of scarce data and use cases in RAN where ML techniques have shown superior performance than other techniques, include CDR data for traffic map prediction [19], [76], MDT data for outage detection [77], cell trace data for performance analysis [78], RSS data for pathloss prediction [79], [80], RF data for radio map generation [81], [82] and configuration data for performance prediction [83], [84], [85], [86], [87], [88], [89], [90].

A. Generative Adversarial Networks

Generative adversarial networks (GANs) success in image processing has been well established [91], [92], [93], [94], [95]. Although this concept has widely been used in image processing, it can also be used in wireless communications. In wireless communications context, the works that utilize GANs are limited to [19], [76], [77], [81], [82], [96], [97], [98]. While GANs have been widely used for image data, its application to tabular data remains relatively limited. The works that use GANs on tabular data in a non-cellular network data context include [99], [100], [101], [102], [103], [104]. However, similar concepts can be applied to wireless data domain too.

The basic idea of GAN illustrated in Fig. 7 is to generate large amount of synthetic data building on small amounts of real data which will not be distinguishable from real data. The intuition behind GANs is to exploit the potential of deep neural networks (DNNs) to both model nonlinear complex relationships (the generator) as well as classify complex signals (the discriminator). In GAN, a two-player minimax game is set between the discriminator DNN and generator DNN as follows:

$$\min_{G}\max_{D} \; V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{37}$$

where V (D, G) is the value function over which training happens, the latent variable z is randomly drawn from prior distribution $p_z(z)$, x is sampled from $p_{\text{data}}(x)$, generator G is a mapping from the latent variable z to data space and the discriminator is a scalar function of data space that outputs probability that input was genuine. Other types of loss functions for the discriminator and generator for different types of GANs are described in [105]. In each training epoch, the generator iterates its weights to produce synthetic data trying to fool the discriminator DNN. The discriminator DNN on the other hand, tries to discriminate between real data and generated data. In theory, when Nash equilibrium is reached between the generator DNN and discriminator DNN, the pair of DNNs will provide us a generator that can exactly duplicate or reproduce the distribution of the real data so that the discriminator would be unable to identify whether a sample is synthetic, i.e., whether it is generated by the generator DNN or it is from the real data. At this point, the synthetic data generated by the generator DNN are indistinguishable from the real data, and are thus as realistic as possible.

To assess the efficacy of GAN-based approach outlined above, as a preliminary study recently published in [19], GAN was leveraged to generate synthetic call data records (CDRs) data and thus increased training dataset size by enriching the real scarce CDR from [106] with realistic synthetic data. CDRs data are selected as preliminary case study because CDR data can be used by a large number of SON solutions such as in [107], [108]. Real network traces with call durations and call start time stamps, provided by one of the leading mobile operators in USA, were used in this study to train the GAN. The discriminator was trained beginning with 20,000 data points (from a record of several hundred thousand). Once the discriminator could reliably differentiate between the real data taken from the record and randomly generated CDR data with two features, i.e., call duration and start time, the generator was trained. After the generator was generating data that the discriminator perceived to be real, we used the trained generator to produce another 20,000 CDR data samples. Figs. 8(a) and 8(c) and represent the distribution of the real data used to train the discriminator. Figs. 8(b) and 8(d) show the distribution of the 20,000 synthetic data points produced by the trained generator. These preliminary results show the high similarity between real and synthetic data produced by the proposed GAN based approach.

Other GAN-based approaches in cellular networks context include the use of GANs to address the imbalance data issue in cell outage detection [77] Authors in [77] use an LTE simulator to get RSRP and RSRQ data and combine GAN with AdaBoost to improve classification performance of imbalanced data for cell outage detection in self-organizing cellular networks.

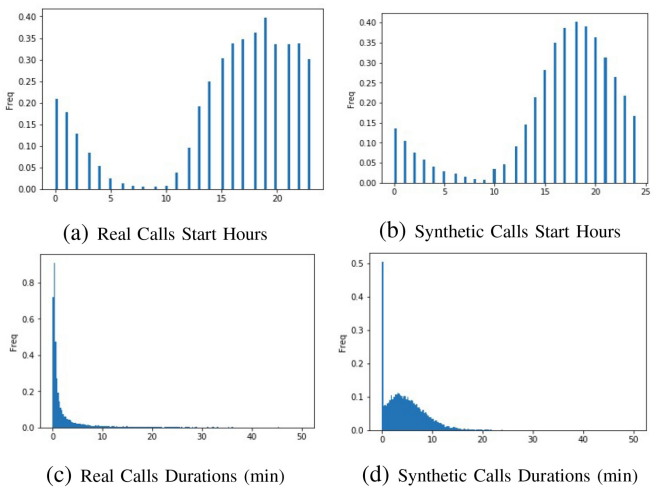


Fig. 8. Leveraging GAN for enriching the scarce training data [19].

A radio environment maps estimation algorithm leveraging a GAN-based pixel regression framework (PRF) for underlay cognitive radio networks using incomplete training data is proposed in [81], [82]. In these works, the authors first transform the radio environment maps estimation task into a pixel regression through color mapping. Then they extract helpful information from the incomplete training data, design a feature enhancing module for the PRF algorithm, which intelligently learns and emphasizes the important features from the training images. Finally, they train the PRF to reconstruct the radio environment maps in the target area. Three indicators are used to test the proposed algorithm: the visual display of the radio environment maps, the estimated power spectrum of primary users, and the average REMs estimating error against different numbers of secondary users. Results are bench-marked with IDW and Kriging with the exponential semi-variogram estimation.

Moreover, authors in [76], while drawing inspiration from image processing design a deep-learning architecture tailored to mobile networking, which combines Zipper Network (ZipNet) and GAN models. Using the open-source Telecom Italia's dataset [106], they infer fine-grained mobile traffic patterns to monitor city-wide mobile traffic via the GAN.

However, GANs suffer from many challenges, such as vanishing gradients, oscillations, modal collapse and the design of suitable evaluation metrics to evaluate their performance.

B. Autoencoders

Unlike GANs, which come in the class of implicit density methods (where the prior distribution of latent features is not known), some generative methods fall under explicit density method, meaning that the distribution of latent features is explicitly defined. One such method is a type of autoencoder, namely variational autoencoder (VAE). Autoencoders are basically neural networks consisting of an encoder and decoder, that encodes the input to a point in latent space, by performing non-linear dimensionality reduction (Fig. 9). The parameters of the encoder and decoder are optimized during training to minimize the reconstruction loss, as the autoencoder learns to reproduce its input. On the other hand, as illustrated in

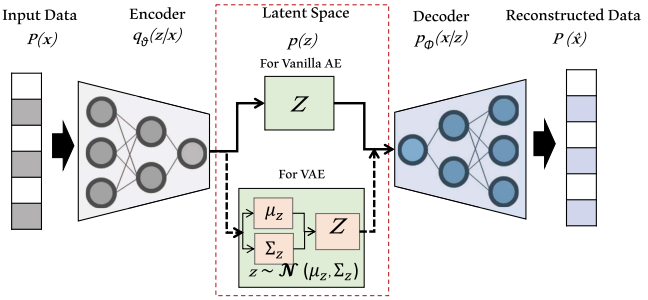


Fig. 9. A conventional vanilla autoencoder and variational autoencoder (whose internal representation is described by a probability distribution).

Fig. 9, variational autoencoders encode the input into a multivariate distribution (e.g., normal distribution) in latent space, described by the mean and variance vector where the length of the vector is equal to the number of dimensions in latent space. This probabilistic representation ensures that the latent space has good properties, such as variability of the latent space, thus making the model more robust and achieve better performance as compared to traditional autoencoders.

VAEs are used in literature [109], [110] to handle labeled training data scarcity problem for anomaly detection use-cases in RAN. In these use-cases labeled training data is severely imbalanced and traditional machine learning techniques are not able to distinguish the anomalies from the majority data. As a case study, authors in [109] used VAEs for anomaly detection and root cause analysis (RCA) in radio access networks. The data used in the analysis includes key performance indicators (KPIs) that indicate network quality of service (QoS), as well as key quality indicators (KQIs) that indicate user quality of experience (QoE). The anomaly detection module focuses on detecting the performance degradation in RAN, whereas the RCA module tries to find the root cause of detected anomalies. The proposed anomaly detection module takes time series of KPIs/KQIs from a cell as an input to the VAE model and outputs their respective anomaly score based on the error from the VAE model when it tries to reproduce its input. The RCA module is trained by auto-labelling the anomaly labels in a semi-supervised fashion using KQI rules, e.g., high PRB usage, over coverage, weak coverage, etc. The proposed AI-based approach is then tested in a live O-RAN compliant network for closed loop automation, resulting in 25% increase in downlink rate and 8% increase in RRC connection establishment with zero human cost in the entire process.

Similarly, adversarial autoencoders are a type of variational autoencoders which combines the architecture of autoencoders with GANs adversarial loss for regularization. Authors in [110] demonstrated the effectiveness of adversarial autoencoders for detecting anomalous behavior in wireless spectrum using power spectral density data. Manual spectrum management, especially in emerging dense and heterogeneous networks is inefficient and can only detect limited anomalies. Therefore, automated spectrum monitoring solutions are becoming more crucial than ever before. Along with anomaly detection, the proposed model in [110] shows a semi-supervised wireless band classification accuracy close to 100% on datasets using only 20% of the labeled samples.

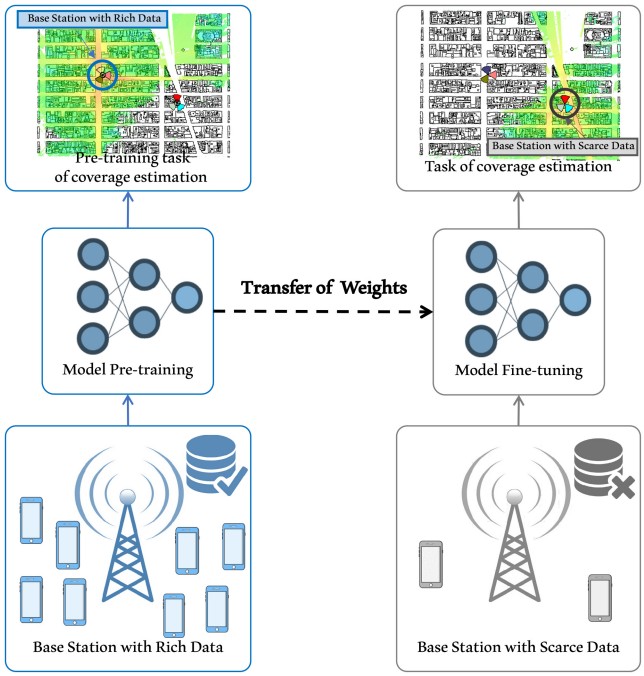


Fig. 10. An example of transfer learning in deep neural networks for coverage estimation. The feature network (source model) is pre-trained on a large dataset (from BS with rich data). The target model is created by transferring the knowledge learned from the source model, e.g., weights of the model. This model is then trained/fine-tuned using the scarce dataset (from BS with scarce data).

C. Transfer Learning

For data streams where latent features are too little to allow the use of GANs, matrix completion or other interpolation techniques identified above, the transfer-learning paradigm [84], [111] can be leveraged.

Transfer learning aims to help improve the learning of the target environment (target model) by transferring the knowledge learned from another similar environment (source model). One way of achieving that is by model fine-tuning, where a larger source dataset is used to pre-train a neural-network based model (source model) and fine-tuned using the target scarce dataset (as illustrated in Fig. 10).

In cellular network context, similarities among cells can be leveraged for determining when to use transfer learning. To quantify similarities among the cells, one approach is to use Wasserstein distance measure [112]. Given two random variables f_i and f_j with marginal distributions $P(f_i)$ and $P(f_j)$ respectively, let ψ denote the set of all possible joint distributions that has marginals of $P(f_i)$ and $P(f_j)$. Then Wasserstein distance between them is defined as:

$$W(f_i, f_j) = \inf_{P_{f_i f_j} \in \psi} \int |f_i - f_j| P_{f_i f_j}(f_i, f_j)\, d_{f_i}\, d_{f_j} \qquad (38)$$

The inf in Equation (38) gives joint distribution with f_i and f_j having smallest distance while maintaining the marginals.

Several works have been carried out in the literature using transfer learning to address data scarcity problem for network performance prediction [84], [85], [86], [87], [88], [89], [90]. As a case study, authors in [85] proposed to use transfer learning for parameter configuration in cellular networks. In this work, contextual bandit algorithm is leveraged along with transfer learning to optimize parameter configurations for uplink power control and user scheduling using cell KPI/counter data. Cell state measurements, e.g., the number of total users within the cell, the number of active users, the average channel quality indicator (CQI) of the cell, etc. are collected for each cell at each hour, and the goal is to minimize the ratio of users with experienced throughput less than 5Mbps for each cell. Live field tests in a real cellular network consisting of 1700+ cells show a significant performance improvement of 20% by optimizing five parameters for two weeks, thereby demonstrating the effectiveness of the proposed scheme.

A transfer actor-critic learning framework for energy saving in cellular radio access networks is proposed in [86]. This work utilizes the transferred learning expertise in historical periods or neighboring regions for predicting traffic load variations for BS ON/OFF switching. The problem of predicting the signal strength in the downlink of a real LTE network, where the antennas can be tuned to operate with different antenna tilt configurations is addressed using transfer learning in [87]. The authors show that augmenting the data from the source domain by adding data available from other tilts configurations of the same antenna improves the performance of the proposed transfer learning approaches. Transfer learning for channel quality and active UEs prediction is proposed in [88], using KPI/counter data from a commercial LTE network. The results show how transfer learning can be carried out across pairs of cells working at different frequencies, or at the same frequency in different locations and how to pick suitable candidate cells across the city for the transfer learning task. Transfer learning is also particularly helpful in tasks that require frequent model retraining, due to changes in the operational environment during execution, such as learning performance model for a cloud service [89]. Authors in [89] show that the number of new measurements required to compute a new model are reduced by an order of magnitude in most cases using transfer learning, as compared to training the new model from scratch, when evaluated on traces collected from a testbed running video-on-demand service, under various load conditions. However, finding suitable transfer candidates, or where to transfer is another challenging research question that remains unfocused in most of the works discussed earlier. Authors in [90] argue that the choice of source domain can either yield 'transfer gain', or further decrease the performance of the baseline model, commonly known as 'negative transfer', and proposed two source selection approaches to mitigate this issue. A key result from their study is that source selection should encourage diversity of the data in source domain rather than similarity between source and target cell, especially in scenarios with few samples in target domain as the similarity between the underlying distributions of both domains cannot be reliably measured.

D. Few-Shot Learning

Few-shot learning (FSL) is another branch of machine learning that addresses the performance degradation problem of deep learning algorithms when the training dataset size is small. Using prior knowledge, FSL can master new tasks from

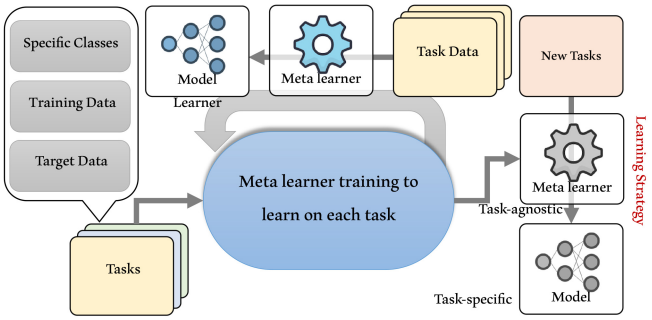


Fig. 11. Meta learning-based methods can learn a learning strategy from a family of tasks by developing a task-agnostic learner. The learning strategy (or task-agnostic knowledge) can then be used to improve the learning of a new few-shot learning task from that task family [128].

a limited number of examples [127]. This type of learning is primarily motivated from the ability of humans to learn from only a few examples. Therefore, FSL can eliminate expensive data collection efforts and help in building suitable models for rare cases of limited supervised data [127].

FSL can be used for classification, regression and even reinforcement learning tasks using only few labeled, input-output and state-action examples respectively. However, the most common application scenario for FSL is *"N-way-K-shot classification"*, where a classifier is built for distinguishing between N classes, each having only K examples per class. When only one example with supervision is available, it is referred to as One-Shot Learning and when no example is available, it is called Zero-Shot Learning.

FSL is a very active area of research these days and the methods being proposed in the literature for solving the few-shot problem can be broadly classified in two different branches: 1) Meta learning, and 2) Metric learning. The key idea in Meta learning-based methods (as shown in Fig. 11) is to distill the experience of multiple learning episodes from a distribution of related tasks. This learning to learn strategy can improve the future learning performance on new few-shot learning tasks, thus developing a task-agnostic learner with improved data and compute efficiency [128], [129]. Examples of methods include Model Agnostic Meta Learning [130], Task-Agnostic Meta Learning [131] and Meta-transfer Learning [132]. These methods are good at out-of-distribution tasks and can handle varying and large shots well, but their model and architecture are intertwined and their optimization process is challenging [133]. On the other hand, Metric learning-based methods learn to compare query set (test set) with support set (few-shot training set) by learning transferable representations in semantic embedding space using a distance loss function (learn to compare). Examples include Siamese Neural Networks [134], Matching Networks [135], Prototypical Networks [136], Relation Networks [137] and Graph Neural Networks [138]. As compared to meta learning-based methods, these are relatively simple, entirely feedforward, computationally fast and easy to optimize, but harder to generalize to varying shots and to scale to very large shots [133].

A few works have been carried out using few-shot learning to address training data scarcity issue in cellular networks. Authors in [78] use prototypical networks, a few-shot learning-based algorithm for performance metrics analysis in LTE networks. They used eNodeB trace data from live network and classified individual eNodeBs into different performance classes based on their KPIs. Their results show an improved performance as compared to baseline DNN, 1-D CNN and 2-D CNN.

Authors in [79] show that meta learning can be used in mmWave smart factory environment to frame the indoor pathloss prediction task as a meta-task comprising of multiple tasks. Authors show that meta-learning based CNN-based model trained on a meta-task of multiple beams can outperform conventional training methods. Specifically, the prediction RMSE of the proposed meta-learning based CNN model show a gain of 70% in terms of prediction accuracy as compared to floating-intercept (FI) model, and a gain of 55% as compared to conventional CNN based model.

Authors in [139] use self-imitation via transfer learning to achieve few-shot learning for the resource management (network power minimization) problem in Cloud Radio Access Networks (C-RAN). Their simulation results show that few-shot learning is able to achieve similar performance even with scarce and unlabeled training data, as compared to a model that is trained without few-shot learning even with labeled data. These results show the power of few-shot learning in scenarios where labeled training data is not available or is very scarcely available.

E. Lessons Learned

Based on the covered literature, we can see that all the above-mentioned ML/DL techniques work well for modeling high-dimensional datasets, however, they differ in terms of their applicability. For instance, both GANs and autoencoders can only generate quality synthetic data if their training data contains some latent information about their environment. In situations where the scarce dataset is not representative of the environment from which it is collected, few-shot learning and transfer learning techniques can be used. Both, however, rely on the availability of auxiliary datasets to help them learn the target environment from unrepresentative training data. Transfer learning requires data from a similar domain or task to gain insights and then transfer that knowledge to the task at hand. few-shot learning requires data from a lot of different (but not necessarily similar) task/domain to learn the unfamiliar environment. These takeaways are also illustrated in Fig. 17 for the benefit of the reader.

V. Synthetic Data Generation

The techniques mentioned in previous sections are likely to work well when the scarce available data is somewhat representative of the whole data or exhibits some degree of correlation. In situations where the available data is scarce and non-representative, the methods presented in preceding sections are likely to perform poorly. Likewise, in other scenarios, the available data can be big, but still not representative. In these cases, the solution lies in either resorting to get real

TABLE II
COMPARISON OF DIFFERENT SIMULATORS FOR SOLVING DATA SCARCITY PROBLEM

Feature	Simulator													
	OpenAirInterface GTEC [113]	5G-K [114]	X. Wang et al. [115]	V.V.Diaz et al. [116]	ns-3 [117]	OMNeT++ [118]	NYUSIM [119]	MATLAB/SIMULINK [120]	C-RAN [121]	OPNET [122]	Vienna 5G [123]	Atoll [124]	SyntheticNET [125]	SyntheticNET [126]
Scheduling support	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓
mm-Wave support					✓	✓	✓	✓		✓		✓	✓	✓
Adaptive numerology									✓			✓	✓	✓
QCI support									✓				✓	✓
Parallelized offline traces and time-independent KPIs pre-generation for reduced online computational cost													✓	✓
Realistic antenna patterns modeling											✓		✓	✓
Signaling overhead modelling														✓
Realistic mobility modeling														✓
AI based pathloss modeling														✓
500+ COPs modeling														✓
Realistic HO management														✓
Realistic mobility pattern														✓
Python based to enable data processing and easy incorporation of ML libraries														✓
Free license*	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓

data or generate synthetic data. In this section, we will present ways to generate synthetic data through simulators.

A. Simulators

System level simulators are widely used in both industry and academia due to limitations of analytical models and field experiments. Apart from the limitation of mounting Base Stations (BSs) on predefined locations, the support of antenna height, tilt, transmission power etc. for individual BSs is absent in the analytical model. Furthermore, stochastic geometry-based models are unable to capture the network dynamics which include mobility management and transmission latency. On the other hand, field trials exhibit the most realistic modeling of network performance, evaluation and tuning. However, this approach is impractical owing to the cost and time effort required to conduct field trials on a large scale, and with the high probability of significant network performance impairment of live mobile network during the trial phase.

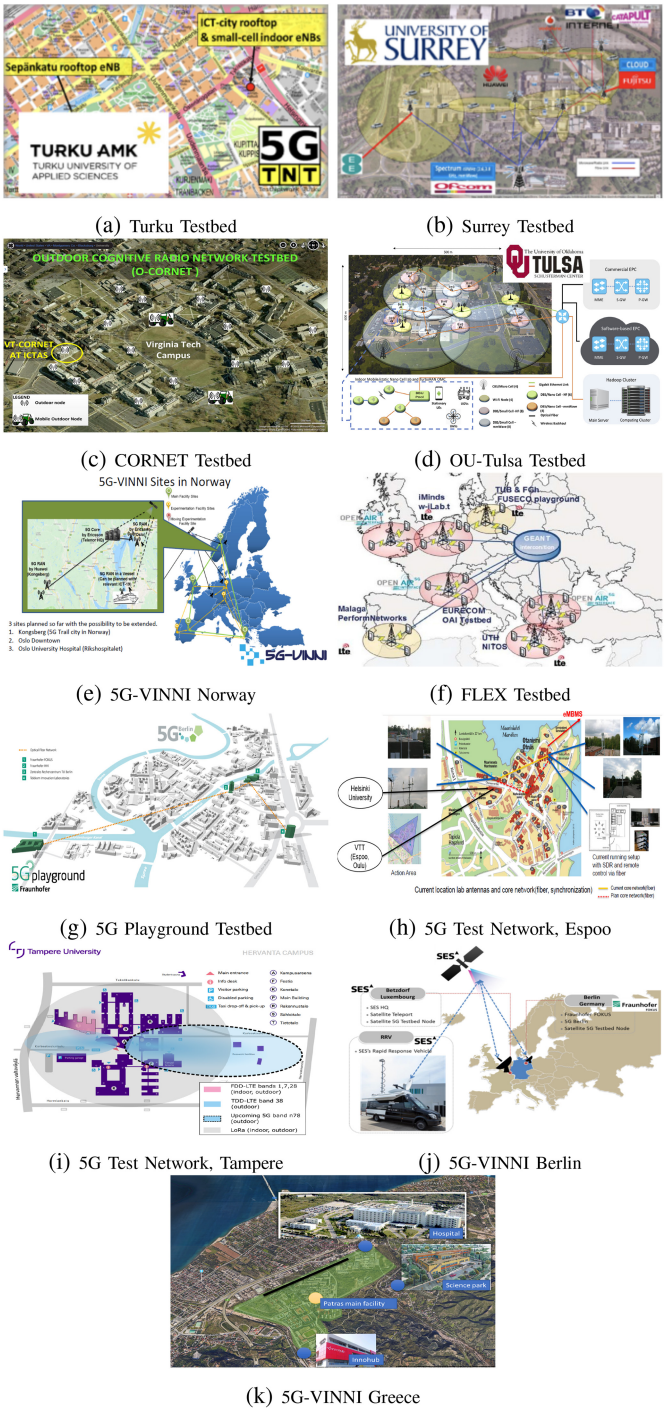
A list of existing simulators along with a comparison of their features is presented in Table II. For more details on these simulators, the reader is referred to two existing surveys on simulators; [28] that compares 4G and 5G simulators, and [29] that gives the summary of the most significant 5G simulators.

As observed from Table II, none of the simulators is based on comprehensive 5G standard incorporating all aspects outlined in the standard. To tackle this problem, SyntheticNET simulator built on Python platform was developed by the AI4Networks Research Center at the University of Oklahoma [126]. The SyntheticNET simulator is modular, flexible, microscopic and versatile, built-in compliance with the 3GPP Release 15. This simulator supports features like adaptive numerology, actual hand over (HO) criteria and futuristic database-aided edge computing to name a few. Instead of an objected-oriented programming (OOP) based structure like existing simulators, SyntheticNET simulator supports commonly used database files (like SQL, Microsoft Access, Microsoft Excel). Site info, user info, configuration parameters, antenna pattern etc. can be directly imported to the simulator. As a result, the simulation environment is more realistic and closer to actual deployment scenarios. For further details of this simulator, the reader is referred to [126].

Python based platform and the flexibility of different input and output data formats in SyntheticNET simulator can assist in solving the data scarcity challenge by generating ample amounts of synthetic data to enrich the available scarce real data, which can then be used to implement different Self Organizing Networks (SON) related features or AI based network solutions [1]. Mobile operators can use it for planning, evaluating or even optimization of beyond 5G networks. Research community can also benefit from it by implementing the new ideas on data generated from this 3GPP-based realistic 5G network simulator.

Fault diagnosis using synthetic data from Atoll simulator is used in [140]. Authors in [140] consider 4 types of faults characterized by cell outage, low transmit power, excessive antenna uptilt, and excessive antenna downtilt. The SINR maps obtained in these scenarios are scarce as shown in Fig. 14. Authors in [140] then analyse the performance of several ML-based algorithms for fault diagnosis in Fig. 15, where the UE density on x-axis corresponds to the network depiction in Fig. 14. As compared to complete coverage maps, a drastic drop in diagnosis accuracy is observed for the ML models on scarce data, where the exact match ratio (EMR) drops from 90.2% to 69% and from 92% to 71.3% respectively, as the density of users drops from 203 to 100 users/cell. Performance continues to deteriorate as the number of users decreases per cell.

Another example of data generated through simulators include system features data (such as BS horizontal/vertical



(a) Turku Testbed (b) Surrey Testbed

(c) CORNET Testbed (d) OU-Tulsa Testbed

(e) 5G-VINNI Norway (f) FLEX Testbed

(g) 5G Playground Testbed (h) 5G Test Network, Espoo

(i) 5G Test Network, Tampere (j) 5G-VINNI Berlin

(k) 5G-VINNI Greece

Fig. 12. Some current and emerging 5G testbeds.

separation, transmit power, operating frequency, antenna beamwidth and gain) and environment features (such as propagation distance, clutter types, BS height, diffraction points, number of building penetrations in each clutter type) to create a machine learning based prediction model for 3D pathloss and received signal strength (RSS) [80] to overcome the challenges of conventional and ray tracing based path loss modeling. This work investigated the model performance under varying data scarcity levels (UE density). Fig. 16 is a key numerical result from this study, which shows how the augmentation of scarce training data (from 400 UE traces/km^2 to 20,000 UE traces/km^2) leads to significant reduction in RMSE (RSS prediction error) for most ML algorithms used for path loss and ultimately RSS prediction.

Another simulator generated data in [141] includes the dataset of RSRP, SINR, and handover success rate (HOSR) against the rarely explored mobility configuration and optimization parameters, namely A5 time to trigger, A5 threshold 1 and 2. The A5 parameters are usually fixed to a gold standard value or adjusted through hit and trial due to the valid reluctance of network operators to test all parameter combinations in the live network. To overcome this issue, synthetic data from a 3GPP-compliant simulator was generated. This type of data was then used to develop a closed loop solution for optimizing seldom explored A5 parameters by jointly maximizing RSRP, SINR and HOSR [141].

B. Lessons Learned

Synthetic data using simulators can be used to augment data in situations where the available data is non-representative. Simulators are also a good candidate to generate training data for transfer learning or meta-learning techniques. Although most simulators are link level, system level simulators are also there. The choice of simulators depends on what features (e.g., scheduling support, mmWave, adaptive numerology, mobility and pathloss modeling, COPs, etc.) are supported and Table II can assist the reader for this purpose. Based on the available literature, SyntheticNET has the most features supported.

VI. Real Data Generation

The preceding techniques, with the exception of using simulators, are likely to work well when the scarce available data is somewhat representative of the whole data or exhibits some degree of correlation. In situations where the available data is scarce or big but non-representative, the solution lies in obtaining real data.

One way of getting access to real data can be utilizing historic logs of data gathered by other researchers. However, these logs might become outdated quickly with the emergence of new technologies, heterogeneous deployments or change in traffic patterns, number of users, construction of buildings and other terrain changes. Another way of generating real data can be through the use of mobile phone applications. However, what if researchers require data for scenarios which are not yet deployed in a real network? The techniques presented in previous sections (except simulators), all require some starting real data but with the advent of AI based next generation networks, there exists the potential of new or anticipated scenarios which do not exist in a real network. In such cases, testbeds to generate real data are going to be the best option for wireless communications community.

A. Phone Applications and Parametric Subscriber/Third-Party Data

Many smartphone applications offer the ability to log parameters such as RSRP, RSRQ, SNR, events occurring (handover, cell re-selection), serving time, speed, height,

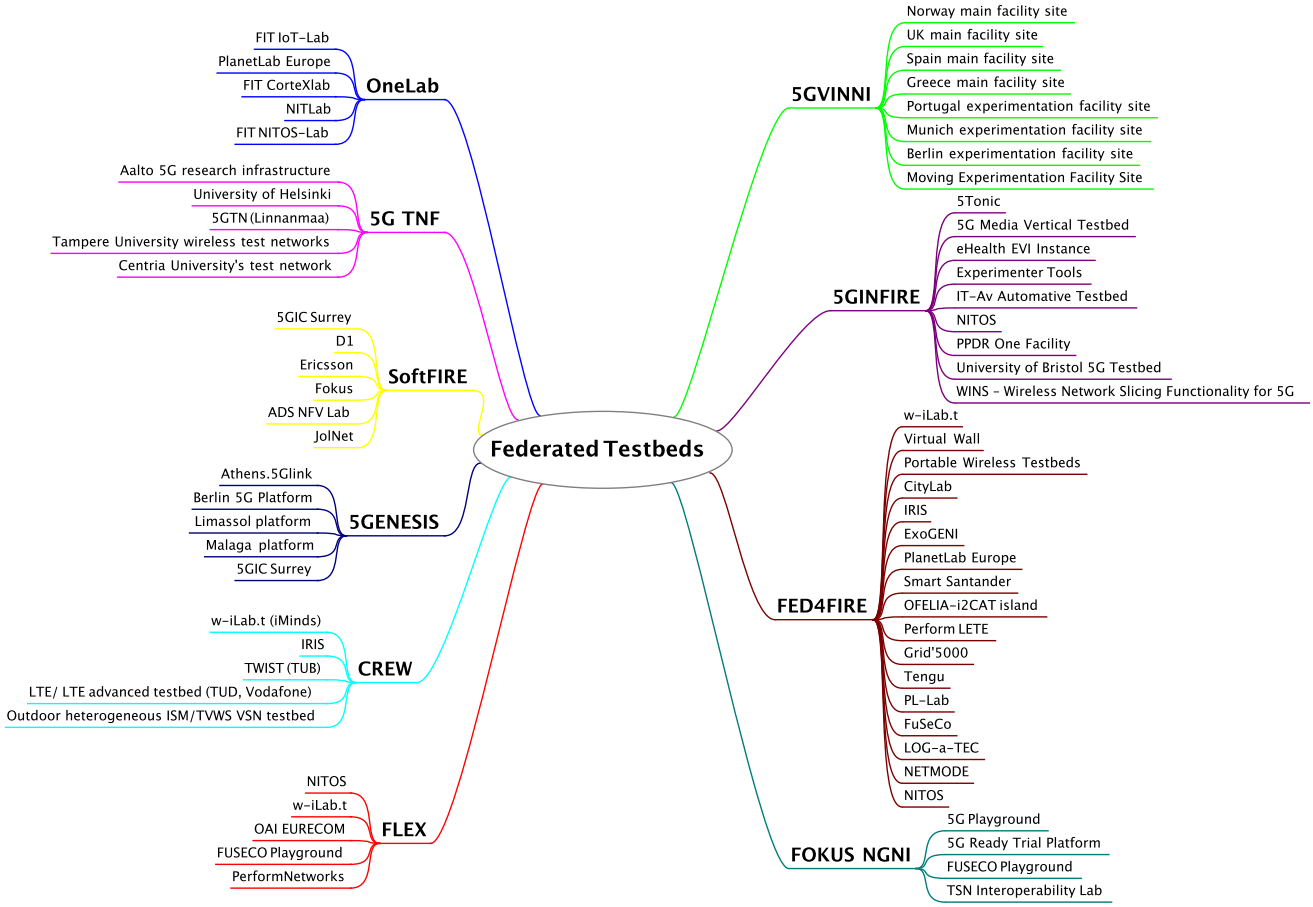


Fig. 13. Federated Testbeds.

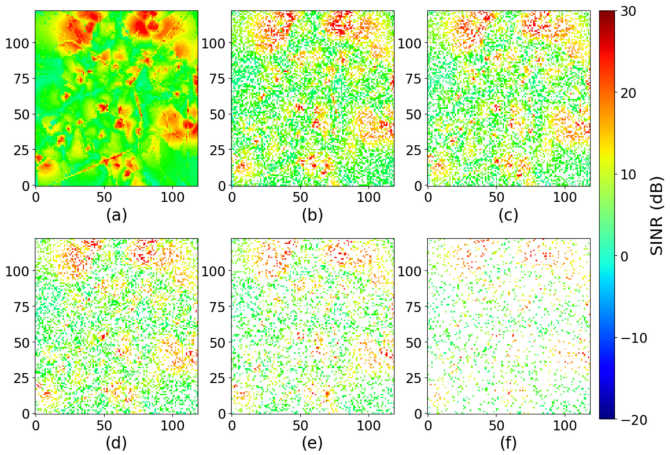


Fig. 14. Network coverage maps with various user densities (a) Full coverage map (203 UEs/cell) (b) 100 UEs/cell (c) 80 UEs/cell (d) 60 UEs/cell (e) 40 UEs/cell (f) 20 UEs/cell [140].

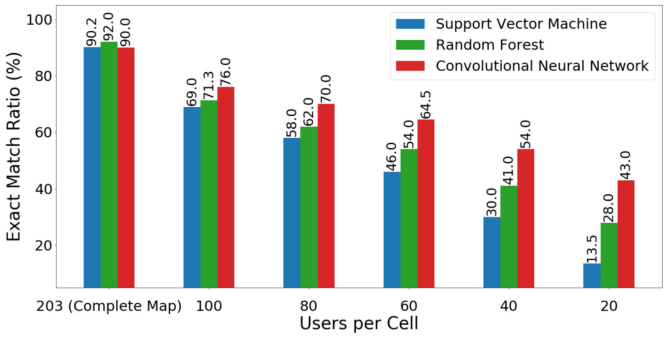


Fig. 15. Performance comparison of ML models on scarce and complete coverage maps data [140].

cell ID, along with timestamp and location (latitude, longitude) information). As an example, one of the studies [108], used a novel methodology of utilizing smartphone application, based on the idea of participatory sensing, to collect real LTE network data for building, training and evaluating the performance of mobility prediction schemes in live network [108]. The data in this case was the handover information of the user. An android application, "LTE Discovery" was installed on the smartphone to log the timestamp and new cell IDs around the OU-Tulsa campus. This information was then used to build a semi-markov model for mobility prediction.

The quality of data gathered through smartphone applications, however, depends on a number of factors, including measurement capabilities of different smartphones and GPS error inaccuracy for measuring heights and positions. Smartphones equipped with barometers are likely to give a better estimate of heights in scenarios with varying terrains. In addition, transmitter parameters, such as type of antennas and their characteristics

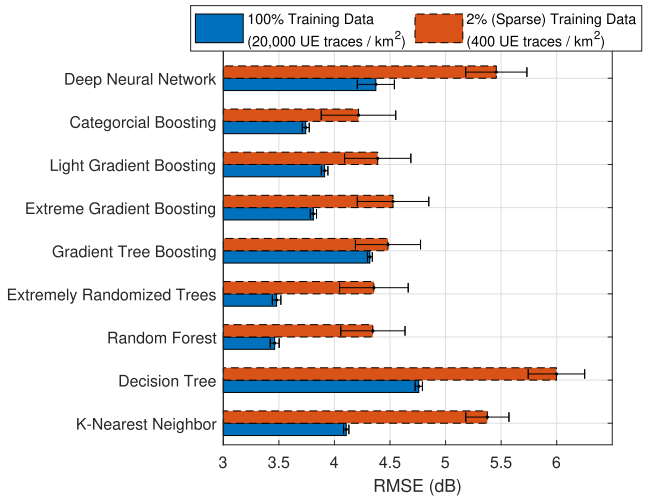


Fig. 16. Comparison of RSS prediction error when the ML based prediction models are trained using scarce and enriched synthetic data. Height of bars represent the mean value and error bar represent the standard deviation using 5-fold Repeated Cross Validation. Enriched synthetic data leads to a reduction in RSS prediction error (RMSE) [80].

remain unknown, unless the network operator is involved. When the network operator is involved, it is possible for the subscriber to obtain parametric data from them. However, that type of data may be limited to a certain number of possible configurations. For this reason and for potential new scenarios, the solution may lie in resorting to testbeds.

B. Testbeds

Field trials using testbeds generate real training data and provide the most realistic picture of the network. An aerial view of some of these testbeds is presented in Fig. 12. We have summarized the existing and emerging testbeds in Table III to make readers aware of current and emerging platforms to access real data. Most of these testbeds are open, i.e., available to external experiments. This will foster collaboration among different academic institutions as well as with industry, which will in turn enable the utilization of these existing facilities to the fullest and accelerate quality research in the field.

Apart from individual testbeds, several federations or consortiums of testbeds have been formed around the world. Some key federated testbeds comprising of the testbeds in Table III are presented in Fig. 13.

Examples of data collected from testbeds include data for scenarios that are not fully and widely deployed yet, e.g., mmWave channel measurement data consisting of direction of user movement with respect to BS-UE link, distance resolution, the number of user locations and whether blockage is present or not [142]. This type of data can be used for building beam tracking algorithms. Other examples of data include received signal strength indicator, electric vector magnitude, packet and bit error rate data from CORNET testbed [143] and massive MIMO data from LuMaMi testbed such as signal to noise ratio (SNR) and bit error rate for different antenna configurations and modulation schemes [144]. These types of data can provide flexibility to researchers for design and testing network scenarios using a much wider range of parameters, which is difficult to obtain from network operators otherwise, due to the high probability of network impairment when varying parameters too much in live networks.

C. Lessons Learned

One way of getting access to real data to augment scarce data can be utilizing historic logs of data gathered by other researchers. However, these logs can become outdated. Lack of diversity in the COP-KPI data is another problem when data is obtained through logs. Testbeds is another way to generate real data and is particularly useful to test new or anticipated scenarios which do not exist in a real network. Key features of several federations and individual testbeds around the world have been presented in Table III that can assist the readers in the choice of testbed for their works.

VII. Conclusion and Discussion

In this paper, we have presented an overview of key techniques in literature to address the data scarcity challenge and presented some emerging new techniques that can be applied to radio access networks in the wireless communication domain to solve this problem.

Table IV summarizes the data augmentation techniques for handling scarce datasets in mobile networks. The typical use cases targeted in existing literature include mobile traffic maps generation using scarce CDR data, spectrum sensing, MDT-based outage detection, CSI/RSS for localization, BS trace data for performance analysis, network power minimization, optimizing BS Tx power using UE SINR data, network parameter configuration optimization for power control and user scheduling, resource allocation, traffic load based energy saving, CQI and RSS prediction, radio environment map reconstruction, channel estimation in Massive MIMO systems and discovering user patterns using user trajectory data. The tools in existing literature to address these use cases include GANs and its variants, transfer learning, autoencoders, interpolation techniques, simulators and testbeds. While these techniques have proved to be beneficial for particular use cases, the generalization ability of a particular technique to different scenarios remains a challenge. Another notable challenge is the applicability of these techniques to highly dynamic or mobile environments. Efforts are also being made to reduce the training time of machine learning based models and modifying them for more robustness.

It should be noted however, that the success of any technique for solving the data scarcity challenge depends on a number of factors, including type of data under consideration, number of transmitter and receivers, distributions of users and base stations in a given area, distribution of measurement data, level of accuracy required, measurement capability of receivers, dynamics of propagation environment, propagation modeling accuracy, time and computational resources available. Also, highly dynamic spatio-temporal environment would greatly hamper the outputs of techniques covered in this paper. In that case, using data through simulations and testbeds may provide the best option. Further options on addressing the data scarcity challenge for highly dynamic

TABLE III
WORLDWIDE EXISTING AND EMERGING TESTBEDS FOR SOLVING DATA SCARCITY PROBLEM

Testbed	Location	Key Features
NITOS [149] [150]	NITlab, University of Thessaly (UTH), Volos, Greece	- Open (facilities available to external experimenters) - Over 100 wireless indoor and outdoor nodes - 45 nodes equipped with a mixture of Wi-Fi and GNU-radios - One Cloud installation with 200-cores - Multiple wireless sensor network deployments - Cameras, temperature and humidity sensors - Software defined radio testbed with 10 USRP devices - Two programmable robots provide mobility - WiMAX/3G/LTE technologies - 5G virtual infrastructure provisioning by 5GINFIRE [151]
6GIC [152] - [155]	ICS, University of Surrey, Guildford, UK	- 4G LTE, 5G NR, 6G (ongoing) - 4km^2 comprising indoor and outdoor environments - Outdoor: 4G ultra-dense C-RAN comprising 3 macro cells, 39 LTE-A TDD small-cell sites, operating at 2.6 GHz, 1x 4G FDD site operating at 700 MHz, 8x 5G NR TDD sites, operating at 3.5 GHz - Indoor: 6x TDD and 6x FDD cells over 2 floors, and Wi-Fi APs - 28 GHz (PtP), 60GHz (PtMP) mmWave and satellite backhauling also supported - Core Network supports separate 4G and 5G core segments - Supports broadband mobile radio - Fixed core network and service platform based on software defined networking - Supports Internet of Things
ORBIT [156] [157] [158]	WINLAB, Rutgers University, USA	- Open: available for remote or on-site access - Radio grid with 20x20 two-dimensional grid of programmable radio nodes - Outdoor ORBIT network provides a configurable mix of both high-speed cellular (WiMAX, LTE) and 802.11 wireless access - SANDBOX networks used for debugging and controlled experimentation - Software defined networking (SDN) resources - Cloud resources
PhantomNet [159] [160]	Flux Group, University of Utah, USA	- Remotely accessible and sharable - Mobility testbed - Built on top of Emulab - EPC/EPS software (OpenEPC), hardware access points (ip.access eNodeB), PC nodes with mobile radios (Nexus 5 phones and SDR-based) - Provides configuration directives and scripts
LuMaMi [161] [144] [162]	Lund University, Sweden	- Real time 128-antenna MIMO test bed - National Instruments USRP RIO SDRs - LabVIEW system design software and PXI platforms - Mobile base stations - Used for channel sounding, high speed data streaming, evaluation of baseband solutions, assessing circuit design - Demonstrated mobile multi-user tests with University of Bristol [163]
Firecycle [164] [165]	Intrusion Detection Systems Group, Columbia University, USA	- Scalable test bed for large-scale LTE security research - Implement, test, analyze impact of security attacks against LTE mobility network - Prototyping and testing attack mitigation strategies for future cellular networks - Implemented on OPNET
Berlin LTE-A [166] [167] [168]	Center of Berlin, operated from Fraunhofer HHI, Deutsche Telekom Laboratories and University of Technology, Berlin	- 3 base station sites with 9 sectors - Incorporates LTE key features: frequency dependent scheduling in 20 MHz bandwidth, adaptive MIMO mode selection for 2x2 MIMO utilizing spatial multiplexing, and low round-trip delay on the PHY layer of 8 m

(continued)

environments is out of the scope of this work and can be considered as part of a future study. Therefore, while a certain technique might work well in a particular scenario, it is likely to perform poorly in other scenarios. It should also be noted that the selection of a performance metric to assess the accuracy of a particular method is important

TABLE III
(Continued) WORLDWIDE EXISTING AND EMERGING TESTBEDS FOR SOLVING DATA SCARCITY PROBLEM

CEWiT LTE and 5GNR [169]	IITMadras Research Park, Chennai, India	- 2 types of testbeds based on: 1) CEWiT hardware 2) TI's multi-core DSPs - Hardware is made using SDR radio nodes - LTE PHY for UE and eNB has been developed in collaboration with IITM - Basic implementation of LTE L1 downlink and uplink chains - L2 MAC, RLC and a thin layer of PDCP - Both eNodeB and UE implementations - End-to-end IP application flow both in DL, UL - Supports 3GPP Release 8 specifications - Supports up to 10 MHz bandwidth and can be extended to 20MHz - 5G NR for sub 6GHz and mm wave under development
TitanMIMO-6 [170] [171]	Nutaq, Qubec, Canada	- Sub 6 GHz wideband Massive MIMO testbed - FDD+TDD capabilities - Up to 56 MHz real-time baseband processing - Radio tumble up to 5 GHz - Nutaq's SDR systems (PicoSDR) can be combined with TitanMIMO system to build up complete HetNet, MUMIMO or CRAN testbed solutions - Enabling evaluation of interoperability behavior for various deployment scenarios
Aalto 5G research infrastructure [172]	Otaniemi, Espoo, Finland	- Network slicing - Support for NB-IOT to be used for IoT hackathon - Mobile and edge computing, VR/AR, Gaming, Industrial Internet - Part of 5G TNF
University of Helsinki Test Network [173]	University of Helsinki, Kumpula campus, (Exactum building), Finland	- 17 Nokia Flexi Zone Indoor Pico BTS (eNBs) - Band: 2600 MHz (E-UTRA 7) FDD - Sync: 1588v2 (PTP) / GPS / Sync-E - 3 connections to cores through VLANs: UH core(s), Aalto core and Nokia core - Part of 5G TNF
VodaPhone Chair [174] [175] [176]	TU Dresden, Germany	- Online Wireless Lab (OWL) testbed - Software Defined Reconfigurable Radio Devices - LabVIEW/LVC in combination with USRPs - Many projects and startups, e.g., 5G Lab Germany, 5GNetMobil, 5G Picture, HPE-5G-Testbed, Airrays GmbH [177]
CORNET [178] [179]	Virginia Tech University, USA	- University-wide testbed - Software-defined radios, cognitive radio and dynamic spectrum access - 48 indoor SDR nodes, 14 fixed outdoor nodes, 6 mobile units (O-CORNET) - A few LTE-capable nodes (LTE-CORNET) - CORNET nodes are remotely accessible - Awarded the grant from DURIP for upgrading to LTE and LTE-A - Outdoor network of 15 radio nodes and 2 mobile nodes
5G Playground [180]	Fraunhofer FOKUS and TU Berlin campus, Germany	- Empowers the 5G Berlin testbed - Support for multi-slicing - Ultra-reliable, low latency communication in Industrial IoT lab of FOKUS - Automotive testbed environment in underground parking of FOKUS building - Coverage of dense urban areas, like portable 5G edge nodes in progress - 3 Toolkits: Open5GCore, OpenSDNCore and Open5GMTC
Tampere University Wireless Test Networks [181] [182]	Tampere University, Hervanta, Finland	- Part of 5G TNF - FDD-LTE operating at band 1, 7, and 28 for mostly indoor coverage - TDD-LTE operating at band 38 to provide campus wide outdoor test network - Upcoming outdoor 5G test network in band n78 with 60 MHz channel - LoRa: Digita's LoRaWAN test network in ISM band at 868 MHz
FUSECO Playground [183]	Fraunhofer FOKUS Institute, Berlin, Germany	- Open IMS Core solution - Heterogeneous indoor and outdoor radio access technologies - DSL/WLAN/2G/3G/4G-LTE/LTE-A and soon 5G - M2M communication, IoT, sensor networks - SDN/OpenFlow, NFV cloud environments - Toolkits: Open5GCore, OpenSDNCore and Open5GMTC, OpenMTC, Open Source IMS Core, OpenStack-based Cloud Testbed, OpenXSP

(continued)

too. As an example, if the metric of mean residual error is used to access Kriging accuracy, it would always yield zero, since this type of interpolant satisfies the unbiased-ness condition, and so some other performance metric, like the average relative error would be more appropriate in this case.

TABLE III
(Continued) WORLDWIDE EXISTING AND EMERGING TESTBEDS FOR SOLVING DATA SCARCITY PROBLEM

5G Ready Trial Platform [184]	Fraunhofer FOKUS, Berlin, Germany	- Consolidated turn-key solution of the Fraunhofer FOKUS software components - Addresses trial needs of emerging network infrastructures - - Edge Instantiation: solution for micro-operators and local networks, provides customized IoT connectivity for x100 devices. - Data Center Instantiation: multi-slice environment, support for multiple parallel instances of IoT and multimedia communication - Technology Elements: Virtual Core network, Network slicing, IoT support, Low delay network, Dynamic spectrum access and management
Ericsson 5G [185] [186]	Ericsson, Stockholm, Sweden	- Live testing of key capabilities, such as multipoint connectivity with distributed MIMO and 5G-LTE dual connectivity - 5G devices and base stations operate in 15 GHz band - TDD and OFDM - Up to 256 QAM modulation in downlink and up to 64 QAM in the uplink - mm-Wave testbeds 15 GHz and 28 GHz - Bandwidth is 80 MHz, centered at 3.5 GHz - Massive MIMO antenna array of 128 cross-polarized antennas
SK Telecom 5G Playground [187] [188] [189]	SK Telecom R&D Center, Bundang, Korea	- Developing a centimeter-wave (cmWave) 5G radio system with Nokia - 5G 3D system level simulator with Nokia and Ericsson - 3D beamforming techniques with large scale array antennas with Samsung - Developing Anchor-Booster Cell and Massive MIMO with C-RAN with Intel - Achieved 19.1Gbps transmission speed over the air - Futuristic services including 4K live broadcast system and AR/VR
5GTN (Linnanmaa) [190] [191] [192]	University of Oulu and VTT Technical Research Centre of Finland	- Multi-access edge computing - Core network in cloud environment - Cloud systems for applications - Secure connection to other 5G sites worldwide, 10 Gb VPN - Part of 5G TNF
TurboRAN [193]	AI4Networks Research Center, University of Oklahoma, Tulsa, USA	- Developing first end to end programmable cellular test bed for enabling AI based SON research towards 5G and beyond - Complete integrated mobile cellular network over 300,000 m^2 area - Tier 1: 4 outdoors macro cells on 1.2-6 GHz HF band - Tier 2: 16 small cells (programmed to pico or femto cells). 8 small cells can operate on the HF band, other 8 can operate on the unlicensed mmWave - Both tier cells are programmable - Both tier cells connected to EPCs and a big data processing Hadoop cluster - Hadoop cluster: 1 high performance master node, 15 slave nodes with high-capacity data modems - Support both high mobility and low mobility users
OAI [194]- [197]	EURECOM, France	- Open-source platform - 8-node testbed, equipped OAI compatible RF front-ends, UEs and VMs - 4 machines that can be used for running OAI as eNodeB - 4 nodes that are equipped with COTS UEs - 2 physical layer emulation modes - 64 antenna Massive MIMO testbed
Munich [198] [199]	TU Munich, Munchen, Germany	- 5G RAN with two sectors, each having carrier frequency: 3.4 GHz, bandwidth: 40 MHz, transmission power: 5 W antennas: up to 8 - 5G Mobile Terminals with vehicular speeds up to 50 km/h, enablingV2X - 5G Core network: HW/SW platform - Hardware: in-house platform of several dozen servers representing a data centre - Software: extended network emulators, controllers, open-source and proprietary switch implementations - Testbed can deploy virtual networks with different topologies as needed - 5G Core network supporting functional split SDN NFV Orchestration - Distributed data centres for mobile edge computing use cases

(continued)

Finally, based on the analysis from literature and domain knowledge, in order to assess the applicability of a particular method, the tree diagram in Fig. 17 is aimed to assist researchers and network operators in choosing the appropriate techniques based on available information. We start the figure by the red box, 'Insufficient data'. The first question

TABLE III
(Continued) WORLDWIDE EXISTING AND EMERGING TESTBEDS FOR SOLVING DATA SCARCITY PROBLEM

Perform Networks [200] [201] [202]	University of Malaga, Spain	- T2010 conformance testing units by Keysight Technologies - LTE release 8 small cells (Pixies) by Athena Wireless working on band 7 - Polaris Core Network Emulator - Several LTE UEs, working on different bands - ExpressMIMO2 and USRP SDR cards - SIM cards from an Spanish LTE operator to be used on commercial deployments
Centria's Test Network [203]	Centria University of Applied Sciences,Ylivieska, Finland	- TDD-LTE operating at band 40 and 42 for both outdoor and indoor coverage - Upcoming 5G test network in band n78 with 60 MHz channel outdoor network - Implementation plan of first 5G Non-Standalone during 2019 - Later 5G Standalone during 2020 - Part of 5G TNF
w-iLab.t [204] [205] [206]	Ghent and Zwijnaarde, Belgium	- w-iLab.t Office testbed: three 90 m x 18 m floors of iMinds office in Ghent - w-iLab.t Zwijnaarde testbed: 5 km away from w-iLab.t Office in Zwijnaarde - Sensor nodes, Wi-Fi based nodes, sensing platforms, and cognitive radio - Heterogeneous wireless/wired experiments - Virtual Walls: Virtual Wall 1 and 2 containing 206 and 159 nodes respectively - OpenFlow experiments - 20 programmable moving robots
5TONIC [207]	Madrid, Spain	- 9 members: Telefonica, Institute IMDEA Networks, Ericsson, Intel, Commscope, Universidad Carlos III de Madrid, Cohere Technologies, Artesyn Embedded Technologies and InterDigital - NFV orchestrator, implemented with Open Source MANO (OSM) - Dedicated NFVI for 5GINFIRE: 3 server computers, each with six cores, 32GB of memory, 2TB NLSAS, network card with 4 GbE ports, DPDK support - Second NFVI: 2 high-profile servers, each equipped with eight cores in a NUMA architecture, 128GB RDIMM RAM, 4TB SAS and eight 10Gbps Ethernet optical transceivers with SR-IOV capabilities
University of Bristol 5G [208]	University of Bristol, England	- Multi-site network connected through a 10 km fibre - Core network is located at HPN Lab at the University of Bristol - Extra edge computing node is available at Watershed - Access technologies are located at Millennium Square for outdoor coverage and We The Curious science museum for indoor coverage - Multi-vendor SDN enabled packet switched network - SDN enabled optical (Fibre) switched network - Nokia 4G and 5G NR - Self-organising multipoint-to-multipoint wireless mesh network - LiFi Access point, Cloud and NFV hosting - 2 different NFV orchestration and management solutions: Open Source MANO , NOKIA CloudBand - 2 cloud/edge computing solutions:Openstack Pike, Nokia MEC - 1 SDN controller: NetOS
D-15 Labs [209]	Ericsson, Santa Clara, CA, USA	- Validation and development platform for 5G use-cases, leverages cloud edge support, core network, and AI-based management and orchestration
ENCQOR 5G [210]	Ontario Region, Canada	- iPaaS Services: 5G connectivity of 5 Gbps Mobile Throughput and sub 5ms latency, cloud services of IoT Accelerator, emulation cloud, edge computing - iPaaS Infrastructure: 5G mobile user equipment (android-based Qualcom terminals operating at 3.5 GHz), 5G radio access technology (NR/LTE/CAT-M1/NB-IoT), 5G transport/backhaul, distributed core network and programmable data plane - Future features expected by 2021 include: 5 Gbps 5G NR, sub 5ms latency, predictive analytics, federated network slicing, real time machine learning / AI - Technology partners: Ericsson, Thales, CGI, IBM, Ciena

in the decision figure is whether the data required is for completely new or unseen scenarios (e.g., 6G drones to terrestrial networks that are not yet deployed) or whether the data required is for scenarios already present in today's networks. In the former case, the only options are utilizing testbeds and simulators to depict new use cases. In the latter case, if the data is non-representative (i.e., very few data points are available that might not represent the scenario very

TABLE IV
REVIEW OF MODELING TECHNIQUES FOR HANDLING SCARCE DATASETS IN RADIO ACCESS NETWORKS (RAN)

Reference	Year	Modeling technique	Use case and data	Data type	Use-case type w.r.t. OSI layer	Use-case type w.r.t. level of analysis
[78]	2020	Few-shot learning	eNodeB performance metric analysis using cell trace data	Tabular data	Network	System
[79]	2021	Few-shot learning	Modeling indoor pathloss model at 28 GHz using RSS data	Tabular data	Physical	Link
[139]	2020	Few-shot learning + Transfer learning	Network power minimization in C-RAN for resource management using UE SINR data	Tabular data	Physical	System
[84]	2019	Transfer learning	Identifying optimal deployment density of the BSs given a BS transmit power w.r.t. spectral and energy efficiency of the network using UE SINR data	Tabular data	Phyical	System
[85]	2019	Transfer learning	Network parameter optimization for uplink power control and user scheduling using Cell KPI/counter data	Tabular data	Application	System
[86]	2014	Transfer learning	BS ON/OFF switching for energy saving using traffic load data	Tabular data	Data Link	System
[87]	2020	Transfer learning	Radio map prediction under different antenna tilt using UE RSS data	Tabular data	Physical	Link
[88]	2021	Transfer learning	Cell performance prediction (CQI and Active UE count) using cell KPI/Counter data	Tabular data	Application	System
[89], [90]	2019-2021	Transfer learning	Network service performance prediction using testbed traces	Tabular data	Network	System
[82]	2020	Transfer learning + GAN	REM generation	Spatial data	Physical	System
[19]	2019	GAN	Synthetic CDR generation using CDR data (call start hour and call duration)	Tabular data	Network	System
[76]	2020	ZipNet-GAN	Infer fine-grained traffic patterns from course aggregates using CDR data	Spatio-temporal data	Network	System
[77]	2020	GAN	Cell outage detection using MDT data	Tabular data	Application	System
[81]	2020	GAN	REM generation	Spatial data	Physical	System

(continued)

well), the options are again to generate more synthetic data through simulators or real data through testbeds and mobile applications.

However, if the data is representative, low dimensional in nature (e.g., spatial only), and exhibits some correlation (e.g., RSRP values that are correlated with distance), the choice

TABLE IV
(Continued) REVIEW OF MODELING TECHNIQUES FOR HANDLING SCARCE DATASETS IN RADIO ACCESS NETWORKS (RAN)

[109]	2020	Variational autoencoder	Anomaly detection and root cause analysis (RCA) in RAN using KPI/KQI data	Tabular data	Application	System
[110]	2018	Adversarial autoencoder	Detecting anomalous behavior in wireless spectrum using power spectral density data	Tabular data	Network	System
[13], [71], [72], [65], [69]	2015-2020	Context-aware interpolation	REM construction using BS location estimated through reverse triangulation	Spatial data	Physical	System
[55], [62], [58], [60], [67]	2018-2020	Kriging interpolation + variants	REM generation	Spatial data	Physical	System
[32]	2019	Correlation-based interpolation	Crowdsourced spatio-temporal REM generation	Spatio-temporal data	Application	System
[211]	2019	Adaptive spatial interpolation	Uplink channel estimation in 3-D massive MIMO systems	Spatial data	Physical	Link
[52]	2019	Adaptive triangulation - induced interpolation	Multiple REM generation	Spatial data	Physical	System
[63]	2019	NN-enhanced, Kriging interpolation	REM generation	Spatial data	Physical	System
[33]	2018	Congregate group pattern	Signaling data (User trajectory data) for discovering congregate group patterns	Spatio-temporal data	Network	System
[34]	2020	Kriging, moving average, matrix completion, IDW, nearest neighbors, natural neighbors, spline interpolation	MDT coverage map (RSRP) construction	Spatial data	Physical	System
[56]	2019	Kriging interpolation	REM generation from crowdsourced data	Spatial data	Application	System
[48]	2011	Kriging, MSM and GIDS interpolation	REM construction from total received signal power	Spatial data	Physical	System
[42]	2012	IDW, adaptive IDW, MSM interpolation	REM construction	Spatial data	Physical	System
[49]	2014	Nearest neighbor, IDW, Kriging interpolation	Interference map estimation of MDT reports in cognitive radio networks	Spatial data	Physical	System

(continued)

of methods depends on whether the propagation environment parameters (e.g., frequency, path loss exponent) are known or not. If these parameters are known, along with knowledge of receivers' SNR and transmit power (through, e.g., operator), then SNR based method in Section III-B4 can be used. If transmit power is known, but receivers' SNR is not known, but

TABLE IV
(Continued) REVIEW OF MODELING TECHNIQUES FOR HANDLING SCARCE DATASETS IN RADIO ACCESS NETWORKS (RAN)

[50]	2012	Nearest neighbor, natural neighbor, triangulation-based interpolation	Interference map generation in cognitive radio networks	Spatial data	Physical	System
[51]	2013	Nearest neighbor, IDW, Kriging	Interference maps for licensed shared access	Spatial data	Physical	System
[14]	2012	Natural neighbor, kriging and spline	Interference cartography generation in cognitive radio networks	Spatial data	Physical	System
[53]	2010	Kriging	Predict network coverage in wireless networks	Spatial data	Physical	Link
[55]	2018	Kriging	REM construction	Spatial data	Physical	System
[57]	2018	Kriging	REM construction in cognitive radio networks	Spatial data	Physical	System
[58]	2019	Kriging, nearest neighbor, IDW	REM construction based on RSSI mobile crowdsensing data	Spatial data	Application	System
[59]	2019	Nearest neighbor, IDW, Kriging	REM construction for spectrum sharing	Spatial data	Physical	System
[60]	2020	Nearest neighbor, IDW, Kriging	REM construction	Spatial data	Physical	System
[61]	2014	Kriging	REM generation for coverage mapping	Spatial data	Physical	System
[62]	2028	Kriging	REM generation for coverage mapping	Spatial data	Link	System
[63], [65]	2019-2020	Hybrid neural networks and Kriging interpolation	REM generation	Spatial data	Physical	System
[66]	2015	Kriging, splines, moving average, triangulation-based interpolation	Coverage extension and prediction with signal strength crowdsourced measurements	Spatial data	Application	System
[67]	2019	Nearest neighbor, IDW, Kriging	REM construction for military cognitive networks	Spatial data	Physical	System
[68]	2018	RSS and RSSD based methods	REM enrichment using RSS measurements from sensors	Spatial data	Physical	System
[69]	2015	STM method, location estimation-based method, IDW, Kriging	REM construction using omnidirectional and directional transmitter antenna	Spatial data	Physical	System
[13]	2015	RSS-based methods	REM construction in fading channels	Spatial data	Physical	System
[71]	2018	RSS-based method, kriging	REM construction	Spatial data	Physical	System

(continued)

antenna characteristics (e.g., antenna tilt, patterns) are known, then the STM method in Section III-B5 can be used. If SNR is not known, and antenna information is also not available, then based on the propagation environment and transmit power information only, three methods described in Section III-B, AOA, RSSD and RSS can be used.

TABLE IV
(Continued) REVIEW OF MODELING TECHNIQUES FOR HANDLING SCARCE DATASETS IN RADIO ACCESS NETWORKS (RAN)

[72]	2010	AOA based and SNR based methods	REM construction	Spatial data	Physical	System
[140]	2022	Synthetic data generation through Atoll simulator	Cell outage detection and diagnosis using SINR-based REM maps	Tabular data	Physical	System
[80]	2022	Synthetic data generation through Atoll simulator	Modeling outdoor propagation model using RSS data	Tabular data	Physical	System
[141]	2022	Synthetic data generation through SyntheticNet simulator	Optimization of A5 mobility parameters using RSRP, SINR, and handover success rate data (HOSR)	Tabular data	RSRP/SINR: Physical, HOSR: Network	System
[108]	2016	Real data generation through smartphone application	Building semi-markov model based mobility prediction schemes using handover data	Tabular data	Network	System
[142]	2020	Real data generation using mmWave testbed	Building beam tracking algorithms using mmWave channel measurement data	Tabular data	Physical	Link
[143]	2014	Real data generation using CORNET testbed	Evaluating real-time radio spectrum access using RSS, packet and bit error rate data (PER/BER)	Tabular data	Physical	Link
[144]	2017	Real data generation using LuMaMi testbed	Design and validation of massive MIMO research using SNR and BER data for different antenna configurations and modulation schemes	Tabular data	Physical	Link

If the low dimensional data is correlated, but we do not have information about propagation environment or transmit power, choice of interpolation method can be done on the based on other contextual information, such as network geometry, which if known, leads to cluster-based interpolation in Section III-A2. If, along with network geometry, transmitter locations are also known, then the triangle method in Section III-A1 can be a possible choice. If, however, the network geometry is also not known, but the data forms a low-rank matrix (e.g., ultra-dense high frequency scenario), then matrix completion in Section II-A can be a choice. Otherwise, decision is made by assessing whether the underlying data surface is mathematically smooth or not. By smooth, we mean differentiable and continuous surface. In case of smooth surface that requires extrapolation of data, kriging, GIDS, MSM, and Splines can be used and where extrapolation is not required, all interpolation methods in Section II can be used with the exception of natural neighbors, which can be used only if all data points are inside the convex hull of location measurements. In the case of non-smooth surface that requires extrapolation, kriging, GIDS, MSM can be used, and if the non-smooth data surface requires interpolation only, then kriging, GIDS, MSM, Nearest neighbors, natural neighbors are the choices, since splines and IDW can be used on smooth data surfaces only. The exception here is again natural neighbors, which can be used only if all data points are inside the convex hull of location measurements.

If the low dimensional data does not exhibit any correlation, we arrive at the decision block that coincides with the case of high dimensional data (e.g., spatio-temporal tabular data with multiple features) nature of data. In these cases, if the data has many latent features, then VAEs in Section IV-B can be used given the prior distribution of latent features is known or can be approximated, otherwise GANs discussed in Section IV-A can be the choice since they do not require the knowledge of prior distribution of latent features. On the contrary, if the low dimensional data does not exhibit any correlation and also does not have enough latent features, then the decision is made based on the availability of

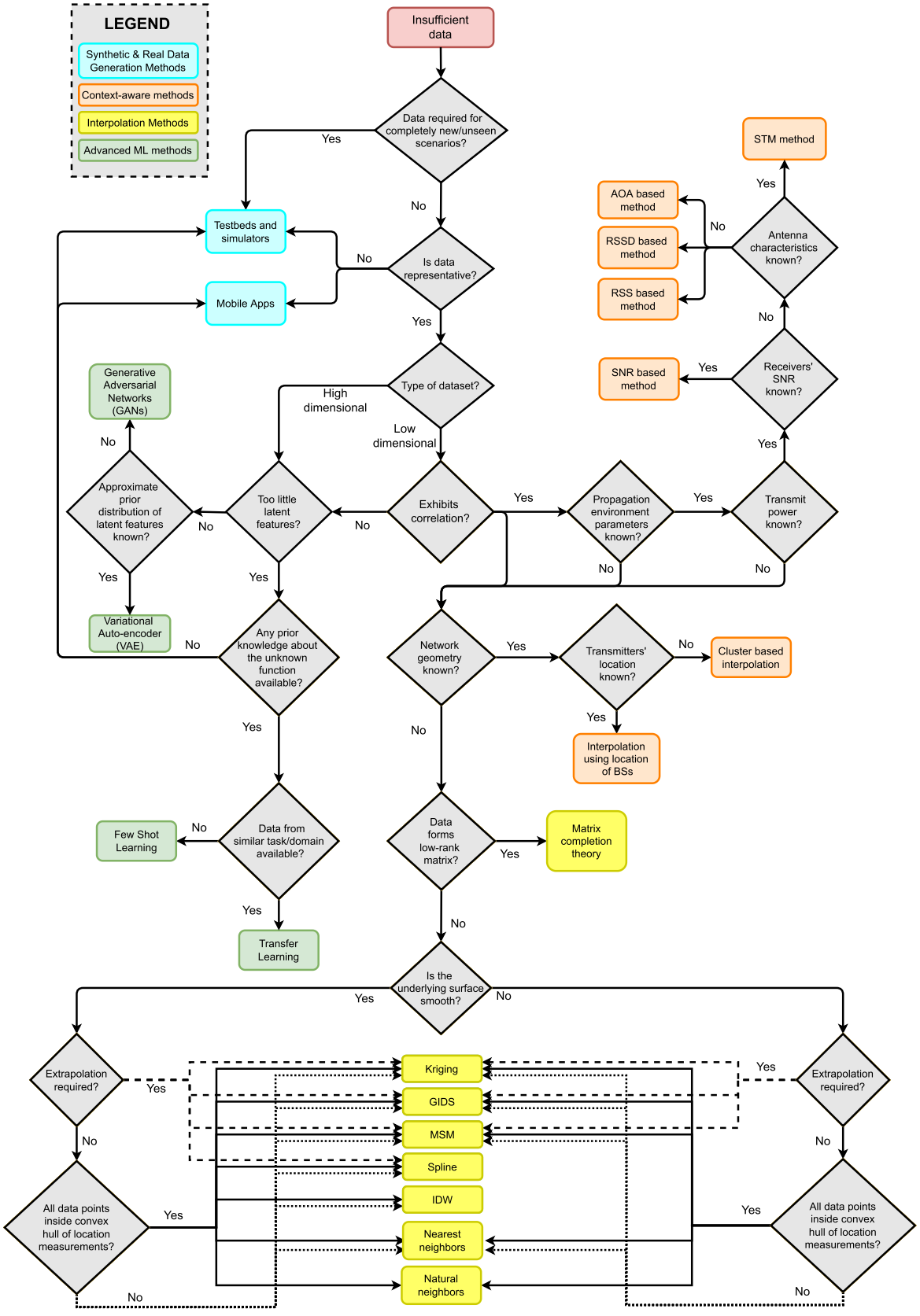


Fig. 17. Decision flowchart for the selection of data augmentation technique for handling scarce datasets in mobile networks.

any prior knowledge about the distribution of data, which if unknown, leads to the augmentation of data through testbeds, simulators and mobile application, and if known, leads to the possible solution of transfer learning (if data from a similar domain is available), otherwise, few-shot learning can be the choice.

VIII. Future Directions

Since the advanced machine learning methods, such as GANs, transfer learning and few short learning are much less explored for different telco use-cases, as compared to techniques such as interpolation methods, more investigation of these techniques in telco domain in needed. Particularly the potential of transfer learning remains unexploited. Future work focused on questions on what to transfer, where to transfer and how transfer while taking into account domain knowledge of RAN may help avail the full potential of transfer learning for wireless networks.

Similarly, in GANs, research questions such as how much minimum data is needed to train a generator for given type of RAN data and problem is an important direction to exploit the full potential of GANs and their limits on synthesizing RAN data. A recent work explores this question [145] indicating significance of this research direction.

Moreover, solutions that have the scalability to generate high dimensional data, robustness to highly dynamic real environments and the capability to take conditional context of the required network conditions into account can also be another future direction.

Another research direction worth exploring to address the data sparsity challenge in wireless communication domain is by leveraging active learning [146], which harnesses the power of machine learning together with the experience from domain expert.

Most current machine learning based approaches to enrich training data are predominately used as black-box models, allowing little interpretability. Therefore, another future direction can be to design gray-box (or hybrid) machine learning models (e.g., GANs) by combining domain knowledge and analytical modeling with machine learning. This can bring model interpretability and therefore improved ability to extrapolate beyond the exposed training data distributions.

Validating the recent and new developed methods and solutions on real data from operators and testbeds can also be a focus of future work.

There is also a need for datasets in this domain to be publicly accessible to enable the research community to devise practical solutions that can be benchmarked. One such initiative in this direction was taken in the form of CRAWDAD repository [147].

Recent advancements in Open RAN might also help the data scarcity challenge as Open RAN introduces a set of open standardized interfaces to interact, control and collect data from every node of the network [148]. However, the issue stemming from sparsity of data (resulting from operators trying a limited range of COPs that leads to a sparse data distribution) will still remain as Open RAN will not allow experimentation on a live network. Consequently, the exploration and advancements of the techniques discussed in this survey will be required.

References

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

[2] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 336–361, 1st Quart., 2012.

[3] A. Asghar, H. Farooq, and A. Imran, "Self-healing in emerging cellular networks: Review, challenges, and research directions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1682–1709, 3rd Quart., 2018.

[4] U. S. Hashmi, S. A. R. Zaidi, and A. Imran, "User-centric cloud RAN: An analytical framework for optimizing area spectral and energy efficiency," *IEEE Access*, vol. 6, pp. 19859–19875, 2018.

[5] H. N. Qureshi and A. Imran, "Towards designing systems with large number of antennas for range extension in ground-to-air communications," in *Proc. IEEE 29th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2018, pp. 1–5.

[6] A. AlAmmouri, J. G. Andrews, and F. Baccelli, "Asymptotic analysis of area spectral efficiency in dense cellular networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2018, pp. 56–60.

[7] H. N. Qureshi, I. H. Naqvi, and M. Uppal, "Massive MIMO with quasi orthogonal pilots: A flexible solution for TDD systems," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, 2017, pp. 1–6.

[8] H. N. Qureshi and A. Imran, "On the tradeoffs between coverage radius, altitude and beamwidth for practical UAV deployments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 6, pp. 2805–2821, Dec. 2019.

[9] O. Onireti, A. Imran, and M. A. Imran, "Coverage, capacity, and energy efficiency analysis in the uplink of mmWave cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 3982–3997, May 2018.

[10] P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza, "A survey of machine learning techniques applied to self-organizing cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2392–2431, 4th Quart., 2017.

[11] E. Balevi and J. G. Andrews, "Online antenna tuning in heterogeneous cellular networks with deep reinforcement learning," 2019, *arXiv:1903.06787*.

[12] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "Data-driven analytics for automated cell outage detection in self-organizing networks," in *Proc. 11th Int. Conf. Design Rel. Commun. Netw. (DRCN)*, 2015, pp. 203–210.

[13] H. B. Yilmaz and T. Tugcu, "Location estimation-based radio environment map construction in fading channels," *Wireless Commun. Mobile Comput.*, vol. 15, no. 3, pp. 561–570, 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcm.2367

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

[15] C. Phillips, M. Ton, D. Sicker, and D. Grunwald, "Practical radio environment mapping with geostatistics," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw.*, 2012, pp. 422–433.

[16] I. Akbari, O. Onireti, A. Imran, M. A. Imran, and R. Tafazolli, "How reliable is MDT-based autonomous coverage estimation in the presence of user and BS positioning error?" *IEEE Wireless Commun. Lett.*, vol. 5, no. 2, pp. 196–199, Apr. 2016.

[17] P.-C. Lin, "Minimization of drive tests using measurement reports from user equipment," in *Proc. IEEE Global Conf. Consum. Electron. (GCCE)*, Oct. 2014, pp. 84–85.

[18] H. N. Qureshi and A. Imran, "Optimal bin width for autonomous coverage estimation using MDT reports in the presence of user positioning error," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 716–719, Apr. 2019.

[19] B. Hughes, S. Bothe, H. Farooq, and A. Imran, "Generative adversarial learning for machine learning empowered self Organizing 5G networks," in *Proc. Int. Conf. Comput. Netw. Commun. (ICNC)*, Feb. 2019, pp. 282–286.

[20] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[21] J. Li and A. D. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecol. Inf.*, vol. 6, nos. 3–4, pp. 228–241, 2011.

[22] L. Mitas and H. Mitasova, "Spatial interpolation," in *Geographical Information Systems: Principles, Techniques, Management and Applications*, vol. 1. Hoboken, NJ, USA: Wiley, 1999.

[23] F. Susanto, P. de Souza, and J. He, "Spatiotemporal interpolation for environmental modelling," *Sensors*, vol. 16, no. 8, p. 1245, 2016.

[24] J. Li and A. D. Heap, *A Review of Spatial Interpolation Methods for Environmental Scientists*. Canberra, ACT, Australia: Geosci. Australia, 2008. [Online]. Available: https://www.researchgate.net/publication/246546630_A_Review_of_Spatial_Interpolation_Methods_for_Environmental_Scientists

[25] M. Pesko, T. Javornik, A. Košir, M. Štular, and M. Mohorčič, "Radio environment maps: The survey of construction methods," *KSII Trans. Internet Inf. Syst.*, vol. 8, no. 11, p. 269, 2014.

[26] M. Höyhtyä et al., "Spectrum occupancy measurements: A survey and use of interference maps," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2386–2414, 4th Quart., 2016.

[27] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, 3rd Quart., 2019.

[28] C. Bouras, A. Gkamas, G. Diles, and Z. Andreas, "A comparative study of 4G and 5G network simulators," *Int. J. Adv. Netw. Services*, vol. 13, no. 1, pp. 1–10, 2020.

[29] P. K. Gkonis, P. T. Trakadas, and D. I. Kaklamani, "A comprehensive study on simulation techniques for 5G networks: State of the art results, analysis, and future challenges," *Electronics*, vol. 9, no. 3, p. 468, 2020.

[30] M. Manalastas et al., "Design considerations and deployment challenges for TurboRAN 5G and beyond testbed," *IEEE Access*, vol. 10, pp. 39810–39824, 2022.

[31] M. A. Azpurua and K. D. Ramos, "A comparison of spatial interpolation methods for estimation of average electromagnetic field magnitude," *Progr. Electromagn. Res.*, vol. 14, pp. 135–145, Sep. 2010.

[32] M. S. Rahman, H. Gupta, A. Chakraborty, and S. R. Das, "Creating spatio-temporal spectrum maps from sparse Crowdsensed data," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7.

[33] T. Chen, Y. Zhang, Y. Tuo, and W. Wang, "Online discovery of congregate groups on sparse spatio-temporal data," in *Proc. IEEE 29th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.

[34] H. N. Qureshi, A. Imran, and A. A Abu-Dayya, "Enhanced MDT-based performance estimation for AI driven optimization in future cellular networks," *IEEE Access*, vol. 8, pp. 161406–161426, 2020.

[35] *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) version 10.2.0*, 3GPP Standard TS 37.320, Jun. 2011.

[36] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, 2009.

[37] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[38] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," *Math. Program.*, vol. 128, nos. 1–2, pp. 321–353, 2011.

[39] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for L1-minimization: Methodology and convergence," *SIAM J. Optim.*, vol. 19, no. 3, pp. 1107–1130, 2008.

[40] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proc. 23rd ACM Nat. Conf.*, 1968, pp. 517–524.

[41] A. S. Sengar, R. Gangopadhyay, and S. Debnath, "On the construction of radio environment map for underlay device-to-device networks," in *Proc. 24th Asia–Pac. Conf. Commun. (APCC)*, Nov. 2018, pp. 413–417.

[42] D. Denkovski, V. Atanasovski, L. Gavrilovska, J. Riihijärvi, and P. Mähönen, "Reliability of a radio environment map: Case of spatial interpolation techniques," in *Proc. IEEE 7th Int. ICST Conf. Cogn. Radio Orient. Wireless Netw. Commun. (CROWNCOM)*, 2012, pp. 248–253.

[43] R. Franke and G. M. Nielson, "Scattered data interpolation and applications: A tutorial and survey," in *Geometric Modeling*. Heidelberg, Germany: Springer, 1991, pp. 131–160.

[44] G. Y. Lu and D. W. Wong, "An adaptive inverse-distance weighting spatial interpolation technique," *Comput. Geosci.*, vol. 34, no. 9, pp. 1044–1055, 2008.

[45] S. Henley, *Nonparamteric Geostatistics*. Dordrecht, The Netherlands: Springer, 1981, p. 145.

[46] D. Weber and E. Englund, "Evaluation and comparison of spatial interpolators," *Math. Geol.*, vol. 24, no. 4, pp. 381–391, 1992.

[47] M. Deng, Z. Fan, Q. Liu, and J. Gong, "A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets," *ISPRS Int. J. Geo Inf.*, vol. 5, no. 2, p. 13, 2016.

[48] M. Angjelicinoski, V. Atanasovski, and L. Gavrilovska, "Comparative analysis of spatial interpolation methods for creating radio environment maps," in *Proc. IEEE 19th Telecommun. Forum (TELFOR)*, 2011, pp. 334–337.

[49] J. D. Naranjo, A. Ravanshid, I. Viering, R. Halfmann, and G. Bauch, "Interference map estimation using spatial interpolation of MDT reports in cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2014, pp. 1496–1501.

[50] S. Üreten, A. Yongaçoğlu, and E. Petriu, "Interference map generation based on delaunay triangulation in cognitive radio networks," in *Proc. IEEE 13th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2012, pp. 134–138.

[51] R. C. Dwarakanath, J. D. Naranjo, and A. Ravanshid, "Modeling of interference maps for licensed shared access in LTE-advanced networks supporting carrier aggregation," in *Proc. IEEE IFIP Wireless Days (WD)*, 2013, pp. 1–6.

[52] Y. Liu, W. Huangfu, H. Zhang, and K. Long, "Multi-criteria coverage map construction based on adaptive triangulation-induced interpolation for cellular networks," *IEEE Access*, vol. 7, pp. 80767–80777, 2019.

[53] A. Konak, "A kriging approach to predicting coverage in wireless networks," *Int. J. Mobile Netw. Design Innov.*, vol. 3, no. 2, pp. 65–71, 2009.

[54] F. Yaseen, U. Masood, A. N. Hassan, and I. H. Naqvi, "Graph signal processing-based network health estimation for next generation wireless systems," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 104–107, 2018.

[55] A. M. Alam, S. Benjemaa, and T. Romary, "Clustering for high accuracy coverage mapping," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[56] X. Wang, M. Umehira, B. Han, P. Li, Y. Gu, and C. Wu, "Online incentive mechanism for crowdsourced radio environment map construction," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[57] D. Mao, W. Shao, Z. Qian, H. Xue, X. Lu, and H. Wu, "Constructing accurate radio environment maps with Kriging interpolation in cognitive radio networks," in *Proc. Cross Strait Quad-Reg. Radio Sci. Wireless Technol. Conf. (CSQRWC)*, Jul. 2018, pp. 1–3.

[58] Z. Han, J. Liao, Q. Qi, H. Sun, and J. Wang, "Radio environment map construction by Kriging algorithm based on mobile crowd sensing," *Wireless Commun. Mobile Comput.*, vol. 2019, Feb. 2019, Art. no. e4064201. [Online]. Available: https://www.hindawi.com/journals/wcmc/2019/4064201/

[59] R. Hosseini Tehrani, "Radio environment map-enabled spectrum sharing in mobile cellular networks," Ph.D. dissertation, Dept. Comput. Sci., Univ. Surrey, Surrey, U.K., Jun. 2019. [Online]. Available: http://epubs.surrey.ac.uk/851962/

[60] H. Xia, S. Zha, J. Huang, and J. Liu, "Radio environment map construction by adaptive ordinary Kriging algorithm based on affinity propagation clustering," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 5, p. 19, May 2020. [Online]. Available: https://doi.org/10.1177/1550147720922484

[61] H. Braham, S. B. Jemaa, B. Sayrac, G. Fort, and E. Moulines, "Coverage mapping using spatial interpolation with field measurements," in *Proc. IEEE 25th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2014, pp. 1743–1747.

[62] A. M. Alam, S. Benjemaa, and T. Romary, "Performance evaluation of covariance tapering for coverage mapping," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.

[63] K. Sato, K. Inage, and T. Fujii, "On the performance of neural network residual Kriging in radio environment mapping," *IEEE Access*, vol. 7, pp. 94557–94568, 2019.

[64] G. Appleby, L. Liu, and L.-P. Liu, "Kriging convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, Apr. 2020, pp. 3187–3194. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5716

[65] N. Mezhoud, M. Oussalah, A. Zaatri, and Z. Hammoudi, "Hybrid Kriging and multilayer perceptron neural network technique for coverage prediction in cellular networks," *Int. J. Parallel Emergent Distrib. Syst.*, vol. 35, no. 6, pp. 682–706, 2020. [Online]. Available: https://doi.org/10.1080/17445760.2020.1805609

[66] M. Molinari, M.-R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," in *Proc. ACM SIGCOMM Workshop Crowdsourcing Crowdsharing Big (Internet) Data*, 2015, pp. 33–38.

[67] M. Suchanski, P. Kaniewski, J. Romanik, E. Golan, and K. Zubel, "Radio environment maps for military cognitive networks: Deployment of sensors vs. map quality," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2019, pp. 1–6.

[68] S. Alfattani and A. Yonzacoglu, "Indirect methods for constructing radio environment map," in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (CCECE)*, 2018, pp. 1–5.

[69] M. Pesko, T. Javornik, L. Vidmar, A. Košir, M. Štular, and M. Mohorčič, "The indirect self-tuning method for constructing radio environment map using omnidirectional or directional transmitter antenna," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 50, Mar. 2015. [Online]. Available: https://doi.org/10.1186/s13638-015-0297-2

[70] A. Mohamed, O. Onireti, M. A. Imran, A. Imran, and R. Tafazolli, "Control-data separation architecture for cellular radio access networks: A survey and outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 446–465, 1st Quart., 2015.

[71] K. Tsukamoto, M. Kitsunezuka, and K. Kunihiro, "Highly accurate radio environment mapping method based on transmitter localization and spatial interpolation in urban LoS/NLoS scenario," in *Proc. IEEE Topical Conf. Wireless Sensors Sensor Netw. (WiSNet)*, Jan. 2018, pp. 5–7.

[72] G. Sun and J. van de Beek, "Simple distributed interference source localization for radio environment mapping," in *Proc. IFIP Wireless Days*, Oct. 2010, pp. 1–5.

[73] A. Pagès-Zamora, J. Vidal, and D. H. Brooks, "Closed-form solution for positioning based on angle of arrival measurements," in *Proc. 13th IEEE Int. Symp. Pers. Indoor Mobile Radio Commun.*, vol. 4, 2002, pp. 1522–1526.

[74] I. Kakalou, K. E. Psannis, S. K. Goudos, T. V. Yioultsis, N. V. Kantartzis, and Y. Ishibashi, "Radio environment maps for 5G cognitive radio network," in *Proc. 8th Int. Conf. Modern Circuits Syst. Technol. (MOCAST)*, May 2019, pp. 1–4.

[75] I. Chahrour and J. Wells, "Comparing machine learning and interpolation methods for loop-level calculations," *SciPost Phys.*, vol. 12, no. 6, p. 187, 2022.

[76] C. Zhang, X. Ouyang, and P. Patras, "ZipNet-GAN: Inferring fine-grained mobile traffic patterns via a generative adversarial neural network," in *Proc. 13th Int. Conf. Emerg. Netw. Exp. Technol. (CoNEXT)*, 2017, pp. 363–375. [Online]. Available: https://doi.org/10.1145/3143361.3143393

[77] T. Zhang, K. Zhu, and D. Niyato, "A generative adversarial learning-based approach for cell outage detection in self-organizing cellular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 171–174, Feb. 2020.

[78] S. Aoki, K. Shiomoto, C. L. Eng, and S. Backstad, "Few-shot learning for eNodeB performance metric analysis for service level assurance in LTE networks," in *Proc. NOMS IEEE/IFIP Netw. Oper. Manag. Symp.*, Apr. 2020, pp. 1–4.

[79] P. Wang and H. Lee, "Indoor path loss modeling for 5G communications in smart factory scenarios based on meta-learning," in *Proc. 12th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, 2021, pp. 438–443.

[80] U. Masood, H. Farooq, A. Imran, and A. Abu-Dayya, "Interpretable AI-based large-scale 3D pathloss prediction model for enabling emerging self-driving networks," *IEEE Trans. Mobile Comput.*, early access, Jan. 31, 2022, doi: 10.1109/TMC.2022.3147191.

[81] X. Han, L. Xue, Y. Xu, and Z. Liu, "A radio environment maps estimation algorithm based on the pixel regression framework for underlay cognitive radio networks using incomplete training data," *Sensors*, vol. 20, no. 8, p. 2245, Jan. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/8/2245

[82] H. Xu, L. Xue, Y. Xu, and Z. Liu, "A two-phase transfer learning-based power spectrum maps reconstruction algorithm for underlay cognitive radio networks," *IEEE Access*, vol. 8, pp. 81232–81245, 2020.

[83] C. Parera, A. E. Redondi, M. Cesana, Q. Liao, and I. Malanchini, "Transfer learning for channel quality prediction," in *Proc. IEEE Int. Symp. Meas. Netw. (M N)*, Jul. 2019, pp. 1–6.

[84] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?" *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.

[85] J. Chuai et al., "A collaborative learning based approach for parameter configuration of cellular networks," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, Apr. 2019, pp. 1396–1404.

[86] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor–critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

[87] C. Parera, Q. Liao, I. Malanchini, C. Tatino, A. E. C. Redondi, and M. Cesana, "Transfer learning for tilt-dependent radio map prediction," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 829–843, Jun. 2020.

[88] C. Parera, A. E. C. Redondi, M. Cesana, Q. Liao, and I. Malanchini, "Anticipating mobile radio networks key performance indicators with transfer learning," in *Proc. IEEE 16th Annu. Conf. Wireless On-Demand Netw. Syst. Services Conf. (WONS)*, 2021, pp. 1–8.

[89] F. Moradi, R. Stadler, and A. Johnsson, "Performance prediction in dynamic clouds using transfer learning," in *Proc. IFIP/IEEE Symp. Integr. Netw. Service Manag. (IM)*, 2019, pp. 242–250.

[90] H. Larsson, J. Taghia, F. Moradi, and A. Johnsson, "Source selection in transfer learning for improved service performance predictions," in *Proc. IFIP Netw. Conf. (IFIP Netw.)*, 2021, pp. 1–9.

[91] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," 2017, *arXiv:1711.04340*.

[92] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[93] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[94] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," 2018, *arXiv:1803.04469*.

[95] C. Bowles et al., "GAN augmentation: Augmenting training data using generative adversarial networks," 2018, *arXiv:1810.10863*.

[96] T. Erpek, Y. E. Sagduyu, and Y. Shi, "Deep learning for launching and mitigating wireless jamming attacks," Dec. 2018, *arXiv:1807.02567*.

[97] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network for wireless signal spoofing," May 2019, *arXiv:1905.01008*.

[98] M. Nabati, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using synthetic data to enhance the accuracy of fingerprint-based Localization: A deep learning approach," *IEEE Sensors Lett.*, vol. 4, no. 4, pp. 1–4, Apr. 2020.

[99] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," Aug. 2020, *arXiv:2008.09202*.

[100] F. Tanaka and C. Aranha, "Data augmentation using GANs," Apr. 2019, *arXiv:1904.09135*.

[101] N. Gao et al. "Generative adversarial networks for spatio-temporal data: A survey." Aug. 2020. [Online]. Available: http://arxiv.org/abs/2008.08903

[102] R. D. Camino, C. A. Hammerschmidt, and R. State. "Improving missing data imputation with deep generative models." Feb. 2019. [Online]. Available: http://arxiv.org/abs/1902.10666

[103] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. "Modeling tabular data using conditional GAN." Oct. 2019. [Online]. Available: http://arxiv.org/abs/1907.00503

[104] L. Xu and K. Veeramachaneni. "Synthesizing tabular data using generative adversarial networks." Nov. 2018. [Online]. Available: http://arxiv.org/abs/1811.11264

[105] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? A large-scale study," 2017, *arXiv:1711.10337*.

[106] G. Barlacchi et al., "A multi-source dataset of urban life in the city of Milan and the Province of Trentino," *Sci. Data*, vol. 2, no. 1, pp. 1–15, 2015.

[107] A. Zoha, A. Saeed, H. Farooq, A. Rizwan, A. Imran, and M. A. Imran, "Leveraging intelligence from network CDR data for interference aware energy consumption minimization," *IEEE Trans. Mobile Comput.*, vol. 17, no. 7, pp. 1569–1582, Jul. 2018.

[108] H. Farooq and A. Imran, "Spatiotemporal mobility prediction in proactive self-organizing cellular networks," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 370–373, Feb. 2017.

[109] Y. Yuan, J. Yang, R. Duan, I. Chih-Lin, and J. Huang, "Anomaly detection and root cause analysis enabled by artificial intelligence," in *Proc. IEEE Globecom Workshops (GC Wkshps*, 2020, pp. 1–6.

[110] S. Rajendran, W. Meert, V. Lenders, and S. Pollin, "SAIFE: Unsupervised wireless spectrum anomaly detection with interpretable features," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Oct. 2018, pp. 1–9.

[111] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[112] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3265–3275, May 2018.

[113] T. Domínguez-Bolaño, J. Rodríguez-Piñeiro, J. A. García-Naya, and L. Castedo, "The GTEC 5G link-level simulator," in *Proc. IEEE 1st Int. Workshop Link Syst. Level Simulat. (IWSLS)*, 2016, pp. 1–6.

[114] OpenAirInterface. "OpenAirInterface: 5G software alliance for democratising wireless innovation." [Online]. Available: http://opnetprojects.com/opnet-simulator/

[115] J. Baek et al., "5G K-simulator of flexible, open, modular (FOM) structure and Web-based 5G K-SimPlatform," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2019, pp. 1–4.

[116] X. Wang, Y. Chen, and Z. Mai, "A novel design of system level simulator for heterogeneous networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–6.

[117] V. V. Díaz and D. M. Aviles, "A path loss simulator for the 3GPP 5G channel models," in *Proc. IEEE XXV Int. Conf. Electron. Elect. Eng. Comput. (INTERCON)*, 2018, pp. 1–4.

[118] NS 3. "mmWave cellular network simulator." [Online]. Available: https://omnetpp.org/

[119] OMNeT++. "OMNeT++: Discrete event simulator." [Online]. Available: https://apps.nsnam.org/app/mmwave/

[120] S. Sun, G. R. MacCartney, and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5G wireless communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–7.

[121] MATLAB. "Why use MATLAB and simulink for 5G?" [Online]. Available: https://www.mathworks.com/solutions/wireless-communications/5g.html

[122] N. Mohsen and K. S. Hassan, "C-RAN simulator: A tool for evaluating 5G cloud-based networks system-level performance," in *Proc. IEEE 11th Int. Conf. Wireless Mobile Comput. Netw. Commun. (WiMob)*, 2015, pp. 302–309.

[123] OPNET. "OPNET: Optimum network performance." [Online]. Available: https://www.openairinterface.org/

[124] M. K. Muller et al., "Flexible multi-node simulation of cellular mobile communications: The vienna 5G system level simulator," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, p. 227, Sep. 2018.

[125] "Atoll." Accessed: Feb. 1, 2023. [Online]. Available: https://www.forsk.com/

[126] S. M. A. Zaidi, M. Manalastas, H. Farooq, and A. Imran, "AI4Networks simulator—A true 3GPP compliant 5G network simulator with support of AI," *IEEE Access*, vol. 8, pp. 82938–82950, 2020.

[127] Y. Wang, Q. Yao, J. Kwok, and L. M. Ni. "Generalizing from a few examples: A survey on few-shot learning." Mar. 2020. [Online]. Available: http://arxiv.org/abs/1904.05046

[128] J. Shtok. "Few-shot learning—State of the art." 2019. [Online]. Available: https://2019.imvc.co.il/Portals/117/Joseph_Shtok.pdf

[129] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[130] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," 2017, *arXiv:1703.03400*.

[131] M. A. Jamal and G.-J. Qi, "Task agnostic meta-learning for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11719–11727.

[132] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 403–412.

[133] C. Finn and S. Levine, "Meta-learning: From few-shot learning to rapid reinforcement learning," in *Proc. ICML*, 2019, pp. 1–6.

[134] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, pp. 1–8.

[135] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3630–3638.

[136] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[137] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[138] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," 2017, *arXiv:1711.04043*.

[139] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "LORM: Learning to optimize for resource management in wireless networks with few training samples," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 665–679, Jan. 2020.

[140] M. S. Riaz, H. N. Qureshi, U. Masood, A. Rizwan, A. Abu-Dayya, and A. Imran, "Deep learning-based framework for multi-fault diagnosis in self-healing cellular networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 746–751.

[141] M. U. B. Farooq et al., "A data-driven self-optimization solution for inter-frequency mobility parameters in emerging networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 2, pp. 570–583, Jun. 2022.

[142] I. K. Jain, R. Subbaraman, T. H. Sadarahalli, X. Shao, H.-W. Lin, and D. Bharadia, "mMobile: Building a mmWave testbed to evaluate and address mobility effects," in *Proc. 4th ACM Workshop Millimeter Wave Netw. Sens. Syst.*, 2020, pp. 1–6.

[143] N. Sharakhov, V. Marojevic, F. Romano, N. Polys, and C. Dietrich, "Visualizing real-time radio spectrum access with CORNET3D," in *Proc. 19th Int. ACM Conf. 3D Web Technol.*, 2014, pp. 109–116.

[144] S. Malkowsky et al., "The world's first real-time testbed for massive MIMO: Design, implementation, and validation," *IEEE Access*, vol. 5, pp. 9073–9088, 2017.

[145] M. H. Naveed, U. S. Hashmi, N. Tajved, N. Sultan, and A. Imran, "Assessing deep generative models on time series network data," *IEEE Access*, vol. 10, pp. 64601–64617, 2022.

[146] K. Sultan, H. Ali, and Z. Zhang, "Big data perspective and challenges in next generation networks," *Future Internet*, vol. 10, no. 7, p. 56, 2018.

[147] "CRAWDAD." Accessed: Feb. 1, 2023. [Online]. Available: https://crawdad.org/

[148] S. D'Oro, L. Bonati, M. Polese, and T. Melodia, "OrchestRAN: Network automation through orchestrated intelligence in the open RAN," in *Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM)*, 2022, pp. 270–279.

[149] D. Giatsios, "FLEX—Fire LTE testbeds for open experimentation: Flex overview," in *Proc. 3RD Int. Nornet Users Workshop (OSLO)*, 2015, pp. 1–28.

[150] "NITOS—Network implementation testbed using open source platforms" Accessed: Feb. 1, 2023. [Online]. Available: http://nitlab.inf.uth.gr

[151] "5G virtual infrastructure provisioning over NITOS testbed." Accessed: Feb. 1, 2023. [Online]. Available: https://5ginfire.eu/nitos/

[152] K. Kondepu, F. Giannone, S. Vural, B. Riemer, P. Castoldi, and L. Valcarenghi, "Experimental demonstration of 5G virtual EPC recovery in federated testbeds," in *Proc. IFIP/IEEE Symp. Integr. Netw. Service Manag. (IM)*, 2019, pp. 712–713.

[153] *5G Innovation Centre*, Univ. Surrey, Surrey, U.K., 2009. [Online]. Available: https://www.surrey.ac.uk/5gic

[154] J. Costa-Requena, A. Poutanen, S. Vural, G. Kamel, C. Clark, and S. K. Roy, "Sdn-based upf for mobile backhaul network slicing," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, 2018, pp. 48–53.

[155] "Surrey platform." Accessed: Feb. 1, 2023. [Online]. Available: https://5genesis.eu/surrey-platform/

[156] M. Ott, I. Seskar, R. Siraccusa, and M. Singh, "Orbit testbed software architecture: Supporting experiments as a service," in *Proc. 1st Int. Conf. Testbeds Res. Infrastruct. Develop. Netw. Commun.*, 2005, pp. 136–145.

[157] T. Chen, M. B. Dastjerdi, G. Farkash, J. Zhou, H. Krishnaswamy, and G. Zussman, "Open-access full-duplex wireless in the ORBIT testbed," 2018, *arXiv:1801.03069*.

[158] ORBIT. "Open-access research testbed for next-generation wireless networks (ORBIT)." Accessed: May 15, 2023. [Online]. Available: http://www.orbit-lab.org/

[159] PhantomNet. "PhantomNet." Accessed: May 15, 2023. [Online]. Available: https://phantomnet.org/

[160] A. Banerjee et al., "PhantomNet: Research infrastructure for mobile networking, cloud computing and software-defined networking," *Mobile Comput. Commun.*, vol. 19, no. 2, pp. 28–33, 2015.

[161] E. Luther, "5G massive MIMO testbed: From theory to reality," Nat. Instrum., Austin, TX, USA, White Paper, 2014.

[162] J. Vieira et al., "A flexible 100-antenna testbed for massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 287–293.

[163] S. Mattisson, "Overview of 5G requirements and future wireless networks," in *Proc. 43rd IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, 2017, pp. 1–6.

[164] J. Jermyn, R. P. Jover, M. Istomin, and I. Murynets, "Firecycle: A scalable test bed for large-scale LTE security research," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2014, pp. 907–913.

[165] J. L. Jermyn, "Discovering network control vulnerabilities and policies in evolving networks," Ph.D. dissertation, Dept. Comput. Sci., Columbia Univ., New York, NY, USA, 2017.

[166] T. Wirth, L. Thiele, T. Haustein, O. Braz, and J. Stefanik, "LTE amplify and forward relaying for indoor coverage extension," in *Proc. IEEE 72nd Veh. Technol. Conf. Fall*, 2010, pp. 1–5.

[167] "Berlin LTE-advanced testbed." Accessed: Aug. 15, 2022. [Online]. Available: https://www.hhi.fraunhofer.de/en/departments/wn/research-groups/software-defined-radio/research-topics/berlin-lte-advanced-testbed.html

[168] T. Wirth, V. Venkatkumar, T. Haustein, E. Schulz, and R. Halfmann, "LTE-advanced relaying for outdoor range extension," in *Proc. IEEE 70th Veh. Technol. Conf. Fall*, 2009, pp. 1–4.

[169] "5G testbed." Accessed: Feb. 1, 2023. [Online]. Available: https://cewit.org.in/testbed/

[170] Nutaq. "TitanMIMO-6 technology." Accessed: Aug. 15, 2022. [Online]. Available: https://www.nutaq.com/products/titanmimo/titanmimo-6/technology

[171] Nutaq. "TitanMIMO-6 sub 6 GHz massive MIMO testbed product sheet." Accessed: Feb. 1, 2023. [Online]. Available: https://www.nutaq.com/wp-content/uploads

[172] "ESPOO aalto 5G research infrastructure." Accessed: Feb. 1, 2023. [Online]. Available: http://5gtnf.fi/sites/espoo/

[173] "HELSINKI." Accessed: Feb. 1, 2023. [Online]. Available: http://5gtnf.fi/sites/helsinki/

[174] "Vodafone chair mobile communication systems." Accessed: Aug. 15, 2022. [Online]. Available: https://www.vodafone-chair.org/
[175] W. Anwar, S. Dev, K. Kulkarni, N. Franchi, and G. Fettweis, "On PHY abstraction modeling for IEEE 802.11 ax based multi-connectivity networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2019, pp. 1–7.
[176] A. H. Mahdi, K. Kulkarni, N. Franchi, and G. P. Fettweis, "On network deployment for ultra-reliable communication using multi-connectivity," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, 2019, pp. 1–5.
[177] "Vodafone chair & research." Accessed: Aug. 15, 2022. [Online]. Available: https://www.vodafone-chair.org/chair+research#projects
[178] T. R. Newman, A. He, J. Gaeddert, B. Hilburn, T. Bose, and J. H. Reed, "Virginia tech cognitive radio network testbed and open source cognitive radio framework," in *Proc. 5th Int. Conf. Testbeds Res. Infrastruct. Develop. Netw. Communities Workshops*, 2009, pp. 1–3.
[179] T. R. Newman, S. M. S. Hasan, D. DePoy, T. Bose, and J. H. Reed, "Designing and deploying a building-wide cognitive radio network testbed," *IEEE Commun. Mag.*, vol. 48, no. 9, pp. 106–112, Sep. 2010.
[180] "5G playground." Accessed: Feb. 1, 2023. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/5g_playground
[181] "TAMPERE tampere university wireless test networks (Hervanta)." Accessed: Feb. 1, 2023. [Online]. Available: http://5gtnf.fi/sites/tampere/
[182] R. Yasmin, J. Petäjäjärvi, K. Mikhaylov, and A. Pouttu, "On the integration of LoRaWAN with the 5G test network," in *Proc. IEEE 28th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, 2017, pp. 1–6.
[183] "FUSECO playground." Accessed: Feb. 1, 2023. [Online]. Available: https://www.fokus.fraunhofer.de/go/en/fokus_testbeds/fuseco_playground
[184] Fraunhofer Fokus. "5G ready trial, platform." Accessed: May 15, 2023. [Online]. Available: https://www.fokus.fraunhofer.de/en/ngni/projects/5grtp
[185] "Ericsson 5G radio test bed biggest contribution to 5G development in Asia." Accessed: Feb. 1, 2023. [Online]. Available: https://www.ericsson.com/en/news/2015/10/ericsson-5g-radio-test-bed-biggest-contribution-to-5g-development-in-asia
[186] B. Halvarsson, A. Simonsson, A. Elgcrona, R. Chana, P. Machado, and H. Asplund, "5G NR testbed 3.5 GHz coverage results," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, 2018, pp. 1–5.
[187] "SK telecom." Accessed: Feb. 1, 2023. [Online]. Available: https://www.sktelecom.com/index.html
[188] "Korean ICT news, SK telecom opens 5G playground to lead innovation towards 5G commercialization." Accessed: Aug. 15, 2022. [Online]. Available: https://www.netmanias.com/en/post/korea_ict_news/8251
[189] *SK Telecom's 5G Architecture Design and Implementation Guidelines (Version 1.35)*, 5G Tech Lab Corp. R D Center, SK Telecom, Seoul, South Korea, 2015.
[190] "OULU 5GTN (Linnanmaa)." Accessed: Aug. 15, 2022. [Online]. Available: http://5gtnf.fi/sites/oulu/
[191] E. Piri et al., "5GTN: A test network for 5G application development and testing," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, 2016, pp. 313–318.
[192] M. Latva-Aho, A. Pouttu, A. Hekkala, I. Harjula, and J. Mäkelä, "Small cell based 5G test network (5GTN)," in *Proc. IEEE Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2015, pp. 231–235.
[193] "TurboRAN." Accessed: Aug. 15, 2022. [Online]. Available: http://bsonlab.com/TurboRAN/
[194] "OpenAirInterface testbed." Accessed: Aug. 1, 2022. [Online]. Available: https://oailab.eurecom.fr/oai-testbed
[195] C. Y. Yeoh, M. H. Mokhtar, A. A. A. Rahman, and A. K. Samingan, "Performance study of LTE experimental testbed using OpenAirInterface," in *Proc. IEEE 18th Int. Conf. Adv. Commun. Technol. (ICACT)*, 2016, pp. 617–622.
[196] "OpenAirInterface massive MIMO testbed: A 5G innovation platform." Accessed: Aug. 1, 2022. [Online]. Available: https://www.openairinterface.org/
[197] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "OpenAirInterface: A flexible platform for 5G research," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 33–38, 2014.
[198] "Munich experimentation facility site." Accessed: Aug. 1, 2022. [Online]. Available: https://www.5g-vinni.eu/munich-experimentation-facility-site/
[199] T. Heyn et al., "Integration of broadcast and broadband in LTE/5G (IMB5)-experimental results from the eMBMS testbeds," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, 2016, pp. 319–324.
[200] "PerformNetworks testbed." Accessed: Feb. 1, 2023. [Online]. Available: http://morse.uma.es/performnetworks
[201] A. Díaz-Zayas, C. A. García-Pérez, Á. M. Recio-Pérez, and P. Merino-Gómez, "PerformLTE: A testbed for LTE testing in the future Internet," in *Proc. Int. Conf. Wired/Wireless Internet Commun.*, 2015, pp. 46–59.
[202] A. Diaz, C. A. Garcia-Perez, A. Martin, P. Merino, and A. Rios, "PerformNetworks: A testbed for exhaustive interoperability and performance analysis for mobile networks," in *Building Future Internet Through FIRE*. Gistrup, Denmark: River, 2017, pp. 1–250.
[203] "YLIVIESKA Centria University of applied sciences test network." Accessed: Aug. 1, 2022. [Online]. Available: http://5gtnf.fi/sites/ylivieska/
[204] S. Verstichel et al., "Distributed ontology-based monitoring on the IBBT WiLab.T infrastructure," in *Proc. Int. Conf. Testbeds Res. Infrastruct.*, 2010, pp. 509–525.
[205] "w-iLab.t (iMinds)." Accessed: Aug. 1, 2022. [Online]. Available: http://www.crew-project.eu/wilabt.html
[206] S. Bouckaert, W. Vandenberghe, B. Jooris, I. Moerman, and P. Demeester, "The wiLab.t testbed," in *Proc. Int. Conf. Testbeds Res. Infrastruct.*, 2010, pp. 145–154.
[207] "5TONIC: An open research and innovation laboratory focusing on 5G technologies." Accessed: Feb. 1, 2023. [Online]. Available: https://www.5tonic.org
[208] "University of Bristol 5G testbed." Accessed: Feb. 1, 2023. [Online]. Available: https://5ginfire.eu/university-of-bristol-5g-testbed/
[209] "Ericsson D-15 labs." Accessed: Aug. 15, 2023. [Online]. Available: https://www.ericsson.com/en/about-us/experience-centers/d-15/ericsson-d-15-labs
[210] "Accessing the 5G innovation platform as a service (IPAAS) testbed." Accessed: Feb. 1, 2023. [Online]. Available: https://ontario.encqor.ca/accessing-5g-innovation-platform-as-a-service-ipaas-testbed/
[211] Y. Wang, A. Liu, X. Xia, and K. Xu, "Learning the structured sparsity: 3-D massive MIMO channel estimation and adaptive spatial interpolation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10663–10678, Nov. 2019.

Haneya Naeem Qureshi received the B.S. degree in electrical engineering from Lahore University of Management Sciences, Pakistan, in 2016, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of Oklahoma (OU), USA, in 2017 and 2021, respectively. She is currently a Postdoctoral Research Fellow with the Artificial Intelligence for Networks Research Center, OU, where she is managing and contributing to several NSF-funded projects and teaching graduate level courses. She has also worked as an ORISE Fellow with the Center for Devices and Radiological Health, U.S Food and Drug Administration, MD, USA; and has significant industrial research experience in wireless communication with Ericsson Research, CA, USA, and in 3GPP standardization with InterDigital, Inc., New York, USA. Her other current research interests include digital smart healthcare, network automation and combination of machine learning and analytics for future cellular systems. She has also been engaged in system design of unmanned aerial vehicles deployment, channel estimation, and pilot contamination problem in massive MIMO TDD systems.

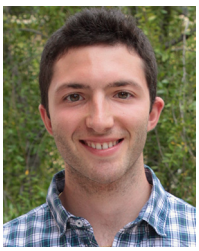
Usama Masood (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in electrical and computer engineering with the AI4Networks Research Center, The University of Oklahoma, USA, where his research focus is on designing novel artificial intelligence-based network modeling techniques for enabling zero touch automation in next generation networks. He is currently working with AT&T Labs, CA, USA, where he is co-leading several projects on network analytics and optimization of AT&T nationwide 5G network. Previously, he worked with T-Mobile USA, where he developed innovative machine learning-based cloud-native applications for RAN automation use cases.

Marvin Manalastas (Graduate Student Member, IEEE) received the B.S. degree in electronics and communication engineering from the Polytechnic University of the Philippines in 2011, and the M.S. degree in electrical and computer engineering from The University of Oklahoma, USA, in 2020, where he is currently pursuing the Ph.D. degree in electrical engineering. He is also affiliated with the AI4Networks Research Center, The University of Oklahoma. Recently, he joined Nokia Standards as a Senior RAN Architecture Research Engineer. He has gained valuable industry experience in cellular network optimization through his work in the Philippines and Japan. He has also completed multiple internships in the USA, including positions as an RF Optimization Intern with Mobilecomm Professionals, TX, USA, an AI/ML Intern with Synopsys, VA, USA, and a Research Fellow with the U.S. FDA, MA, USA. His research interests center on machine learning applied to optimize 5G and beyond networks.

Syed Muhammad Asad Zaidi received the B.Sc. degree in information and communication engineering from the National University of Science and Technology, Pakistan, in 2008, the M.S. degree from Ajou University, South Korea, in 2013, and the Ph.D. degree in electrical engineering from the AI4Networks Research Center, University of Oklahoma in 2021. With almost 15 years' experience in telecom industry, he has worked in Mobilink, Pakistan; KoreaElectronics and Technology Institute, South Korea; MOTiV Research, Japan; ATT, USA; Sprint, USA; and T-Mobile, USA. Currently, he is leading 5G radio-frequency optimization team in T-Mobile networks. His research domain is mobility robustness and optimization of futuristic ultra-dense base station deployment.

Hasan Farooq received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, the M.Sc. degree (by Research degree) in information technology from Universiti Teknologi PETRONAS, Malaysia, and the Ph.D. degree in electrical and computer engineering from the University of Oklahoma, USA, where he was a Postdoctoral Fellow. He is a Senior AI Researcher with Ericsson Research, Santa Clara, USA. His background is AI/ML-driven zero-touch automation algorithms for radio access networks. He has authored/coauthored over 50 publications in high impact journals, book chapters, and proceedings of IEEE flagship conferences on communications. He also has patents in the area of SON algorithms.

Julien Forgeat received the M.Eng. degree in computer science from the National Institute of Applied Sciences, Lyon, France. He is an Artificial Intelligence Principal Researcher with Ericsson Research. In 2010, he joined Ericsson after spending several years working on network analysis and optimization. At Ericsson, he has worked on mobile learning, Internet of Things, and big data analytics before specializing in machine learning and AI infrastructure. His current research focuses on the software components required to run AI and machine learning workloads on distributed infrastructures as well as the algorithmic approaches that are best suited for complex distributed and decentralized use cases.



Maxime Bouton received the M.S. degree in aeronautics and astronautics as part of a double degree between École Centrale Paris and Stanford University, and the Ph.D. degree from Stanford University, where he worked on safety and scalability of intelligent autonomous systems. He is an Artificial Intelligence Researcher with Ericsson Research. His research interests lie in applying reinforcement learning to network optimization problems. He also works on topics related to AI safety, multiagent systems, and decision-making problems with partial observability.

Shruti Bothe is an Artificial Intelligence Researcher with Ericsson. With several years of academic and industry experience and a proven track record of identifying issues in and achieving solutions in domains combining AI/ML to telecommunication networks, she was a main contributor of Ericsson's entrepreneurial effort "Ericsson Routes" that brings autonomous and unmanned vehicles the awareness required to have consistent and reliable connectivity throughout the whole journey. She also heads Ericsson Research's multiyear collaboration with MIT CSAIL related to neuromorphic computing and Lithionics. This research is aimed to produce the next generation of more efficient algorithms and hardware that enable more efficient computing and substantial energy savings. She has several research publications and over 18 pending patents and has recently been honored with a "Key Contributor" Award at Ericsson.

Per Karlsson is the Director of Media Research with Ericsson, focusing on A/V Coding, content analytics, and how new XR experiences will be enabled by the rollout of 5G networks. He is also the Director of Ericsson Research, Silicon Valley, focused on the areas of radio, AI, networking, media, cloud, and security. The Research is performed together with Academia, Customers, Partners, and Universities. His team is currently actively engaged in collaborative projects focused on exploring new opportunities that the 5G networks will bring to the entertainment, manufacturing, and automotive industry. He has been in the industry since 1993 working in the intersection of Research and Products mainly with Ericsson but also leading the Networking Research at the Swedish Research Institute Acreo.

Ali Rizwan received the bachelor's degree in applied and theoretical mathematics and the MBA-IT degree from Bahauddin Zakariya University, Pakistan, in 2006 and 2008, respectively, the M.Sc. degree in big data science from Queen Mary University of London, U.K., in 2016, and the Ph.D. degree from the University of Glasgow, Glasgow, U.K, in 2021. He worked as a Research Scientist with Qatar Mobility Innovations Center, Qatar University, Qatar, from 2020 to 2022. He currently serves as the Chief Technical Officer and the Co-Founder of Artificial Intelligence for Life, Pakistan. His work primarily focuses on the research and development of AI-enabled screening solutions in healthcare.

Ali Imran (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology Lahore, Pakistan, in 2005, and the M.Sc. degree (Hons.) in mobile and satellite communications and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2007 and 2011, respectively. He is a Professor of Cyber Physical Systems with the James Watt School of Engineering, University of Glasgow. He is currently on leave from the University of Oklahoma, where he is a Williams Presidential Professor in ECE and the Founding Director of the Artificial Intelligence (AI) for Networks (AI4Networks) Research Center. His research interests include AI and its applications in wireless networks and healthcare. His work on these topics has resulted in several patents and over 150 peer-reviewed articles including some of the highly influential papers in the domain of wireless network automation. On these topics, he has led numerous multinational projects, given invited talks/keynotes and tutorials at international forums and advised major public and private stakeholders and co-founded multiple startups. He is an Associate Fellow of the Higher Education Academy, U.K. He is also a member of the Advisory Board to the Special Technical Community on Big Data, the IEEE Computer Society.