IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 April 2024, accepted 22 April 2024, date of publication 26 April 2024, date of current version 13 May 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3394312*

RESEARCH ARTICLE

# Cellular Operator Data Meets Counterfactual Machine Learning

**SRIKANT MANAS KALA[1], MALVIKA MISHRA[1], VANLIN SATHYA[2], (Member, IEEE), TATSUYA AMANO[1], (Member, IEEE), MONISHA GHOSH[3], (Fellow, IEEE), TERUO HIGASHINO[4], (Senior Member, IEEE), AND HIROZUMI YAMAGUCHI[1], (Senior Member, IEEE)**

[1]Mobile Computing Laboratory, Graduate School of Information Science and Technology, Osaka University, Osaka 565-0871, Japan
[2]Celona Inc., Campbell, CA 95008, USA
[3]University of Notre Dame, Notre Dame, IN 46556, USA
[4]Kyoto Tachibana University, Kyoto 607-8175, Japan

Corresponding author: Srikant Manas Kala (manas_kala@ist.osaka-u.ac.jp)

This work was supported by the National Institute of Information and Communications Technology (NICT), Japan, through the Commissioned Research (Data Extraction Effort, Predictive Modeling, and Counterfactual Analysis) under Grant 22608. The data collection was supported by the National Science Foundation (NSF) under Grant CNS-1618920.

**ABSTRACT** Unlicensed cellular networks and spectrum-sharing standards assist operators in meeting the ever-increasing demand for mobile data. However, several incumbents are already operational in these frequencies, rendering the wireless environment extremely dynamic and unpredictable. The challenges associated with unlicensed Licensed Assisted Access (LAA) operations in the 5 GHz band and New Radio in Unlicensed (NR-U) in the 6 GHz band are best addressed through a data-driven approach. This requires operator data from current cellular deployments. Further, from an operator's perspective, the precision and reliability of predictive models must be analyzed before deployment. Counterfactual machine learning is ideal for quantifying causal impact in a dynamic, unlicensed cellular environment. However, the literature lacks a framework that combines data-driven solutions, counterfactual analysis, and conventional optimization. This work contributes a dataset from the LAA networks of three major cellular operators in Chicago consisting of 15 features and 9676 samples. Additionally, it proposes a framework for analyzing the performance of unlicensed networks that leverages machine learning for predictive modeling, employs counterfactual analysis for model explainability and network performance enhancement, and utilizes optimization for validation. We show that operator data is necessary to build reliable prediction models for network throughput, and signal strength, among others. Further, the impact of network parameters is shown to differ in unlicensed and licensed cellular network models. Next, a counterfactual machine learning framework is proposed to explain and analyze the predictive models. The framework proposes counterfactual policies to enhance unlicensed cellular network performance. Finally, we validate the suggested counterfactual policies through joint network optimization.

**INDEX TERMS** Unlicensed spectrum, NR-U, cellular networks, operator data, machine learning, counterfactual analysis, explainable AI, optimization.

## I. INTRODUCTION

Licensed spectrum is a limited and expensive resource. Thus, there has been a consistent push from industry and standardization bodies such as 3GPP and ETSI for greater utilization of the unlicensed spectrum by Long-Term Evolution (LTE) cellular networks. Consequently, Cellular operators have deployed Licensed Assisted Access (LAA) services – the first public unlicensed cellular deployments in the 5 GHz band. However, a deeper understanding of the

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

existing spectrum utilization is crucial for fair sharing and facilitating the coexistence of multiple radio technologies in the unlicensed band. For example, Federal operations are the primary incumbent in the Wi-Fi 5 GHz UNII-2 band. Soon, Wi-Fi 6E will coexist with NR in unlicensed (NR-U) in the 6 GHz unlicensed band, and Wi-Fi 7 and 8 will enhance these capabilities by a factor of 40x with multi-link aggregation [1], [2]. Thus, the performance of unlicensed cellular networks in the shared spectrum must be studied with respect to parameters such as bandwidth, signal quality, density, and resource allocation, among others [3], [4], [5].

However, the model-based optimization approach is unsuitable for the analysis and management of unlicensed network performance. The presence of currently operational incumbents in these frequencies, such as military, radar, and navy systems, renders the wireless environment extremely dynamic and unpredictable. Consequently, greater probabilities of transmission conflicts and differing quality-of-service (QoS) requirements present new challenges in harmonious coexistence and spectrum sharing. Despite their versatility, the classical network optimization models seem ill-equipped to offer solutions in real-time in unlicensed systems due to (a) a large number of constraints from coexisting incumbents and (b) much longer computational times required for the non-linear models to converge than licensed cellular systems.

Thus, the typical unlicensed band problems such as fair coexistence, performance prediction, and resource allocation are best solved through a data-driven approach. Machine learning (ML) has emerged as a powerful tool for harnessing data, which makes it suitable for network analysis, performance prediction, and anomaly detection. ML algorithms derive insights from raw data gathered through measurements and are better suited for comparing cellular network scenarios and contexts [6]. For example, two cell selection scenarios were identified in current unlicensed networks, which was not evident from the measurement-based analysis [7]. Network performance optimization can also benefit from data-driven inputs. A hybrid approach that combines AI/ML with a theoretical constraint-based optimization formulation can significantly enhance network performance. For example, feature relationship equations, learned from network data for response variables such as SINR or Throughput, can serve as constraints in an optimization model [8], [9]. Thus, the publicly available LAA dataset will pave the way for new research on data-driven network optimization.

Furthermore, network operators already employ AI/ML to optimize their processes, such as cell selection [10]. However, AI/ML models can often act as a black box, and operators must understand how network data shapes the performance of prediction and classification models. Thus, model *explainability* is becoming essential to analyze the parameters most significantly affecting network performance. Creating data-driven policies to configure these parameters enables operators to achieve optimal performance.

This work employs counterfactual machine learning (CFML) to analyze network performance prediction models. Counterfactual analysis is a great machine learning paradigm for an under-the-hood understanding of AI/ML models as it offers causal inference through ''what-if'' scenarios [11], [12]. The proposed counterfactual framework explains the role of critical network Quality of Service (QoS) indicators in performance prediction, offers alternative network policies that will enhance network performance, and validates these policies through joint network optimization.

## II. MOTIVATION AND CONTRIBUTIONS

Developing machine learning systems for network analysis and optimization requires data from network deployments.

### A. MOTIVATION AND RESEARCH PROBLEMS

Without access to operator data, it is difficult to investigate how efficiently the spectrum is being used and to identify the practical challenges to fair coexistence. Thus, access to cellular network data is vital for the research community. Unfortunately, the democratization of data access is constrained by geographical proximity to state-of-the-art networks and the high cost of network monitoring tools and applications.

Moreover, researchers face typical challenges in data sharing, including but not limited to data management, data security, and regulatory constraints. The challenges in data management include collation, sanitization, and making it readily available for analysis. Data security is important as there may be a concern that network data may hint at the strategic business side operations of cellular operators. For example, number of active subscribers can be estimated through RB allocation. Therefore, it may be necessary to anonymize operator data. Finally, data from primary incumbents may be subject to restrictions by authorities.

Furthermore, although cellular operators and industry professionals have access to their own network data, this creates data silos that reduce the benefits of collaborative research. Overcoming these barriers to democratic and universal access to cellular network data requires low-cost gathering, accurate extraction, and periodic release of data.

As machine learning systems are being deployed to inform decisions that have a real-world impact, it is imperative to not only understand their decision-making process but also be able to provide satisfactory explanations to the people affected by the decision. To provide a more comprehensive understanding of how captured data features can affect the decision-making process, we apply interpretability mechanisms to elucidate the black-box models using counterfactuals [13]. This section describes how we leverage counterfactual explanations for the classification problems defined on our dataset.

Counterfactual explanations simplify the understanding of complex machine learning models by showing 'feature-perturbed versions' of a sample that would result in a different (opposite, or targeted) outcome. Counterfactuals explain the

model output by providing "what-if" explanations. Our study utilizes counterfactuals due to their independence from classifier decision boundaries and their capacity to directly reflect model predictions following feature adjustments. This makes counterfactual examples more human-interpretable than other explanation methods. CFML has been applied by researchers as an optimization technique for wireless power control [14], agreement violations [15] and cellular responses [16].

### B. CONTRIBUTIONS

This work aims to solve the above problems through the following contributions. The first major contribution of this work is the gathering and extraction of cellular network data for three operators and the release of a public dataset of close to ten thousand samples on LAA networks [17]. It then utilizes several ML algorithms to analyze and predict the network performance of unlicensed cellular networks through reliable prediction models for network throughput, resource allocation, signal strength, and more. Further, it shows that system parameters such as the number of carriers, modulation coding scheme, and channel quality indicator can also be determined with high confidence. Through feature importance techniques, we demonstrate that important parameters with the highest impact on a cellular network QoS differ in different network environments, such as LTE-only, Licensed LTE-LAA, or LAA. It also shows that the network environment can be predicted with near-perfect accuracy, which is particularly useful for device-initiated cell selection. This paper then presents a counterfactual machine learning framework that introduces an element of explainability to the ML models utilized for predictive analyses. The CFML framework offers potential network configurations that can enhance network performance and end-user QoS. Finally, the counterfactual outcomes are validated through a joint optimization model proposed in this work that maximizes resource allocation with specific constraints. To the best of our knowledge, this is the first study on unlicensed cellular networks to propose a counterfactual framework that uses real-world network data and is validated through well-known techniques such as optimization.

Please note that cellular operators have limited access to network data of coexisting systems (Wi-Fi and other incumbents), and even less control over them to optimize performance. Thus this work focuses on collecting and analyzing unlicensed cellular (LAA) data. The objective is to find actionable insights that operators can benefit from while deploying 5G NR-U networks.

### C. PAPER ORGANIZATION

The rest of the paper is organized as follows. Section III, presents the recent developments and opportunities in the unlicensed band focusing on allocation and technologies. Thereafter, Section IV outlines the relevance and need for an LAA dataset and the challenges in creating such a dataset.

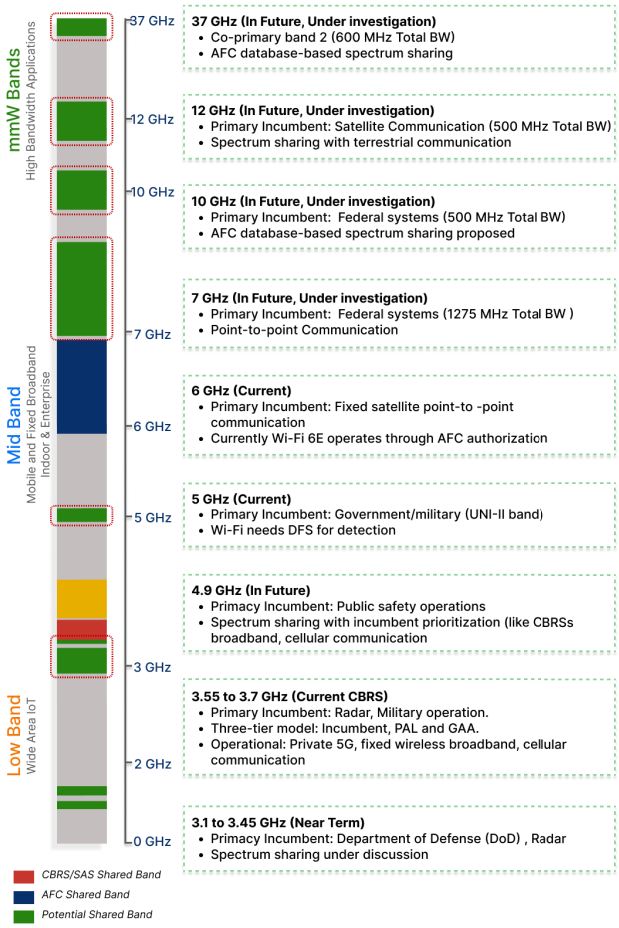



**FIGURE 1.** Opportunities in the Unlicensed Spectrum.

Section V discusses the methodology of the paper in stepwise detail. Section VI presents the cellular data collection and extraction process in great detail. The extracted cellular data is then used for network analysis and performance prediction in Section VII. Next, a counterfactual framework is proposed in Section VIII that analyzes prediction models and suggests alternate policies for performance enhancement. Outcomes of the proposed counterfactual framework are validated through conventional network optimization model in Section IX. Finally, Section X summarizes the major contributions and findings of this work.

## III. THE UNLICENSED BAND

The opportunities in the unlicensed band are presented in Figure 1. Unlicensed cellular networks and spectrum sharing are two paradigms that can help operators tap into this unharnessed potential.

### A. UNLICENSED CELLULAR NETWORKS

To ensure the fair coexistence of existing incumbents with unlicensed cellular operations, two LTE-WiFi coexistence standards were prescribed and adopted: *LTE license assisted access* (LTE-LAA) proposed by 3GPP and *LTE in unlicensed*

*spectrum* (LTE-U) proposed by industry forum. The Federal Communications Commission of the United States of America (FCC) deregulated 500 MHz spectrum in the 5 GHz band for unlicensed cellular operation in coexistence with Wi-Fi. After a prolonged LTE-U vs. LAA debate, cellular operators globally backed LAA. LAA was hailed as a 3GPP benchmark for the fair coexistence of LTE with existing incumbents, such as Wi-Fi networks in the unlicensed bands. It also became a successful technological precursor to the 5G New Radio-Unlicensed (5G NRU) [18]. Encouraged by these outcomes, toward the end of 2018, the FCC issued an NRPM for unlicensed operation in the ''greenfield'' 6 GHz band (5925-7125 MHz) [1], [19]. It prescribed guidelines for the unlicensed coexistence of cellular services with existing incumbents, such as Wi-Fi access points (APs) in the 6 GHz band (5925-7125 MHz). These prescriptions were adopted in April 2020. Likewise, the European Commission allocated 480 MHz (5925- 6725 MHz) of spectrum in the 6 GHz band to harmonize the unlicensed coexistence of cellular and Wi-Fi systems. Industry leaders such as Qualcomm expect more unlicensed bands, shared spectrum standards, and technologies like mmWave to be available [19]. Thus, researchers in both industry and academia are investigating deployment scenarios far more complex than the current 5 GHz unlicensed operation. These include:

- Carrier aggregation between licensed band NR Primary Cell (P-cell) and NR-U Secondary Cell (S-cell). This includes (i) Downlink (DL) and Uplink (UL) on NR-U S-cell (ii) Only DL on NR-U S-cell.
- Dual connectivity between licensed LTE P-cell and NR-U P-cell/S-cell.
- Dual connectivity between licensed band NR P-cell and NR-U P-cell/S-cell.
- Unlicensed band DL and licensed band UL within the same NR cell.
- NR-U standalone operation.

### B. SPECTRUM SHARING

Traditionally, regulators have allocated spectrum to mobile operators only after clearing out the incumbent users. However, in addition to coexistence, the unlicensed bands can also be utilized through spectrum sharing where the primary incumbent(s) or primary user(s) can be military communications, radar or satellite transmissions, and Broadcast Auxiliary Services (BAS). During times or in areas where the primary incumbent is inactive, secondary incumbents/users can operate in the medium. Examples of secondary incumbents include Wi-Fi in 5 GHz UNII-2 or Citizens Broadband Radio Service in 3.5 GHz (3550 MHz to 3700). Three spectrum-sharing mechanisms are operational or under consideration [5], [20]. The most prominent is the Citizens Broadband Radio Service (CBRS) approach, currently operational in the 3.5 GHz in the US. The other two are Licensed Shared Access and Concurrent Shared Access, such as *club licensing* [20].

CBRS uses dynamic sharing to support three tiers of prioritized or controlled access to the spectrum. The highest-priority tier with the most protection comprises the incumbents such as radars and satellite services. Prioritized Access License (PAL) holders form the secondary tier. PALs purchase the rights to use the available spectrum (up to a maximum of 40 MHz) when the top-tier incumbent is not using it. The lowest tier offers General Authorized Access (GAA) to any service willing to use the spectrum when available with the least protections. In areas where the top-tier incumbent is not utilizing the spectrum, PAL and GAA tiers get access to reserved portions of the spectrum. Further, the FCC has mandated all tiers to look up the Spectrum Access System (SAS) database which facilitates the spectrum sharing model by regulating and managing access. In case a band is not registered as being used in the SAS database, PAL and GAA tiers can access each other's reserved portions in the band.

Second, is the Licensed Shared Access [20]. It has a two-tiered structure where primary incumbent(s) are licence holders who can sub-license the spectrum to secondary service providers such as mobile operators. The secondary tier can use the shared spectrum when not in use by the incumbent. The first such spectrum-sharing model was operational in Europe in the 2.3 GHz band and more sophisticated models are under development.

The third spectrum sharing mechanism is the Concurrent Shared Access such as *club licensing* [20]. Unlike the first two mechanisms, this approach considers a single tier of users and permits them to coordinate and share the spectrum. Thus, mobile operators can share spectrum to enhance the quality of services (QoS) and overall spectrum-usage efficiency.

### C. OPPORTUNITIES IN THE UNLICENSED BAND

The successful initial implementations of the above two paradigms have encouraged regulators, standardization bodies, and industry organizations to initiate discussions on opening other underutilized frequencies for coexistence and sharing. A detailed overview of ongoing discussions with band-specific highlights is presented in Figure 1.

Although high-band terahertz range frequencies such as mmWave have fewer incumbents and offer high bandwidths, the midband frequencies offer a more balanced combination of transmission range and bandwidth. Consequently, most potential bands under consideration are in the 1 GHz – 12 GHz range, essentially making the mid-band the primary driver of coexistence and spectrum-sharing systems. However, it will also make it the most crowded, exacerbating the challenges and bottlenecks observed in the 5 GHz unlicensed operation. The Federal Communications Commission (FCC) in the U.S. recently created rules for the 6 GHz band that would allow unlicensed services to coexist with existing incumbents in the band, mainly high-power fixed microwave links and low-power broadcast auxiliary services. It is expected that in addition to WiFi, this band will

**TABLE 1. Cellular networking monitoring tools used by researchers.**

| Monitoring App | Subscription Cost | Iperf Traffics | Root Access | QXDM Logs | 4G and 5G | CSV Export | Spectrum |
|---|---|---|---|---|---|---|---|
| XCAL [21] | High | Yes | Yes | Yes | Yes | Yes | Licensed and Unlicensed |
| Qualipoc [22] | High | Yes | Yes | Yes | Yes | Yes | Licensed and Unlicensed |
| NSG [23] | High | Yes | Yes | Yes | Yes | Yes | Licensed and Unlicensed |
| SigCap [18] | Open-source | In Progress | No | No | Yes | Yes | Licensed and Unlicensed |
| CellInfo [24] | Open-source | No | No | No | Yes | Yes | Licensed and Unlicensed |
| FCC [25] | Open-source | Yes | No | No | Yes | Yes | Licensed and Unlicensed |

also be used by cellular systems deploying 5G NR-U, similar to the 5 GHz band by LAA.

Let us consider the recently deregulated 1200 MHz spectrum (5925 MHz–7125 MHz) in the UNII bands 5, 6, 7, and 8 in the US. In the 6 GHz band, apart from Wi-Fi 6E, the unlicensed cellular services will coexist with other existing incumbents, primarily the high-power fixed point-to-point microwave services and lower-power BAS. Despite these additional constraints, reliable and improved QoS is expected from the unlicensed services. To that end, IEEE and 3GPP organized a workshop on coexistence networks in 2019 to discuss existing challenges and propose feasible solutions for the next generation of standards covering Wi-Fi 6E and 5G NRU operation in the 6 GHz band.

Thus, the current deployments in the 5 GHz band serve as the best training ground for improved 6 GHz unlicensed and spectrum-sharing operations. Lessons learned from LAA deployments will pave the way for robust and low-friction spectrum-sharing systems enabling NR-U operation in 6 GHz and beyond. However, these lessons can be learned only through collaborative research, for which access to network data is of utmost importance. Likewise, without LAA network data analysis, empirical ground truth regarding the efficiency of various network mechanisms will be hard to ascertain. Findings such as the type of services coexisting in the unlicensed band or the impact of cell selection on LAA performance [8], [26] would not have been possible without access to network data.

## IV. LAA DATASET: RELEVANCE AND CHALLENGES

### A. RELEVANCE AND VALUE-ADD

The released LAA deployment data [17] was collected from different areas of Chicago. This includes the three University campuses viz., the University of Chicago, the University of Illinois at Chicago, and the Illinois Institute of Technology. Further, data was also gathered from downtown areas such as the Loop, South Loop, and River North. Chicago is considered for the data collection exercise, as all three major operators, viz., AT&T, Verizon, and T-Mobile, had deployed their LAA networks, which allows for a comprehensive sample space. A diverse sample space from multiple operators allows for a comprehensive evaluation of unlicensed band processes such as cell selection and handover. The dataset consists of LAA with LTE, and LTE-only datasets with variations in the number of available carriers for all three operators. It will add much value by helping the broader research community (a) Identify new challenges in unlicensed coexistence and spectrum sharing and (b) Propose data-driven solutions for existing and future unlicensed networks. Innovative AI/ML-based solutions may also be included in future specifications by standardization bodies. The LAA network dataset we are releasing may potentially complement the primary incumbent network database. This seems necessary according to the report of the Global System for Mobile Communications Association, the international body that represents the interests of cellular operators. GSMA states, "(Spectrum) Sharing will only be useful for operators if the proposed band is harmonized for mobile use." [20]. It also calls upon the regulators for simple and investment-friendly coexistence and sharing frameworks that (a) Support reliable and high cellular QoS (b) Allow operators to voluntarily share their spectrum, and (c) Incentivise incumbents to share unused bands with high demand from other users [20]. To achieve these objectives, regulators and primary incumbents can use the released dataset. It will help develop a more comprehensive understanding of spectrum sharing, vis-a-vis other operational wireless technologies. Further, the gathered data is from multiple operators, mitigating the problem of data silos. It may also encourage government entities to release their data in the unlicensed or shared spectrum in the spirit of collaborative research. Open-source applications like Sigcap, CellInfo, and FCC APP offer some Phy layer information such as RSRP, RSRQ, EARFCN, and PCI. However, for AI/ML-based network analysis, detailed and accurate network information (e.g., Resource Block (RB), SINR, Throughput) is required, which is difficult to get at scale. Thus, creating such a dataset is a non-trivial exercise.

### B. CHALLENGE: EXPENSIVE NETWORK MONITORING APPLICATIONS

To determine the efficiency of spectrum usage by an operator through AI/ML models, values of essential network parameters (features) are required, e.g., SINR, Throughput, RB, Channel Quality Indicator (CQI), and Modulation Coding Scheme (MCS). Only a handful of applications can extract features such as RB, Channel Quality Indicator (CQI), and Modulation Coding Scheme (MCS) on mobile devices with the latest chipsets that need to be rooted. Thus creating an LAA dataset at scale is a non-trivial exercise. A list of subscription-based and open-source tools, along with their features, is presented in Table 1. Paid applications generally

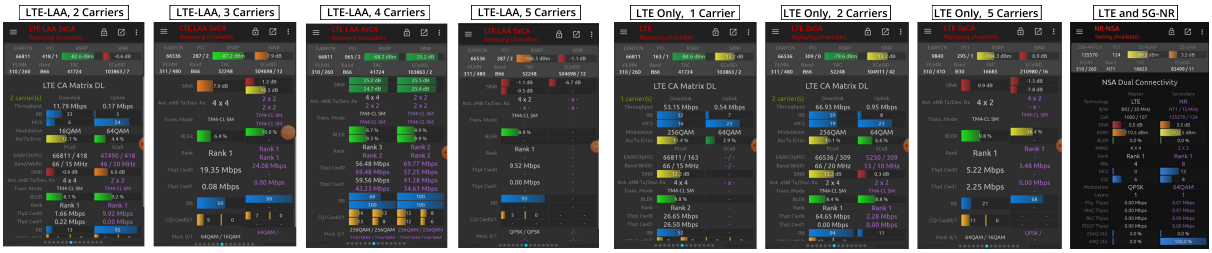



**FIGURE 2.** Relevant Data screen-types of NSG App.

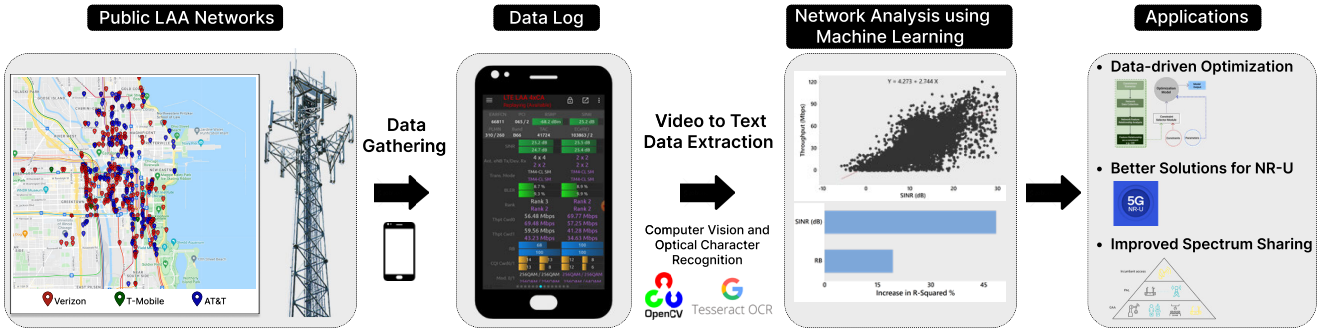



**FIGURE 3.** Data gathering and extraction for AI/ML Analysis.

offer tiered freemium subscriptions for monitoring and exporting network data. The free versions have limited utility, e.g., Network Signal Guru (NSG) does not currently allow 5G-NR monitoring in the free tier. The basic subscriptions come with monitoring capabilities, ranging from tens to hundreds of dollars per month. However, for AI/ML analysis and modelling, network data is required at scale, typically in multiple thousands of samples. To be able to export the gathered data into a format suitable and sufficient for AI/ML analysis these applications charge up to tens of thousands of dollars every year. A plausible reason for the high cost is that cellular operators are secretive about their data as it could be leveraged to draw inferences on their proprietary technology and strategic processes [26]. There may also be a concern that network data may hint at their strategic business side operations, e.g., number of active subscribers through RB allocation, or cell density through SINR or RSRP.

Most chipset vendors, such as Qualcomm and cellular operators themselves, are often equipped with state-of-the-art network monitoring tools such as QCAT, QXDM, and other applications. On the other hand, academic researchers, especially from the developing world, are denied access to data from the latest unlicensed deployments due to the exorbitant cost of the applications. It is noteworthy that Mobile Insight [27] and SigCap [18] are very useful open-source applications. SigCap, in particular, is easy to install, displays real-time network information on the mobile device, and allows passive network monitoring [18]. We chose NSG, as it delivers precise information on a

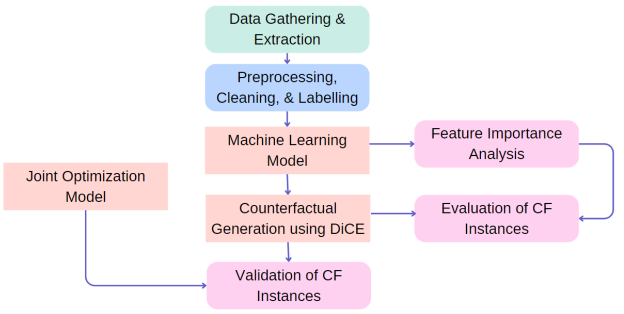



**FIGURE 4.** Methdology of the study.

larger set of PHY layer parameters and has an overall better interface. Thus, NSG offers a better user experience and functionality for LAA data collection at scale. It also allows for capturing mobility data within the background mode with high stability. Nevertheless, expensive subscriptions to NSG and other applications make it virtually impossible for the wider research community to have access to LAA deployment datasets. Our innovative computer-vision-based solution makes it possible to overcome this obstacle by extracting data from NSG screens (even in the free tier) and releasing it to the broader community for AI/ML analysis.

## V. METHODOLOGY

This study introduces a novel Counterfactual Framework designed to enhance cellular network performance through advanced explainable machine learning techniques. The

methodology encompasses four primary components: data collection, the establishment of a machine learning model, the generation of initial predictions, and the application of counterfactual analysis to identify potential performance improvements. The detailed process followed in the methodology is presented in Figure 4 and explained below.

The cellular network data is collected across downtown Chicago, focusing on LAA deployments from three major operators: AT&T, T-Mobile, and Verizon. Data collection is conducted using both stationary and mobile devices to capture a wide array of network conditions. The Network Signal Guru (NSG) application, recognized for its comprehensive coverage of cellular standards, such as LTE, LAA, and 5G-NR, is the primary data collection tool. Following collection, we perform a meticulous extraction process. A computer vision and optical character recognition (OCR) based system is developed to process encrypted logs into a structured format suitable for analysis. This innovative approach allows for the extraction of detailed network parameters, including bandwidth, signal strength, Physical Cell ID (PCI), allocated resource blocks (RB), throughput, Modulation Coding Scheme (MCS), and Rank among others, thereby overcoming the limitations of NSG's subscription model for data export.

We perform exploratory analyses on the extracted data and after a thorough cleaning and preprocessing phase, we utilize it for machine learning prediction models. The dataset is divided based on different network environments to examine the effect of network parameters on performance metrics through a data-driven approach. The models utilize the Number of Carriers, Antenna Configuration (ANT), Transmission Rank (TRANS), Channel Quality Indicator (CQI), Modulation and Coding Scheme (MCS), Signal-to-Noise Ratio (SINR), Resource Blocks (RB), Block Error Rate (BLER), and Throughput. In our research, we have employed a suite of low-cost machine learning algorithms that are well-suited for our data characteristics and the objectives of the study. These include regression and classification algorithms, such as Ordinary Least Squares Linear Regression, Ridge Regression, Kernel Ridge Regression, Polynomial Regression, K-nearest neighbours (KNN), Random Forests (RF), Decision Trees (DT) and Support Vector Machines (SVM). We train and evaluate these models on segmented data Our dataset is segmented into various subsets reflecting different network scenarios, which helps in understanding the model's performance under varied conditions- Combined (entire dataset), LTE (Standalone LTE network data), LAA-L (LTE coexisting with LAA), LAA-U (Unlicensed LAA), and LAA (LAA-L and LAA-U combined). Each model is trained on 80% of the data from each segment, ensuring that it learns to predict or classify based on a comprehensive set of examples, and then tested on the remaining 20%. We also explore the impact of different features in the decision-making process through a feature importance study.

The next step of our analysis is the application of counterfactual analysis to explore potential network configurations that could lead to performance enhancement. This process involves generating counterfactual instances for each data point in the dataset, evaluating their validity, and assessing their feasibility and proximity to original instances. The counterfactual framework utilizes binary classification and empirically estimated means as the threshold, simplifying the prediction of SINR and throughput as network performance prediction metrics. This data-driven counterfactual approach offers a deep understanding of the impact of network parameters on performance. Classified samples are input to the counterfactual generator model, using multiple classifiers to generate policies and scenarios for maximal signal strength and throughput. The proposed framework uses the DiCE CF generator to incorporate diversity and proximity into the synthesized counterfactual instances through a unified loss function. It also learns the interrelationships between network parameters and performance metrics suggesting potential improvements in signal strength and reduction in error rates that could boost overall throughput. We also cross-validate our findings with the feature importance results generated from the classifiers.

The final step is to validate our findings from the counterfactual analysis using a joint optimization model. The joint optimization model presented in the documents aims to maximize radio resource allocation efficiency in a multi-point network comprising LTE (Long-Term Evolution) and NR (New Radio) technologies. The objective of the model is to maximize the sum of the rewards for all devices in the network by allocating radio resources effectively. This validation step demonstrates that the counterfactual generation and selection step leads to policies for network parameter configuration that will yield optimal network performance and end-user experience.

## VI. DATA COLLECTION, EXTRACTION, AND EXPLORATORY ANALYSIS

The data was collected in different areas of downtown Chicago. LAA deployments of three major cellular operators, viz., AT&T, T-Mobile, and Verizon, were considered. Data were collected with both stationary and mobile devices. The overview of the data gathering and extraction process is shown in Figure 3. The initial observations were made using multiple tools, some are presented in Table 1. Network Signal Guru (NSG) application developed by Qtrun Technologies was selected as the primary data gathering tool for this dataset [23]. The ''Data Log'' in Figure 3, shows one of the many screens of the NSG app, which is discussed in detail ahead. NSG supports multiple cellular standards such as LTE, LAA, and 5G-NR. Some of the relevant data screens for LTE-only, LTE-LAA, and 5G-NR are presented in Figure 2. It's discernible that NSG provides more detailed information than open-source alternatives such as SigCap or FCC Speed Test. This includes information on bandwidth, signal strength, Physical Cell ID (PCI), resource blocks allocated (RB), throughput, Modulation Coding Scheme (MCS), Rank, and others. While it may suffice to make a

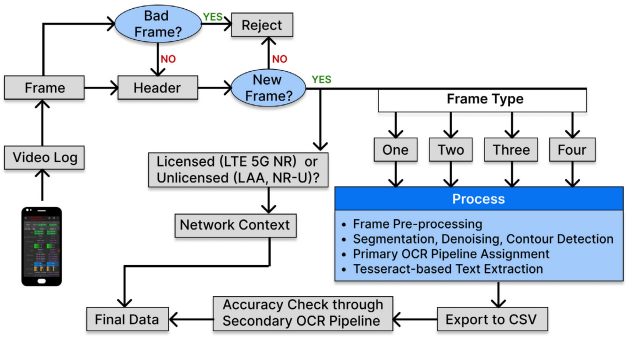



**FIGURE 5.** Schema of the data extraction process.


**FIGURE 6.** Challenges in data extraction.

note of the observed data manually for network monitoring purposes, data-driven network solutions require thousands of samples for reliable modelling and prediction of network performance. Unfortunately, NSG doesn't permit extracting this data in the free tier of its freemium model. In the basic subscription tier, NSG allows capturing the monitoring session as an encrypted log in DLF format which can only be decoded by the NSG application on another device with a basic subscription. However, retrieving data in a file format desirable for data analysis, such as ".CSV" or ".txt," is only possible with a more expensive subscription. This prevents easy and affordable access to PHY layer data for AI/ML modelling and analysis.

To extract LAA data at scale for research and make it available to the broader research community, we engineered a solution based on computer vision and optical character recognition. The high-level schema of the data extraction system is shown in Figure 5. The rendering of the encrypted log was converted into a video and then into individual frames. Next, duplicate and undesirable frames were filtered out to speed up the extraction process. Unique frames were processed through the extractor using multiple pipelines dedicated to the number of carriers. Thereafter, the frames were pre-processed and subjected to various image processing techniques for better extraction performance. Finally, values for several network parameters on each frame were captured through computer vision and a deep-learning-based Tesseract OCR engine. Two OCR pipelines were considered with different image processing techniques for verification, and a field-specific comparison was made. Finally, the extracted data was exported into a format such as CSV or Excel that can serve as input to the data processing pipeline of an AI/ML model.

Building the extraction system was a difficult task for several reasons. The constant change in network parameters causes frequent toggling of the NSG screen. Likewise, the constant fluctuations in parameter values drastically alter the user interface making the frame too dynamic for a generic solution to process accurately. Our LTE-LAA monitoring logs consist of seven different frame screens/layouts with similar fields. Thus network context, such as LTE or LTE-LAA or 5G-NR, number of carriers, needs to be learned on the fly and data must be extracted based on the context. Further, most parameter fields on the NSG frame are unique entities from the perspective of image processing and OCR pipelines, requiring specific techniques for near-perfect accuracy. A sample frame highlighting the major challenges to high extraction accuracy, numbered from one to six, is presented in Figure 6. The challenges include (1) Different background colours with varying levels of overlap for a single field requiring fine-tuned image processing and contour detection (2) Different text colours requiring different image-processing pipelines (3) Elements of UI overlapping parameter values requiring custom-tailored denoising (4) Identifying narrow strips of data fields in a toggling frame (5) Partial or cut-off fields (6) Multiple problems in the same field: missing data, different colours, varying background overlaps over values and noise.

The data extractor shown in Figure 5, was designed to overcome these constraints, ensure high accuracy, reduce data wastage, work across network types and frame types, and allow extraction at scale. Despite the multiple challenges, the extraction system works remarkably well with close to 100% accuracy. It also weeds out irrelevant and duplicate frames, speeding up the extraction process. A total of 9676 samples are extracted with the following twelve network parameters, viz., network type, number of carriers, the band, SINR, Ant. eNB Tx, Trans. Mode, BLER, RANK, Throughput Cwd0/Cwd1, RB, CQI Cwd0/Cwd1, and MCS [17]. A small subset of the dataset has helped identify performance bottlenecks in existing LAA deployments and prescribe potential solutions [8], [26]. The dataset in its current form includes seven additional network parameters and has a much larger sample size [17].

### A. EXPLORATORY DATA ANALYSIS

This section presents an analysis of several important network parameters in different types of cellular networks. The distributions of nine important network variables viz., SINR, RB, BLER, Throughput, Throughput (CWD0 and CWD1), Number of carriers, and CQI (CWD0 and CWD1) are

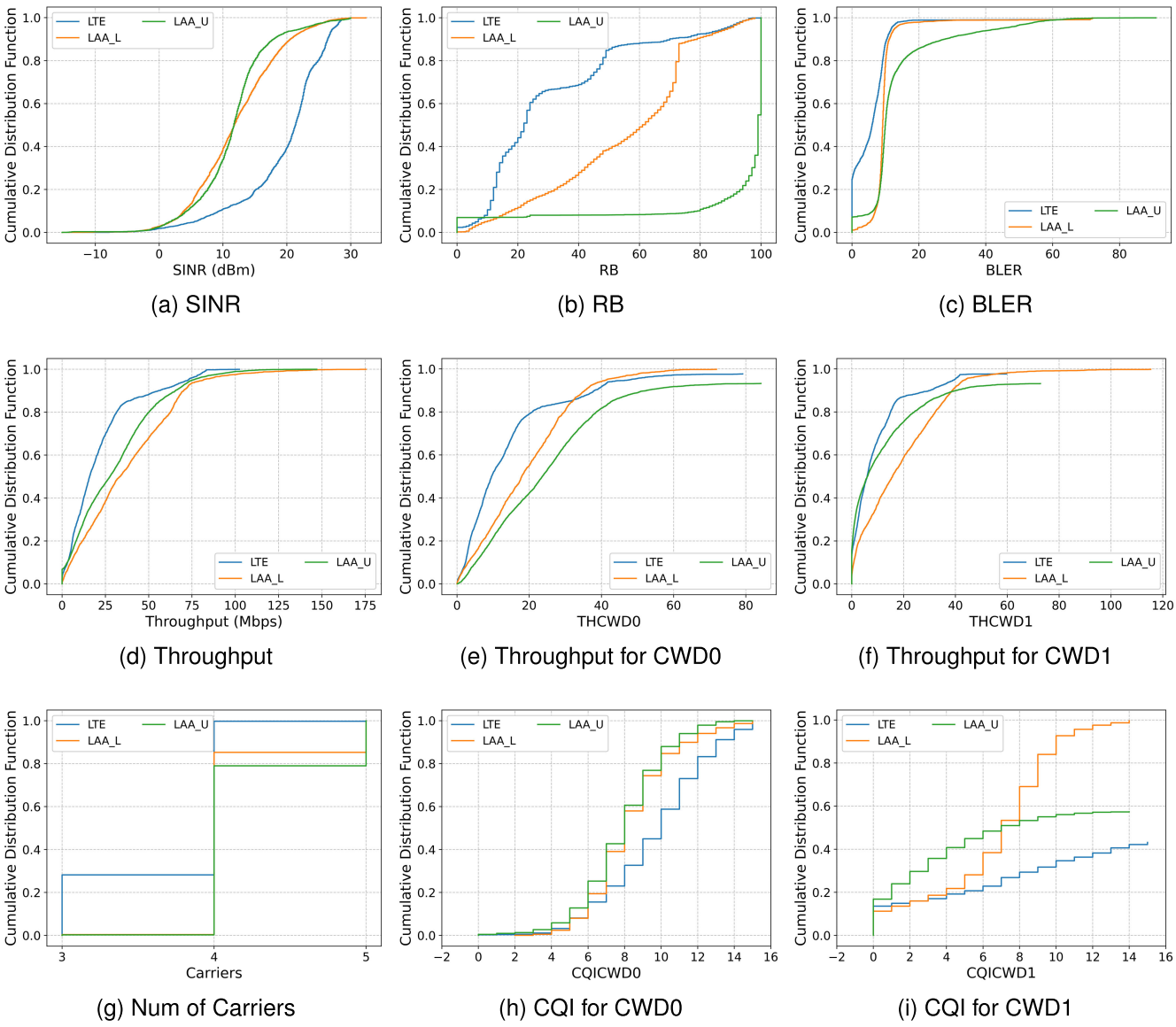



**FIGURE 7.** Distribution of important network variables.

presented in Figure 7. Below, the performance of LTE, $LAA_U$, and $LAA_L$ is analyzed in terms of SINR, RB, BLER, and Throughput.

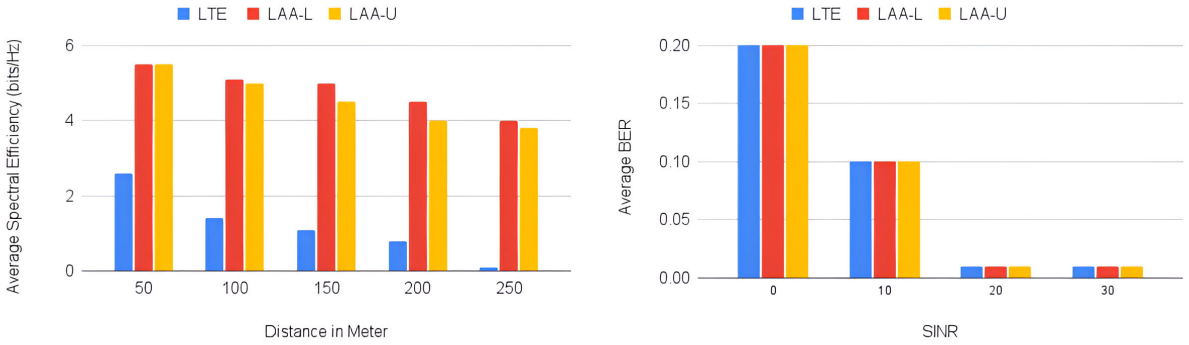
### 1) SINR

Only in the LTE scenario, all transmissions are in the licensed spectrum. As the specific operator exclusively controls a licensed band, they are carefully used in the RF planning, reducing the co-channel interference. Hence, a higher SINR can be observed for LTE in Figure 7 (a). However, for $LAA_L$, where Licensed (LTE) exists with Unlicensed (LAA) carriers, the SINR is slightly lower. The reason is that the unlicensed spectrum is non-exclusive *i.e.,* not clean. Every Wi-Fi AP and LAA small cell of other operators is free to transmit in the same spectrum. This invariably leads to more co-channel interference and lower SINR for $LAA_L$. Further, for pure LAA ($LAA_U$), the impact of co-channel interference is much higher than $LAA_L$, resulting in even lower SINR for $LAA_U$. Since SINR is a critical determinant for the channel quality (CQI), a similar trend can be observed for CQI on both antenna ports CWD0 and CWD1[1] as shown in Figure 7 (h) & (i).

### 2) RESOURCE BLOCK ALLOCATION

Typically, there is no relation between SINR and RB. The RB allocation entirely depends upon the bandwidth available per radio or base station. In LTE, obtaining a greater chunk of the licensed spectrum is difficult as the mid-band frequencies are scarce, and the licenses are costly. This is the primary

[1]The average of CWD0 and CWD1 will resemble the SINR plot.




(a) Spectral Efficiency

(b) BER Comparison

**FIGURE 8.** Spectral efficiency of LTE, LAA-L and LAA-U.

reason that operators rely on the LAA unlicensed spectrum. It is relatively much less expensive to operate but is not a clean spectrum like the licensed band. Hence, operators optimize the network performance by diverting certain traffic transmissions on the LAA network. Thus, it is unsurprising that there is a higher RB allocation in $LAA_U$ and $LAA_L$ than in LTE, as observed in Figure 7 (b). A much higher allocation in $LAA_U$ than $LAA_L$ could be attributed to fewer LAA devices being connected to the small cells and a higher number of carriers in $LAA_U$ as can be seen in Figure 7 (g).

### 3) BLER

The high BLER in the network leads to more packet drops and, eventually, to more re-transmissions. This mainly depends on the same RB allocation on the same frequency or the same channel nearby. The impact will be higher in the LTE scenario because these transmissions are possible on the high-power Macro network, where the transmission can reach several miles or kilometres. The adverse impact of the same channel interference affects the nearby transmitter, which eventually increases the BLER on the system, as shown in Figure 7 (c). On the other hand, in $LAA_U$ than $LAA_L$, the LAA operations are low-power small cells, and the contribution from the neighbouring cell transmission may be lesser compared to the LTE scenario.

### 4) THROUGHPUT

The throughput performance depends upon the number of RBs allocated and the SINR. Data shows that fewer RBs are allocated in $LAA_L$ as compared to $LAA_U$. However, the SINR is high for $LAA_L$ compared to $LAA_U$. This in turn increases the modulation coding scheme (MCS), *i.e.,* which helps the base station to push more bits. This will increase the overall system performance *i.e.,* throughput as shown in Figure 7 (d). Though the SINR is good in the LTE scenario, the RB allocation is less for the LTE, so the throughput is low compared to $LAA_L$ and $LAA_U$. Figure 7 (e) and Figure 7 (f) show the throughput allocation on each antenna port *i.e.,* CWD0 and CWD1. Figure 7 (d) shows the average allocation from Figure 7 (e) and Figure 7 (f).

### 5) SPECTRAL EFFICIENCY

We determine the spectral efficiency based on the throughput observed with respect to channel bandwidth in Hz. Figure 8 (a) shows the average spectral efficiency for LTE, $LAA_L$ and $LAA_U$. Small cells like LAA have different and often lower coverage than the LTE Macro cell. Further, depending on the number of users connected to the base station, the throughput received by the user will vary over time. We notice that LTE has low spectral efficiency, which is due to the urban deployment setting where we can expect more users to connect to the base stations. Typically, it is difficult to determine the user density through real-time measurement. For LAA, during the time of the experiment, there were only a few LAA-capable devices and most of the time our devices were the only user devices connected to the LAA base station (this can be verified by looking at the max RB allocation on the rooted NSG device). Hence, for $LAA_L$ and $LAA_U$ scenarios, the spectral efficiency is higher compared to the LTE. Also, the range of cells is comparatively smaller in LAA, which is a contributing factor to the users in that range transmitting with higher MCS. This, in turn, leads to high throughput and spectral efficiency.

Careful RF planning with good SINR and optimal resource block allocation can significantly improve the spectral efficiency of the network by maximizing the throughput. Adverse impacts of channel impairment, such as noise, fading, and attenuation, can be mitigated by optimal placement of the radios, focused beam or MIMO transmission, power control, and electric tilt adjustment of the radio antennas.

### 6) BIT ERROR RATE (BER)

Figure 8 (b) shows the average BER based on the SINR thresholds. The number of bit errors is calculated depending on the SINR range experienced by mobile devices. A higher SINR represents the UE is within a good signal range to effectively decode the symbols or bits transmitted, which

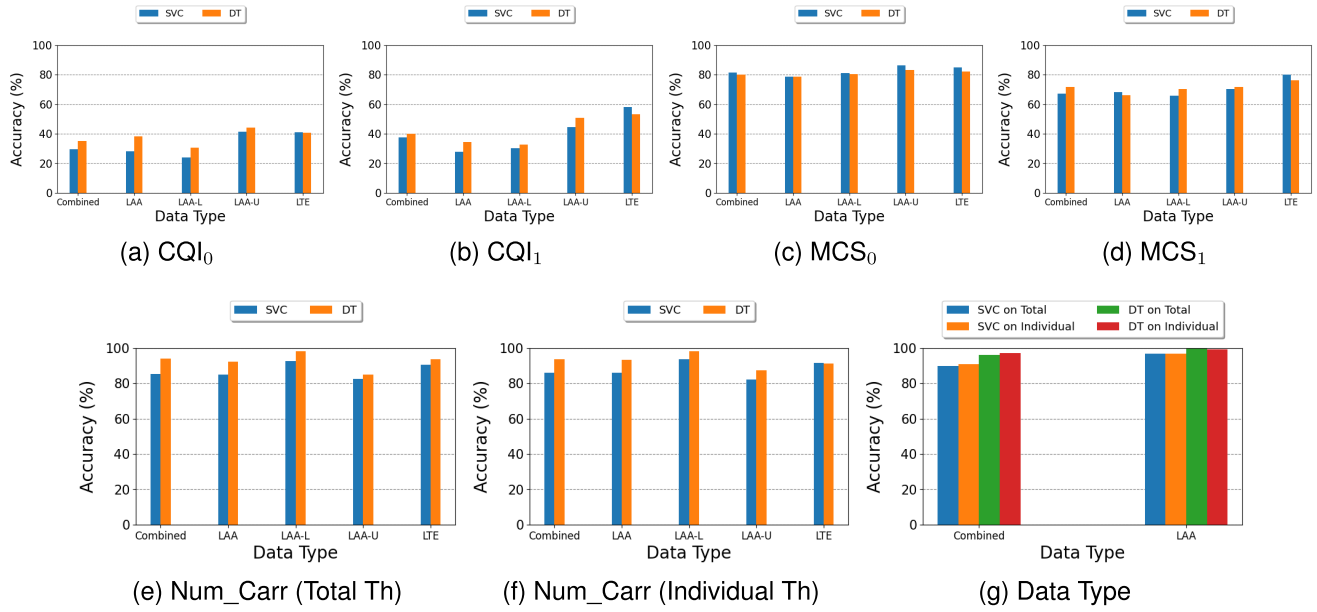



**FIGURE 9.** Prediction Accuracy of Support Vector (SVC) and Decision Tree (DT) Classifiers for various network prameters.

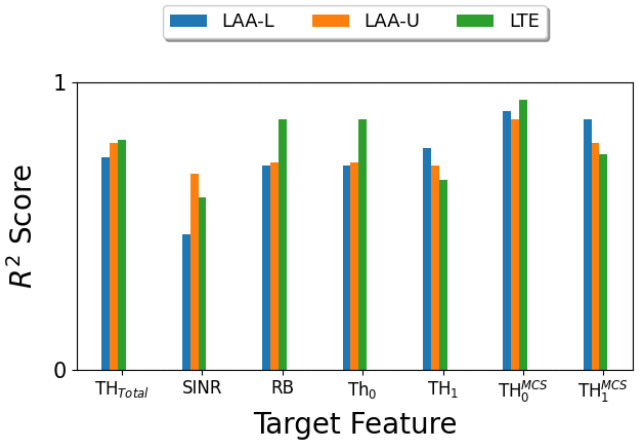



**FIGURE 10.** Network performance prediction using random forests.

results in a lower bit error rate. We do not observe much difference in the BER of LTE, LAA$_L$ and LAA$_U$. The potential reason is that LTE is the underlying technology (protocol stack) for LAA, and it is possible that the traffic from both cellular standards utilizes the same decoder.

The reduction in interference can improve the SINR signal, which can drastically improve the BER of the particular network. In general, this can be achieved by proper RF planning with different frequencies and channels assigned to nearby radios or base stations to minimize co-channel interference. In addition, intelligent RB allocation based on coordination or Physical Resource Block (PRB) muting during the silence period further minimizes the impact of co-channel interference. Both these approaches reduce the BER at the receiver by improving the SINR.

## VII. UTILIZING OPERATOR DATA FOR NETWORK PERFORMANCE ANALYSIS AND PREDICTION

A primary objective of this work is to offer a data-driven performance analysis of unlicensed networks by going beyond the classical measurement-based analysis. Machine learning algorithms make it possible to formulate the problem of estimating unlicensed band performance as a classification or a regression (prediction) problem.

### A. DATA-DRIVEN INSIGHTS FROM LAA NETWORKS

We now leverage AI/ML to analyze the LAA operator dataset and answer various questions on network performance (e.g., expected network Throughput) with high accuracy. We consider 15 features, including several Phy layer parameters such as SINR, Block Error Rate (BLER), ANT eNB T/Rx, Transmission Mode, Rank, Resource Blocks (RB), Channel Quality Indicator (CQI), and Modulation Coding Scheme (MCS). We also consider feature importance so that network bottlenecks can be easily identified. For instance, we find that it is challenging to predict network throughput by looking at SINR or RB allocation alone. Further, for different locations, small cells, or operators, the importance of SINR or RB in determining network capacity differs (sometimes operators aggregate more than one channel with different bandwidths). For example, the feature importance of SINR in predicting network capacity is far higher for Unlicensed networks than for Licensed networks. The reason is that LAA networks were characterized by high resource block allocation and lower user density, making device throughput more dependent on the signal quality it received. Another problem of interest is to predict the expected throughput or SINR with high confidence for a given set of network features. This is modelled as a regression problem. Further,

we build classification models to solve various significant problems such as identifying the carrier (LTE or LAA), determining the number of carriers aggregated, detecting the modulation coding scheme, and predicting channel quality.

The solutions utilize low-cost ML algorithms from the family of regression algorithms viz., Ordinary Least Squares Linear Regression, Ridge Regression, Kernel Ridge Regression, Polynomial Regression, K-nearest neighbours (KNN), and Random Forest (RF). For classification, we considered Decision Trees (DT) and Support Vector Machines (SVM). Further, multiple categories of the LAA dataset are considered. They are denoted by, Combined (entire dataset), LTE (Only standalone LTE network data), $\text{LAA}_L$ (LTE coexisting with LAA), $\text{LAA}_U$ (Unlicensed LAA), and LAA ($\text{LAA}_L$ and $\text{LAA}_U$ combined). Data categorization helps understand how existing licensed networks perform standalone and when they coexist with Unlicensed networks. This is particularly important for future unlicensed operations as the number of incumbents will increase.

The discussion is categorized into network performance analysis, prediction of relevant network parameters, and identification of the critical network variables influencing performance.

### B. EXPECTED NETWORK PERFORMANCE PREDICTION

Data-driven cellular network performance analysis entails evaluating the reliability of machine learning models in predicting critical performance metrics that affect the end-user quality of experience. These include the total device throughput ($\text{TH}_{Total}$), SINR at the device (SINR), and resource blocks allocated to the device (RB). Multiple ML algorithms are run for all five data categories, and the prediction performance of models is analyzed in terms of R-sq and the root mean squared error (RMSE). The results are presented in Figure 10.

We observe that for a given input of network variables (e.g., SINR, RB, BLER, CQI, number of carriers, transmission mode) the total network throughput can be predicted with moderate to high reliability ($\approx$ 0.8). $\text{TH}_{Total}$ is most accurately predicted in LTE with an R-sq of 0.8 and an error of 7.69. Surprisingly enough, predicting total throughput is the most challenging for $\text{LAA}_L$ (LTE coexisting with LAA), with the lowest R-sq and highest error. Further, throughput prediction in $\text{LAA}_U$ is almost as reliable as LTE (R-sq of 0.79), although the mean error is twice as much. This implies a higher dispersion of values from the fit line, which in turn means more fluctuation in network performance. Estimating expected SINR when the total throughput and other network variables are known is more challenging, with prediction model R-sq in the 47%–68% range. SINR Prediction models SINR are more reliable in the unlicensed band ($\text{LAA}_U$), and least reliable for LTE coexisting with LAA ($\text{LAA}_L$), with LTE in the middle (R-sq of 60%). Predicting the RB allocated to the UE can be done most accurately for LTE (R-sq of 87%) and a low mean error of 4.19. Estimating RB allocation in the unlicensed band is relatively challenging (R-sq of 72%), and there is a higher variability in RB allocation (mean error = 10.51). $\text{LAA}_L$ models are again the least reliable.

An interesting conclusion from these findings is that licensed cellular network characteristics performance in standalone deployments (licensed only) and coexistence deployments (LTE-LAA) differs substantially. The network environment and performance prediction for the licensed carriers of an LTE-LAA system are the most challenging.

To further understand this phenomenon, we analyze network performance at the granularity of aggregated traffic streams. These are represented by 'Codewords', which combine the network metrics from individual data streams. In uplink or downlink, a codeword-to-layer mapping is performed in spatial multiplexing. In LTE-Advanced and 5G-NR (as monitored on NSG), these aggregated streams are represented by 'Cwd0' and 'Cwd1' values of metrics such as throughput, modulation coding scheme (MCS), and channel quality indicator (CQI). We first compare the performance of codeword-specific models in predicting the two aggregate throughputs, viz., ($\text{TH}_0$ and $\text{TH}_1$), with the total throughput estimation models. Further, the impact of codeword-specific parameters such as MCS on the aggregate throughputs ($\text{TH}_0^{MCS}$ and $\text{TH}_1^{MCS}$) is studied.

The throughput prediction performance trends change significantly for the codeword-specific models for aggregated traffic streams. First, the throughput estimation ability depends on the codeword or stream-aggregate data and varies by up to 10%. In the LAA dataset, $\text{TH}_0$ can be predicted more reliably than $\text{TH}_1$, although with a slightly higher error. Secondly, while LTE models are the most reliable for $\text{TH}_0$, LAA and LAA $\text{TH}_1$ models are most reliable for $\text{TH}_1$. Surprisingly, for $\text{TH}_0$, LTE models perform the worst with an R-sq of 66%. This is in sharp contrast with the total throughput models, where prediction in LTE networks is the most accurate and with the lowest error. Thus, traffic-stream data analysis can yield more significant insights into network parameter performance.

Further, codeword-specific MCS values improve overall throughput prediction and reduce error. What is particularly interesting is that the inclusion of MCS makes $\text{LAA}_L$ throughput estimation remarkably accurate, with 90% and 87% R-sq for $\text{TH}_0^{MCS}$ and $\text{TH}_1^{MCS}$ prediction, respectively. Trends for $\text{LAA}_U$ remain unchanged when compared to total throughput models. This further underscores our inference earlier that LTE in coexistence with unlicensed cellular networks is characteristically different from standalone LTE. Analysis and performance prediction in the coexistence environment is highly contextual for LTE. Further, context-specific variables such as MCS only slightly improve $\text{LAA}_U$ or unlicensed cellular model performance. The impact of LTE on unlicensed cellular operation is not as pronounced as the impact of $\text{LAA}_U$ on the licensed operation. This finding is crucial for 5G, 6G, and other cellular standards that coexist with unlicensed band networks like 5G-NRU.

### C. NETWORK PARAMETER PREDICTION

Now, we consider the challenge of accurately determining four network parameters, viz., the number of carriers aggregated, channel quality indicator, modulation coding scheme, and network environment or network type. These prediction problems were considered for both total throughput and stream-aggregated throughput data for all five data categories as before. The results are presented in Figures 9(a) and 9(b) for CQI, Figures 9(c) and 9(d) for MCS, Figures 9(e) and 9(f) for Number of Carriers, and Figure 9(g) for cellular network type.

Predicting channel quality through the values of the CQI metric is challenging (Figures 9(a) and 9(b)). The prediction accuracy varies from 32% for $LAA_L$ to 58% for LTE. $LAA_U$ lies in the middle with 44% and 50% accuracy for the two aggregated-stream models. A plausible explanation is that channel quality in LAA depends on the neighbouring LAA cells and other Wi-Fi APs on the same channel. From our observations, when a Wi-Fi AP detects LAA on the same channel, it usually moves to a better channel with less contention. We also notice that LAA UE stays on the same channel as it does not currently support dynamic channel selection. A latent consequence is a more accurate channel quality or CQI prediction in LAA networks. However, in $LAA_L$ networks, coexistence with unlicensed cellular networks makes CQI prediction challenging. LTE networks operating in the licensed spectrum operation try to reuse the same channel as much as possible to effectively reuse the spectrum. So, improper RF planning on the cellular network may pose difficulties in determining CQI by feeding the network performance metrics as input to the prediction model.

Another network parameter of interest is the modulation coding scheme (Figures 9(c) and 9(d)). The transmission from an LAA base station happens at 23 dBm, so the operator typically deploys small cells. Consequently, the users connected to the LAA small cell receive strong signal strength measured in RSRP and SINR. This translates to a high MCS, and in turn, a high data rate in LAA. It also makes it possible to accurately predict MCS in the unlicensed band for a given feature, including codeword-specific throughput and other network metrics. This explains the high prediction accuracy of up to 86%, for $LAA_U$. Further, LTE or licensed operation requires high transmit power at the Macro base station to provide greater coverage with robust signal quality. However, potential scenarios such as a user connected to the Macro base station at the edge of the cell may lead to low RSRP and low SINR, which directly translates to a lower MCS, and in turn, low throughput. We also find a similar variation in prediction accuracy in LTE (41% to 58%) for the two codeword-specific traffic streams. The case of $LAA_L$ continues to be peculiar, with prediction accuracy hovering around 30%. Clearly, it seems extremely challenging to estimate channel quality by looking at the network data in LTE when it coexists with LAA.

Determining the number of carriers aggregated at the UE can be done with a near-perfect accuracy of 98% (Figures 9(e) and 9(f)). This is true for both a smaller feature set (6 features) including $TH_{Total}$ or an expanded feature set (15 features) including aggregated parameters. The interesting aspect is that the prediction is the most reliable in $LAA_L$, followed by LTE, and least accurate in $LAA_U$. LTE is in the licensed spectrum, where only authorized cellular providers are operational. Hence, the carrier or channel is 'clean' with no co-channel interference. On the contrary, $LAA_U$ networks exist in the unlicensed spectrum, where existing incumbents employ the Listen Before Talk or the Wi-Fi CSMA protocol to use the spectrum freely and without restrictions. Thus, external factors (features) other than the core network metrics influence the accuracy of models analyzing network data.

Finally, we consider the problem of predicting the network environment or type at the device (Figure 9(g)). This is particularly important for use cases such as device-initiated cell selection or handover. Only two data categories are considered in this solution, viz., Combined data and LAA. The models for Combined data are able to distinguish between the three network types i.e., LTE, $LAA_L$, and $LAA_U$, with 97% accuracy. In LAA data, the accuracy of predicting whether the device is transmitting on a licensed carrier ($LAA_L$) or unlicensed carrier ($LAA_U$) is 99.3%. It is encouraging that a high level of accuracy can be achieved with low-cost ML algorithms which makes delegating the cell selection and handover decision to the mobile devices feasible.

### D. IDENTIFYING IMPORTANT NETWORK PARAMETERS

Measurement-based cellular network studies provide empirical trends on many network variables. A limitation of this measurement-focused approach is that it offers little insight into which variables have the most significant impact on metrics associated with end-user QOE, such as throughput or signal quality. From the cellular operator's perspective, it is pertinent to identify important network variables so as to remove performance bottlenecks and manage the network better.

We employ *permutation feature importance* to identify the variables that contribute the most to prediction model performance for all metrics considered in Section VII-B. Permutation importance [28] is a technique used to measure the importance of features in a machine learning model. It helps understand which features have the most significant impact on the model's performance and predictions. Proposed as a measure of variable importance in random forests, permutation importance involves randomly shuffling the values of a single feature in the test set and observing the effect on the model's performance. The percentage change in the performance metric is identified as the permutation importance of that feature. The three most important network variables that affect network performance are listed in Table 2.

**TABLE 2. Most important features that affect network performance.**

| Performance Metric | Combined Data | LTE Data | LAA Data | $LAA_L$ Data | $LAA_U$ Data |
|---|---|---|---|---|---|
| $TH_{Total}$ | RB, SINR, BLER | RV, Num_Carrier, ANT | RB, SINR, ANT | RB, SINR, RANK | SINR, BLER, RANK |
| SINR | ANT, $TH_{Total}$, Data Type | Num_Carrier,$TH_{Total}$, RB | $TH_{Total}$, ANT, RB | ANT, $TH_{Total}$, RB | $TH_{Total}$, RB, BLER |
| RB | $TH_{Total}$, Data Type, Num_Carrier | $TH_{Total}$, ANT, SINR | $TH_{Total}$, Data Type, SINR | $TH_{Total}$, Num_Carrier, SINR | Num_Carrier, BLER, SINR |
| $TH_0$ | RB, SINR, BLER | RB, Num_Carrier, ANT | RB, SINR, BLER | RB, SINR, BLER | SINR, BLER, RB |
| $TH_1$ | RB, SINR, ANT | RB, RANK, CQI | RB, ANT, SINR | RB, SINR, CQI | SINR, BLER, RB |
| $TH_0^{MCS}$ | RB, MCS, SINR | RB, MCS, SINR | RB, MCS, BLER | RB, MCS, SINR | MCS, RB, BLER ≈ SINR |
| $TH_1^{MCS}$ | RB, MCS, RANK | RB, RANK, MCS | MCS, RB, SINR | RB, MCS, SINR | MCS, RB, SINR |

[1]For each data type, the three variables with the highest feature importance are presented for Random Forest models.

An interesting observation is that in models with data from licensed cellular networks, total throughput is primarily controlled by RB. However, in $LAA_U$, SINR and BLER are more important than RB in predicting network throughput. So, different network variables need to be fine-tuned for optimal performance in licensed and unlicensed bands for similar performance metrics. A similar trend can be observed for RB as well, where $TH_{Total}$ is the most important predictor in all data types, except for $LAA_U$, where SINR contributes most to model performance, with $TH_{Total}$ as a close second. Further, it can also be noticed that the data type itself is an important feature in aggregated data types, viz., Combined data and LAA data. This further demonstrates that network environment or "context" differs in licensed and unlicensed bands, and performance prediction models must be trained on network-specific data. These trends persist for the codeword-specific models too. RB is the most important predictor of $TH_0$, $TH_1$, $TH_0^{MCS}$, and $TH_1^{MCS}$, for data-types comprising licensed data. Whereas for purely unlicensed data, SINR for $TH_0$ and $TH_1$, and MCS for $TH_0^{MCS}$ and $TH_1^{MCS}$, are the most important network variables to predict system performance. Further, adding new variables to the dataset can often help arrive at the specific parameter with the maximum impact on network performance. For example, MCS is determined by the radio link quality, which includes signal strength, BLER, CQI and more. Thus, the unlicensed system performance is best enhanced by improving MCS rather than individual variables that have a high correlation with it. In the case of other data types, RB is the predominant network variable that governs all types of throughput. Finally, the impact of incumbents coexisting and operating in the unlicensed band is clearly visible on $LAA_U$ performance. Variables associated with the link and medium quality, e.g., SINR, BLER, MCS, and TRANS, significantly impact unlicensed network performance more than licensed networks.

**Algorithm 1** Counterfactual Framework for Cellular Network Performance Analysis

**Require:**
1. Cellular Network data $X$
2. Low-cost machine learning model $M$
3. Predictions for each input instance $Y_{\text{original}}$
4. Desired prediction $D$

**Ensure:**
Initialize an empty set $C_{\text{valid}}$
**for** each input instance $x_i$ in $X$ **do**
  Generate a counterfactual instance $CF_i$ using the counterfactual machine learning framework
  **if** $Y_{CF_i} \in R$ and $Y_{CF_i} \neq Y_{\text{original}_i}$ where $R$ is the desired range **then**,
    Add $CF_i$ to $C_{\text{valid}}$
  **end if**
**end for**
Initialize an empty set$C_{\text{feasible}}$
**for** each valid counterfactual in $C_{\text{valid}}$ **do**
  **if** FeasibilityCheck($x_i$, $CF_i$, $M$) = True **then**
    Add $CF_i$ to $C_{\text{feasible}}$
  **end if**
**end for**
**for** each feasible counterfactual in$C_{\text{feasible}}$ **do**
  Initialize two similarity scores for continuous and categorical features:
  $similarity_{\text{continuous}} \leftarrow$ CalculateCosineSimilarity($x_i$, $CF_i$)
  $similarity_{\text{categorical}} \leftarrow$ CalculateCategoricalSimilarity($x_i$, $CF_i$)
  Calculate the final similarity score:
  $similarity \leftarrow (w_{\text{continuous}} \cdot similarity_{\text{continuous}}) + (w_{\text{categorical}} \cdot similarity_{\text{categorical}})$
**end for**
Use joint optimization to validate the results of the counterfactual analysis

## VIII. COUNTERFACTUAL FRAMEWORK FOR NETWORK PERFORMANCE ENHANCEMENT

This work leverages machine learning to demonstrate how operators can make accurate predictions and decisions for optimal cellular network performance. However, to present a more comprehensive understanding of how the features in network data can affect the operator decision-making process, causal analysis of network performance prediction models is needed. Counterfactual explanations are a great tool for causal inference. Instead of focusing on existing outcomes, counterfactual analysis shines light on desirable outcomes by posing the question, "What would happen to the outcome if a specific input were altered?". This section describes how we leverage counterfactual explanations for the classification problems defined on our dataset. We apply interpretability

mechanisms to comprehend the black-box models using counterfactuals [12], [13].

### A. PROPOSED FRAMEWORK

Our counterfactual analysis adopts a binary classification framework, where the positive class threshold aligns with the mean value of the relevant performance metric, which is SINR and throughput, for our analysis. The dataset is divided into three segments based on the different network environments to study the effect of network parameters on performance metrics through a data-driven approach. In the context of each subset, the computational procedure involves the calculation of the mean scalar for SINR. This quantification is then utilized to create a novel binary target attribute, bifurcating samples into those exceeding or falling short of the statistically computed mean. Concurrently, the summation of throughput for CWD0 and CWD1 is done to obtain a comprehensive aggregate throughput magnitude. We ascertain the mean of this composite throughput value and introduce an additional binary target feature discerning whether the cumulative throughput is above or below the mean. Therefore, by using the empirically estimated average as the threshold for the binary classifier, we simplify the prediction problem and set the grounds for further analysis using counterfactuals. Using a random forest classifier, we apply this methodology to appraise signal strength and throughput.

Specific constraints are incorporated into the model's architecture to engineer a counterfactual generator explainer model based on the chosen classifier, thus defining an acceptable spectrum within which the feature perturbations transpire. Concurrently, attributes limited by their immutability are identified to ensure their preservation in the counterfactual generation process. We leverage DiCE [11], Diverse Counterfactual Explanations, that incorporates diversity and proximity into a unified loss function to synthesize counterfactual instances. The objective of this analysis can be illustrated through this hypothetical instance- consider a specific sample demonstrating inadequate throughput, i.e., its throughput falls below the mean of the pertinent network environment. A counterfactual of this would suggest that to have more throughput, the error rate can be reduced, and the signal strength can be improved. This analysis determines whether the classifier model has learned the interrelations among the various network parameters and performance metrics.

A test set of samples is created for subsequent counterfactual synthesis. The trained explainer model is then made to generate counterfactual explanations for each individual sample encompassed within this test set. The objective of this analysis can be illustrated through this example - consider a specific sample demonstrating low throughput, i.e., its throughput falls below the mean of the given network environment. A counterfactual of this would suggest that to have more throughput, the error rate can be reduced, and the signal strength can be improved. This analysis determines whether the classifier model has learned the interrelations among the various network parameters and performance metrics. Once we have generated a set of counterfactual instances for our test set, we evaluate them using the metrics of validity, feasibility and similarity. These are designed specifically for this research to gauge the effectiveness of our analysis and the counterfactual framework. The validity of a counterfactual instance aims to check whether the permutations in the features are sufficient to alter the decision of the prediction model. Feasibility is a parameter that is rooted in the domain of the research problem and highlights the model's understanding of the interrelations among the features and the target variables. Similarity aims to measure the distance of the original input from the synthetic counterfactual instance, which uncovers insights about the decision boundaries of the classification. These metrics are discussed in detail in Section VIII-C. The high-level process flow is depicted in Figure 11.

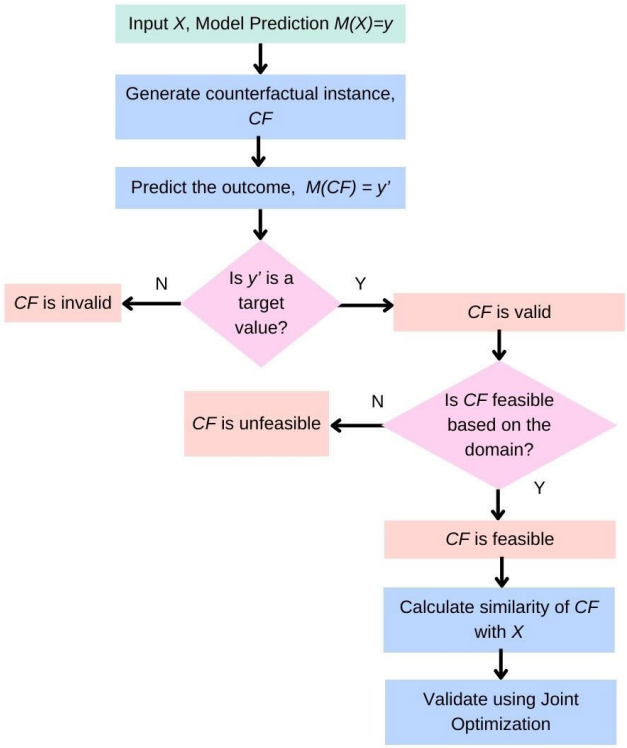



**FIGURE 11. High-level Schema of the proposed counterfactual framework.**

### B. APPLICATION IN CELLULAR NETWORKS

Algorithm 1 presents a Counterfactual Framework designed to be an advanced explainable machine-learning method for cellular network data analysis and performance enhancement. It requires four primary elements: a cellular network dataset $X$, a cost-effective machine learning model $M$, initial predictions $Y_{\text{original}}$ for each data point, and a target prediction $D$. The procedure initiates by creating an empty set $C_{\text{valid}}$ to accumulate valid counterfactual instances. For every instance

**TABLE 3. Validity and feasible counterfactuals.**

| Data | Metric | Validity | Feasibility | Continuous Similarity | Categorical Similarity |
|---|---|---|---|---|---|
| LAA (U) | SINR | 98.49% | 76.33% | 97.35% | 75.12% |
| | TH | 98.49% | 71.62% | 99.17% | 64.16% |
| LAA (L) | SINR | 89.32% | 67.58% | 87.64% | 80.55% |
| | TH | 96.98% | 78.70% | 89.42% | 75.92% |
| LTE | SINR | 99.34% | 62.91% | 94.96% | 65.35% |
| | TH | 92.10% | 87.14% | 87.71% | 79.60% |

$x_i$ in $X$, a corresponding counterfactual instance $CF_i$ is generated using a dedicated counterfactual machine learning framework. These instances are evaluated to ascertain that they conform to a specified range $R$ and differ from the original predictions $Y_{\text{original}_i}$. Those meeting these criteria are stored in $C_{\text{valid}}$.

In the following stage, the algorithm filters these valid counterfactuals through a feasibility assessment, ensuring their practicability in the context of the model $M$, and places feasible ones in a new set $C_{\text{feasible}}$. Two types of similarity scores are computed for each feasible counterfactual: one for continuous features using cosine similarity and another for categorical features using a specific method. These scores are integrated into a final similarity score, achieved by a weighted sum of the two, thereby balancing the influence of different feature types. The process culminates with a joint optimization phase, validating the outcomes of the counterfactual analysis. This ensures that the generated counterfactuals are valid and feasible and closely resemble the original data points, enhancing their practical value and interpretability in cellular network data analysis.

### C. EVALUATION

The generated counterfactuals are now subjected to a comprehensive assessment of their practical utility, focusing on three primary aspects: validity, feasibility, and proximity.

The validity of a counterfactual is ascertained through its impact on the classification outcome, with the pivotal question centring on its potential to transform the prediction to the desired class or range of values. The quantification of validity is interpreted by the computation of the proportion of valid counterfactuals in the generated set. Studying the validity provides deeper insights into the interpretability of the machine learning model. For each network environment, the fraction of valid counterfactuals for each performance metric is illustrated in Table 3.

The features in the dataset do not function as independent entities but exhibit intricate interconnectedness. Given the intrinsic dependencies among features, altering one feature could have cascading effects on others. Consequently, any counterfactual generated should not only achieve the intended prediction alteration but also ensure that these intricate interdependencies are preserved. This adds a layer of complexity to the evaluation process with feasibility as another desirable property of counterfactuals. They must not only satisfy individual conditions for prediction alteration but also align with the real-world interplay of features. This imposes constraints on their modifications and simply perturbing each feature individually may lead to unrealistic examples, suggesting conflicting changes, such as simultaneously increasing signal strength and error rate to improve the throughput. Consequently, counterfactual analysis necessitates comprehensive checks on multiple feature alterations to ensure that generated counterfactuals are not only valid in terms of prediction shifts but also adhere to the nuanced relationships inherent in the dataset. The feasibility of a generator can be conceptualized as the proportion of counterfactual instances within the set that adheres to the inherent relationships not only among the features themselves but also between the features and the target variable. To evaluate the feasibility, we utilize a filtering methodology for generated counterfactual examples based on causal constraints and present the results in Table 3. An in-depth analysis of counterfactuals that meet the validity criteria is conducted, delving into the extent to which the implicated modifications uphold the intrinsic interdependencies embedded within the dataset. The altered feature set is scrutinized to ascertain its fidelity to the underlying network of realistic and attainable configurations. Subsequently, we compute the proportion of counterfactual instances that exhibit the characteristic of feasibility in the set of valid counterfactuals.

$$\begin{aligned} \forall i : (Y_{\text{cf}} > Y_{\text{orig}}) &\implies (\rho(X_i, T) > 0 \implies \Delta X_i > 0) \\ &\wedge (\rho(X_i, T) < 0 \implies \Delta X_i < 0) \\ \forall i : (Y_{\text{cf}} < Y_{\text{orig}}) &\implies (\rho(X_i, T) > 0 \implies \Delta X_i < 0) \\ &\wedge (\rho(X_i, T) < 0 \implies \Delta X_i > 0) \end{aligned}$$

Across all three datasets, the observed feasibility rate demonstrates a significant level of efficacy. This underscores the classifiers' capability to discern the intricate interplays between the target attributes and the determinant features. The classifier adeptly captures the complex interdependencies among diverse performance metrics and network parameters. Consequently, perturbing the data yields outcomes that align with the intended class assignment, attesting to the random forest classifier's comprehensive grasp of the underlying relationships. This further highlights the

**TABLE 4. Most frequently changed features.**

| Performance Metric | LTE | LAA_U | LAA_L |
|---|---|---|---|
| SINR | Num_carriers, CQI, TH | CQI, TH, BLER, | ANT, CQI, TH |
| Throughput | RB, ANT, SINR | RB, BLER, CQI | RB, CQI, SINR |

application of counterfactual analysis as an explainability mechanism for complex machine learning models with otherwise difficult interpretability and comprehension.

From a logical standpoint, the utility of counterfactual examples is significantly amplified when they closely resemble the original input. This is because drastic changes to alter a prediction to attain a desirable outcome suggest that the model may have captured a rudimentary data abstraction, lacking finer-grained comprehension of intricate feature interactions. Simultaneously, the consideration of drastic modifications raises the spectre of implementational impracticability. The very nature of significant changes to input features demands significant alterations to real-world circumstances, which may not be operationally tenable or economically feasible. The cosine similarity index across the continuous numerical attributes is computed to quantify the proximity between counterfactual instances and their corresponding original samples. The aggregate proximity measure is obtained through the arithmetic mean across all instances, thereby encapsulating the comprehensive feature space.

While categorical features are amendable to level encoding for numerical treatment, such a procedure might not inherently capture the nuanced disparity in altering individual categorical attributes or the genuine ''distance'' between distinct values. To tackle this intricacy, we introduce the concept of categorical similarity, an evaluative measure of how closely a counterfactual's categorical attributes align with their respective original inputs. This metric bestows a similarity score of 0 upon any deviation in the counterfactual's categorical feature value from the original input, whereas a score of 1 signifies consistency. In a broader context, when considering an ensemble of counterfactual instances, the notion of proximity can be quantified as the mean similarity value spanning the entire set.

$$\text{Categorical Feature Similarity} = \frac{1}{d}\sum_{i=1}^{d} I(c^p = x^p) \quad (1)$$

### D. COMPARISON WITH FEATURE IMPORTANCE

Section VII discusses the use of permutation feature importance to uncover the most valuable features during the decision-making process. An intuitive assumption might link feature importance, observed in prior stages, to counterfactual analysis, where more important features are more prone to alterations, aligning with their frequency of change. Notably, permutation importance, akin to counterfactuals, involves feature variation or shuffling, reinforcing this conjecture. The most frequently changed features are discussed in Table 4.

We can compare these findings to the results of the feature importance calculation in Table 4 and notice that among the top three features, only two overlap for each case. This divergence is intriguing as it presents valuable insights into the complexity of the model's decision-making process. It also highlights the unique nature of counterfactuals. While permutation importance quantifies the features' impact on the model's predictive performance, counterfactual analysis considers the collective perturbation of multiple features to alter predictions. This discrepancy highlights that while certain features may be individually pivotal, a more nuanced interplay of features might be required to effectively influence the model's outcomes through counterfactual manipulation.

In the context of SINR analysis, two key parameters that often change across all three data types are throughput and Channel Quality Indicator. Specifically, when considering the LAA_L data type, variations are observed in the Antenna Configuration (ANT) in 31% of cases, throughput in 26%, and CQI in 29%. In LAA_U, throughput experiences alterations in 31% of the test cases and BLER in 15% of cases, while CQI is subject to variation in 49% of instances. Meanwhile, the LTE data type is modified in 53% of cases, CQI in 19%, and throughput in 11% of the generated counterfactuals. Examining the total throughput in LAA_L, it becomes evident that RB is changed 61% of the time, CQI at 24%, and SINR at 11%. In LAA_U, RB is changed in 15% cases, BLER in 13% cases and CQI in 17% cases. Similarly, for LTE, RB is varied in 80% cases, ANT in 22%, and SINR in 11%.

## IX. VALIDATING COUNTERFACTUAL POLICIES THROUGH JOINT OPTIMIZATION MODEL

The final step in the proposed counterfactual framework is to validate the counterfactual outcomes through well-established techniques such as network optimization. We devise a joint optimization model, where the goal is to maximize the radio resource allocation or scheduling algorithm in the 4G LTE or LAA LTE-U or 5G NR Time Division Duplexing (TDD) frames with the constraints of the device signal, application deadline, and user allocation fairness. The proposed model is one form of replicating the classical proportional fair scheduling algorithm [29], which has the QoS and priority as a constraint on the radio resource allocation. The notation and definition of the problem formulation are illustrated in Table 5.

**Objective:** The joint optimization model presented aims to maximize radio resource allocation in a network that utilizes both Long-Term Evolution (LTE) and New Radio (NR) technologies, primarily focusing on Time Division Duplexing (TDD) frames. The main objective is to maximize the sum of the weights of the radio resources allocated in each LTE or NR frame, represented below, ensuring that devices receive the necessary radio resources to fulfil their service

**TABLE 5.** List of notations used in the problem formulations.

| Notation | Definition |
|---|---|
| $\hat{O}_o$ | Maximum allowable delay for a device $o$. |
| $O$ | Number of devices |
| $T$ | Number of time slot. |
| $x_{dt}$ = | 1 if device $o$ gets access at timeslot $t$ else 0. |
| $O_{dt}$ | Delay of device $o$ at timeslot $t$. |
| $R_{dt}$ | Reward for device $o$ at timeslot $t$. |
| $\bar{R}_o$ | Penality for device $o$ (calculated at the end of each frame). |
| $P_o$ = | 1 if request of device $o$ is not served within $\hat{O}_o$. |
| $\alpha_{dt}$ = | 1 if device $o$ receives packet on or before $\hat{O}_d$ at timeslot $t$ else 0. |
| $\beta_{dt}$ = | if device $o$ receives packet after $\hat{O}_d$ at timeslot $t$ else 0. |
| $\lambda_j$ | SINR threshold achieved by a device when connected to the AP or BS or Node |
| $N_{0t}$ | System Noise |
| $G_{wj}$ | Channel gain from the near-by LTE or LAA or LTE-U BS |
| $P_{max}$ | The maximum transmission power at the BS |
| $z$ | Set of spectrum such as licensed and un-licensed in the system |

requirements.

$$Max(\sum_{o=1}^{|O|}\sum_{t=1}^{T} R_{dt} + \sum_{o=1}^{|O|} \bar{R}_o)$$

**Constraints:** It is ensured that each device receives access to the network exactly once within a given frame, preventing any device from being overlooked or being given multiple access within the same frame. There are also penalties for delays and rewards for timely packet reception. The model incorporates the Signal-to-Noise Ratio (SNR) as a significant aspect of resource allocation, ensuring that each device's SNR is maintained above a predefined threshold.

The allocation of resources is based on the devices' observed transmission outcomes in previous frames, determining if a packet should be sent or withheld (1 for sending, 0 for withholding). A packet can only be sent in the current frame if it was not sent in the previous frame, ensuring efficient use of the radio resources and avoiding unnecessary retransmissions.

The equation below ensures that exactly one LTE or NR device accesses the channel at each time frame. The constraint below can be extended to dual connectivity between two technologies such as LTE and NR, or coordinated multi-point (COMP) technologies between the same 4G or 5G NR base stations.

$$\sum_{o=1}^{|O|} x_{dt} \leq 1 \qquad \forall t \in [T] \tag{2}$$

Based on the SINR signal level at the device, traffic requirement, and fairness, the MAC scheduling algorithm determines the modulation coding scheme and the number of radio resources that need to be allocated in each frame. In this work, fairness is considered in terms of deadline, reward, and penalty.

*$SINR_{ij}$ Threshold:* The L.H.S. of Equation (3) is the $SINR_{ij}$ received at the device $j$ due to transmissions from radio or base station or AP (LTE/LAA/LTE-U) $i$, and $N_0t$ represents the system noise. To ensure a reliable connection, each device link's $SINR_{ij}$ is maintained above a predefined threshold $\lambda_j$, which may vary across mobile nodes.

$$\frac{Inf \times (1 - Q_{ij}^z) + G_{ij} p_i^z P_{max}^w}{N_{ot} + \sum_{w \in W_k} G_{wj} P_{max}^i + \sum_{i' \in I \setminus i} G_{i'_j} p_{i'}^z P_{max}^w} \geq \lambda_j$$
$$\forall i \in I, j \in J, z \in Z \tag{3}$$

Here, $W_k$ is the set of all nodes using the spectrum $z$ in a given LTE or NR frame duration. Similarly, $G_{wj}$ is the channel gain from the other near-by LTE or NR node $w$ to $j$ (operating on the same spectrum chunk), and $G_{ij}$ is the channel gain from $i$ to $j$.

The use of $Inf \times (1 - Q_{ij}^z)$ ensures that if $Q_{ij}^z = 0$, then $Inf \times (1 - Q_{ij}^z)$ amounts to a very large value, which allows for the expression to be conveniently ignored. Through the virtual infinite value $Inf$, Equation (3) ensures that all relay nodes provide a minimum $SINR_{th}$ to a particular mobile node. The proposed joint optimization model will be impractical without the SINR consideration through $Inf$. The Equation (3) can be rewritten as follows,

$$SINR_{ij} \leq \frac{Inf \times (1 - Q_{ij}^z) + G_{ij} p_i^z P_{max}^w}{N_{ot} + \sum_{w \in W_k} G_{wj} P_{max}^i + \sum_{i' \in I \setminus i} G_{i'_j} p_{i'}^z P_{max}^w}$$
$$\forall i \in I, \forall j \in J, \forall z \in Z \tag{4}$$

The below three constraints help in classifying whether a packet received by device $o$ at time $t$ ($x_{dt} = 1$) is served before $\hat{O}_o$ then $\alpha_{dt} = 1$. Based on the LTE or LAA or LTE-U and NR QCI priority, the maximum allowable delay $\hat{O}_o$ is determined. Suppose if the packet is served after $\hat{O}_o$ then $\beta_{dt} = 1$.

$$\alpha_{dt} + \beta_{dt} = x_{dt} \qquad \forall o \in [O], t \in [T] \tag{5}$$

$$O_{o,t-1} \leq \hat{O}_o - 1 + \bar{O}(1 - \alpha_{dt}) \qquad \forall o, t \tag{6}$$

$$\bar{O}(1 - \beta_{dt}) + O_{o,t-1} \geq \hat{O}_o \qquad \forall o, t \tag{7}$$

The reward $R_{dt}$ for device o at time t is calculated only if the packer is served within $\hat{O}_o$. Otherwise, the penalty $-\hat{R}$ is observed in each frame.

$$R_{dt} = \alpha_{dt}(\frac{1 + O_{o,t-1}}{\hat{O}_o}) + \beta_{dt}(-\hat{R}) \qquad \forall o \in [O], t \in [T] \tag{8}$$

$$O_{o,t} = (O_{o,t-1} + 1)(1 - \alpha_{dt}) \qquad \forall o \in [O], t \in [T] \tag{9}$$

Calculate the delay for device $d$ at time $t$:

$$O_{o,t} = \begin{cases} O_{o,t-1} + 1 & if \quad x_{dt} = 0 \\ 0 & otherwise \end{cases}$$

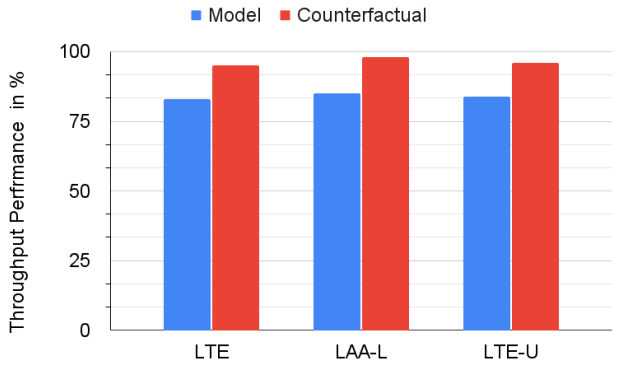



**FIGURE 12.** Model Vs Counterfactual.

To calculate the reward of device $d$ at time $t$,

$$R_{o,t} = \begin{cases} 0 & if \quad \alpha_{dt} = \beta_{dt} = 0 \\ \dfrac{1 + O_{o,t-1}}{\hat{O}_o} & if \quad \alpha_{dt} = 1 \\ -\hat{R} & otherwise \end{cases}$$

where $\hat{R}$ is a large value. This large value ensures that device $d$ is not scheduled in a frame if the delay exceeds $\hat{O}_o$. **Note:** $\hat{R}$ can be replaced with other functions if we wish to schedule a device even after its delay exceeds the allowable delay.

$$M * P_o \geq \frac{O_{o,T}}{\hat{O}_o - 1} - 1 \qquad o \in [O] \tag{10}$$

The above equation ensure that $P_o = 1$ if $O_{o,T} \geq \hat{O}_o$ and $M$ is a large value.

$$\bar{R}_d = -\hat{O}_o * P_o \qquad o \in [O] \tag{11}$$

$\rightarrow$ Penalty for device $d$ if $P_o = 1$

$x \in \{0,1\}^{O*T}, \alpha \in \{0,1\}^{O*T}, \beta \in \{0,1\}^{O*T}, P \in \{0,1\}^{O}$
$D \in \mathbb{R}^{O*T}, R \in \mathbb{R}^{O^+*T}, \bar{R} \in \mathbb{R}^{O^+}$

### A. LINEARIZING THE ABOVE MODEL

Bilinear product $\alpha_{dt}\beta_{dt-1}$ makes the above model non-linear. Hence, we linearize the model as follows,

$$\begin{aligned} \bar{O}(1 - \alpha_{dt}) + y_{dt} &\geq O_{o,t-1} \\ y_{dt} &\leq O * \alpha_{dt} \\ y_{dt} &\leq O_{o,t-1} \end{aligned}$$

These three constraints together ensure that $y_{dt} = \alpha_{dt} * O_{o,t-1}$. Hence, the bi-linear term $\alpha_{dt} * O_{o,t-1}$ can be replaced with $y_{dt}$ subject to adding the above set of constraints. The above linear optimization model can be solved using commercial solvers such as CPLEX and GUROBI.

### B. COMPARING OPTIMAL AND COUNTERFACTUAL MODELS

We validate counterfactual outcomes through optimal values generated by the proposed optimization model. Fig. 12 shows the throughput performance comparison between LTE, $\text{LAA}_L$ and LTE-U. Two main observations can be made. First, counterfactual outcomes are comparable to the optimal values and show similar trends for all three network types. Thus, theoretical network models support the idea that the network configuration change proposed by counterfactual models will lead to enhanced network QoS. Second, the counterfactual outcomes are higher than the projected values from the optimization models. This is primarily because counterfactual models are data-driven and allow greater flexibility in network configuration change (e.g., in SINR or MCS), to achieve the maximum potential performance. In contrast, the theoretical constraint-driven optimization model doesn't allow much flexibility in varying the SINR, MCS, and QCI. The model optimizes resource allocation based on the MAC scheduling algorithm by considering the QoS or QCI, fairness, and application deadline. Although some features, such as the resource block allocation, are more or less comparable in the optimization model and counterfactual model, the difference in MCS and SINR, unlocks a greater potential network performance in the counterfactual model compared to the optimization model.

The proposed optimization model will help small cells to allocate radio resources effectively by considering SINR, QOS and application requirements. The model helps to formulate policies on SINR, which is a key part of BER and spectral efficiency. A higher SINR will ensure optimal radio resource usage by the scheduling algorithm.

## X. CONCLUSION AND WAY FORWARD

This paper presents a dataset from the LAA networks of three major cellular operators in Chicago consisting of 15 features and 9676 samples. Additionally, we create a novel framework to analyze and comprehend the complex interplay of features in a network environment. This work sought to facilitate greater access to unlicensed network data through data that was extracted through an innovative low-cost and scalable solution. A subset of the dataset has led to insightful findings on LAA network operation and led to data-driven solutions. This study reports that predicting network throughput in unlicensed bands, especially in Licensed Assisted Access (LAA) networks, is more complex compared to licensed networks. Variables like Signal-to-Noise Ratio (SINR) and Resource Blocks (RB) allocation have varying importance in different network environments. Additionally, the importance of network variables like SINR, BLER, and RB varies significantly based on the network type (licensed or unlicensed), emphasizing the need for context-specific modeling in performance prediction. Models can predict network parameters like the number of carriers aggregated and network environment type with high accuracy, demonstrating the potential of machine learning in network management.

Building on the successes of the ML models, this research introduces counterfactual explanations as a pivotal tool for causal analysis in cellular network performance, enhancing understanding of feature impacts on network operator decisions. For the same, a novel Counterfactual Framework is

proposed, tailored for cellular network data analysis, requiring a dataset, a machine learning model, initial predictions, and a target prediction. This algorithm focuses on generating valid and feasible counterfactual instances for better decision-making. Additionally, the paper compares feature importance (derived through permutation feature importance) with the frequency of feature alterations in counterfactual analysis, revealing insights into the model's decision-making process. The study highlights how counterfactual analysis can serve as an effective explainability mechanism for complex machine learning models, especially in scenarios where traditional interpretability approaches might fall short. This adds a new dimension to understanding and improving network performance prediction models.

The data set in this study is gathered from macro LTE base stations using 4 × 4 and LTE-LAA small cell 2 × 2 MIMO antenna configurations. Carrier aggregation was present in the licensed band, while the unlicensed band constitutes a combination of frequencies in the mid-band spectrum, i.e., < 6 GHz spectrum. Based on our observation, the deployment of radios by all three major operators in the US (at the time of measurements/experiments) does not signify a massive MIMO architecture. However, the proposed data-driven methodology and counterfactual analysis can be extended to massive MIMO architecture with ease.

We are currently analyzing the full released dataset through advanced AI/ML techniques such as multi-task learning to find solutions for unlicensed coexistence and spectrum sharing in the 6 GHz band and beyond. We also intend to periodically release 5G NR and NR-U datasets in the future to facilitate democratic universal access to data from state-of-the-art cellular networks.

## ACKNOWLEDGMENT

The authors thank Karan Bhukar, IBM Research, for his inputs and insights.

## REFERENCES

[1] G. Naik and J. J. Park, "Coexistence of Wi-Fi 6E and 5G NR-U: Can we do better in the 6 GHz bands?" in *Proc. IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[2] E. Reshef and C. Cordeiro, "Future directions for Wi-Fi 8 and beyond," *IEEE Commun. Mag.*, vol. 60, no. 10, pp. 50–55, Oct. 2022.

[3] X. Lu, M. Lema, T. Mahmoodi, and M. Dohler, "Downlink data rate analysis of 5G-U (5G on unlicensed band): Coexistence for 3GPP 5G and IEEE802.11ad WiGig," in *Proc. Eur. Wireless 23th Eur. Wireless Conf.*, May 2017, pp. 1–6.

[4] X. Lu, V. Petrov, D. Moltchanov, S. Andreev, T. Mahmoodi, and M. Dohler, "5G-U: Conceptualizing integrated utilization of licensed and unlicensed spectrum for future IoT," *IEEE Commun. Mag.*, vol. 57, no. 7, pp. 92–98, Jul. 2019.

[5] Y. Lin, X. Sun, Y. Gao, W. Zhan, X. Wang, and X. Chen, "Fair and efficient spectrum sharing in unlicensed bands: Does number of links matter?" *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9459–9471, 2023.

[6] U. Challita, H. Ryden, and H. Tullberg, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 12–18, Jun. 2020.

[7] S. M. Kala, V. Sathya, K. Dahiya, T. Higashino, and H. Yamaguchi, "Identification and analysis of a unique cell selection phenomenon in public unlicensed cellular networks through machine learning," *IEEE Access*, vol. 10, pp. 87282–87301, 2022.

[8] S. M. Kala, K. Dahiya, V. Sathya, T. Higashino, and H. Yamaguchi, "LTE-LAA cell selection through operator data learning and numerosity reduction," *Pervas. Mobile Comput.*, vol. 83, Jul. 2022, Art. no. 101586.

[9] S. M. Kala, V. Sathya, K. Dahiya, T. Higashino, and H. Yamaguchi, "Optimizing unlicensed coexistence network performance through data learning," in *Mobile and Ubiquitous Systems: Computing, Networking and Services*. Cham, Switzerland: Springer, 2022, pp. 128–149.

[10] V. Huang, A. Bertze, and S. Corroy, "Adaptive cell selection in heterogeneous networks," U.S. Patent 10 264 496, Apr. 16, 2019.

[11] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 607–617.

[12] E. Carrizosa, J. Ramírez-Ayerbe, and D. R. Morales, "Generating collective counterfactual explanations in score-based classification via mathematical optimization," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121954.

[13] C. Fernández-Loría, F. Provost, and X. Han, "Explaining data-driven decisions made by AI systems: The counterfactual approach," 2020, *arXiv:2001.07417*.

[14] N. Naderializadeh, M. Eisen, and A. Ribeiro, "Wireless power control via counterfactual optimization of graph neural networks," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, May 2020, pp. 1–5.

[15] A. Terra, R. Inam, P. Batista, and E. Fersman, "Using counterfactuals to proactively solve service level agreement violations in 5G networks," in *Proc. IEEE 20th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2022, pp. 552–559.

[16] Y. Wu, R. A. Barton, Z. Wang, V. N. Ioannidis, C. De Donno, L. C. Price, L. F. Voloch, and G. Karypis, "Predicting cellular responses with variational causal inference and refined relational information," 2022, *arXiv:2210.00116*.

[17] S. M. Kala, V. Sathya, M. Ghosh, T. Higashino, and H. Yamaguchi, (2023), "Unlicensed cellular operator dataset from public LAA networks," *IEEE DataPort*, doi: 10.21227/e1na-0454.

[18] V. Sathya, M. I. Rochman, and M. Ghosh, "Measurement-based coexistence studies of LAA & Wi-Fi deployments in Chicago," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 136–143, Feb. 2021.

[19] V. Sathya, S. M. Kala, M. I. Rochman, M. Ghosh, and S. Roy, "Standardization advances for cellular and Wi-Fi coexistence in the unlicensed 5 and 6 GHz bands," *GetMobile, Mobile Comput. Commun.*, vol. 24, no. 1, pp. 5–15, Aug. 2020.

[20] Mobile Communications Association. (2021). *Spectrum Sharing GSMA Public Policy Position*. [Online]. Available: https://www.gsma.com/spectrum/wp-content/uploads/2021/06/Spectrum-Sharing-Positions.pdf

[21] *XCAL-Mobile, Handheld Air Interface Field Testing Solution*. [Online]. Available: https://accuver.com/sub/products/view.php?idx=10&ckattempt=2

[22] *QualiPoc Android, The Premium Handheld Troubleshooter*. [Online]. Available: https://www.rohde-schwarz.com/us/products/test-and-measurement/network-data-collection/qualipoc-android63493-55430.html

[23] (2020). *Network Signal Guru*. [Online]. Available: https://play.google.com/store/apps/details?id=com.qtrun.QuickTest&hl=enU.S

[24] *Network Cell Info Lite & Wifi*. [Online]. Available: https://www.appbrain.com/app/network-cell-info-lite-wifi/com.wilysis.cellinfolite#

[25] *FCC Speed Test App*. [Online]. Available: https://play.google.com/store/apps/details?id=com.samknows.fcc

[26] S. M. Kala, V. Sathya, E. Yamatsuta, H. Yamaguchi, and T. Higashino, "Operator data driven cell-selection in LTE-LAA coexistence networks," in *Proc. 22nd Int. Conf. Distrib. Comput. Netw.*, Jan. 2021, pp. 206–214.

[27] Y. Li, C. Peng, Z. Yuan, J. Li, H. Deng, and T. Wang, "MobileInsight: Extracting and analyzing cellular network information on smartphones," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2016, pp. 202–215.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[29] V. Sathya, A. Ramamurthy, M. I. Rochman, and M. Ghosh, "QoS guaranteed radio resource scheduling in stand-alone unlicensed MulteFire," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Sep. 2020, pp. 86–91.

**SRIKANT MANAS KALA** is currently a Visiting Assistant Professor with the Mobile Computing Laboratory, Osaka University, Japan. He is also the Co-Founder and the CEO of Veritus, a research companion for researchers. His research interests include extended reality, unlicensed and 5G networks, applied AI/ML, and venture capital investment. He received the IIT Hyderabad Research Excellence Award, in 2016 and 2017. He was a Semifinalist of the 2020 Ericcson Innovation Awards and the Impact Summit Finalist of the 2021 Hult Prize.

**MALVIKA MISHRA** received the degree in computer science and engineering from Manipal University, Jaipur. She is currently a Research Intern with Osaka University. Her research interests include machine learning, explainable AI, and NLP.

**VANLIN SATHYA** (Member, IEEE) received the Ph.D. degree in CSE from IIT Hyderabad, India, in 2016. He is currently working on cutting-edge research and new technology initiatives with the CTO Group, Cleona Inc., USA, where he is primarily responsible for private 5G deployment. Before this, he was a Postdoctoral Scholar with the University of Chicago, USA, where he primarily focused on the issues faced in the 5G real-time coexistence.

**TATSUYA AMANO** (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information and computer sciences from Osaka University, Japan, in 2016, 2018, and 2021, respectively. He is currently an Assistant Professor with Osaka University. His research interests include 3D sensing and spatial computing.

**MONISHA GHOSH** (Fellow, IEEE) received the B.Tech. degree from Indian Institute of Technology, Kharagpur, India, in 1986, and the Ph.D. degree in electrical engineering from the University of Southern California, in 1991. She is currently a Professor with the Electrical Engineering Department, University of Notre Dame, and the Policy Outreach Director of the NSF Spectrum Innovation Center, SpectrumX. Prior to this, she was the Chief Technology Officer with FCC, the Program Director of NSF, and a Research Professor with the University of Chicago.

**TERUO HIGASHINO** (Senior Member, IEEE) is currently a Professor and the Vice President with Kyoto Tachibana University, Japan. He is also a specially appointed Professor with the Graduate School of Information Science and Technology, Osaka University, Japan. Since 2018, he has been the PI of the Society 5.0 Project of Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. His research interests include localization, ultra-low power IoT devices, future smart and connected communities, and disaster mitigation. He was a member of the Science Council of Japan (SCJ), from 2014 to 2020, and the Vice President of the Information Processing Society of Japan (IPSJ), from 2016 to 2018.

**HIROZUMI YAMAGUCHI** (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in information and computer science from Osaka University, Osaka, Japan, in 1994, 1996, and 1998, respectively. He is currently a Full Professor with Osaka University and leading the Mobile Computing Laboratory. His research interests include mobile and pervasive computing and communication networks. He was awarded the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, in 2018. He has served on ICDCN-2021 and Mobiquitous-2021 as the General Co-Chair and at many conferences, such as IEEE PerCom as a Technical Committee Member.