IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 26, 2021, accepted November 3, 2021, date of publication November 18, 2021,
date of current version December 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129281

A Zero-Touch Network Service Management Approach Using AI-Enabled CDR Analysis

ALI RIZWAN[1], MONA JABER[2], (Senior Member, IEEE), FETHI FILALI[1], (Senior Member, IEEE), ALI IMRAN[3], (Senior Member, IEEE), AND ADNAN ABU-DAYYA[1], (Senior Member, IEEE)

[1]Qatar Mobility Innovations Centre, Qatar University, Doha, Qatar
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K.
[3]Department of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK 74135, USA

Corresponding author: Ali Rizwan (arizwan@qmic.com)

This work was supported by the Qatar National Research Fund (QNRF) (a member of The Qatar Foundation) under Grant NPRP12-S 0311-190302.

ABSTRACT The detection of cells with sub-optimal performance and the identification of the root-cause of such performance is a crucial and challenging task in Network Performance Management (NPM). The contemporary NPM approaches, being reactive, silo-based, and highly expert-reliant, are not viable options for such tasks anymore, particularly in the emerging complex heterogeneous mobile networks. The state-of-the-art research in the field of data-driven Artificial Intelligence (AI) is a ray of hope for developing innovative solutions for such NPM tasks. However, the scarcity of holistic and detailed real network data limits the potential of this approach. In this study, we present a comprehensive AI-driven framework for the auto-diagnosis of cells with sub-optimal performance in a real network. We have explored and shared insight about an untapped comprehensive Call Detail Record (CDR) dataset from a real network operator. The outcome is anonymous and annotated data made public to encourage further research in this domain. We employ a $K - means$ clustering method that exploits CDR data and domain experts' input for the identification of particular types of cell performances. Next, a support vector machine-based classifier is developed for real-time applications which classifies the network nodes based on their performance with an accuracy of 97.69%. Subsequently, we introduce an algorithm that uses the classification results for the root-cause analysis of sub-optimal performance by leveraging network topography and area knowledge. The method succeeds in reaching the outcomes of an expert-led root-cause analysis and beyond. At the same time, the algorithmic approach limits the manual root-cause analysis to 30 possible scenarios per hour as opposed to analysis of 759 cells, thus it reduces the workload of an expert significantly. In the broad picture, the proposed AI framework lays the foundation towards zero-touch mobile network and service management starting with automated NPM and root-cause analysis.

INDEX TERMS ZSM, zero-touch, mobile network, service management, performance management, CDR, machine learning, automation.

I. INTRODUCTION

Mobile networks have experienced a meteoric transformation since their first phase of worldwide commercialization in the late 1990s. That phase was dubbed the second generation of mobile networks or 2G. Today's expedited deployment of the Fifth Generation (5G) and the research on the next generation (6G) are the outcomes of this transformation. From a user's perspective, this evolution is translated into a seamless improvement in Quality of Experience (QoE) for human and machine centered applications. Mobile networks must undertake disruptive changes at a fast pace to enable the delivery of the expected QoE by offering adaptability in service provisioning. To this end, networks must be re-thought by taking into consideration the explosive demand in traffic, differing QoE expectations, and large variations in traffic volume on small-scale temporal units. In order to sustain cost-effectiveness, mobile networks are required to support fast

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Abdur Razzaque.

automated adaptation in view of changing circumstances to reshuffle the limited resources for optimum allocation.

Artificial Intelligence (AI) can transform network management into a cognitive process through which the network can self-adapt and self-react to changing conditions with minimal manual intervention (zero-touch). It promises to facilitate in automating the complex tasks such as planning, maintenance, and optimization of the network [1]. Towards making networks intelligent and autonomous, different initiatives have been taken in recent times. For example, *SelfNet1* is the first phase of a project by 5G-PPP that mainly aims at the autonomous management of Network Function Virtualization (NFV) in NFV/ Software Defined Network (SDN)-enabled 5G networks. Similarly, *SliceNet3* is the phase II project from 5G-PPP that focuses to build a framework for End-to-End cognitive network slicing and slice management [2].

The 3rd Generation Partnership Project (3GPP) release 17 employs an NFV/SDN approach to the network architecture and introduces a new function called Network Data Analytics Function (NWDAF), that leverages AI for data-driven network automation [3]. In parallel, the Open RAN Alliance proposes a radio access network architecture based on NFV/SDN and defines the Radio Intelligence Controller (RIC) to enable the data-driven AI for network automation [4]. Zero-touch Network and Service Management (ZSM) refers to the next phase of automation which embodies an end-to-end framework for network orchestration that leverages data-driven AI and does not require human intervention [2], [5]. ETSI ZSM industry specification group was formed in 2017 with the goal to accelerate the process of defining the required end-to-end architecture and solutions [6].

ZSM may be seen as the advanced manifestation of Self Organising Networks (SON), an older 3GPP effort that aimed at automating some aspects of mobile networks [7]. SON, however, saw limited success for two reasons. The first pertains to the ''black box'' feature of the SON algorithm that is vendor-specific and often opaque. As a result, network operators are hesitant to adopt SON on a large scale as they prefer to keep the human-in-the-middle when it comes to abiding by Service Level Agreements (SLA). The second factor is related to the design of the SON algorithm which is generally based on manually defined rules. The drawback here is, such rules are often network-centered and specific to particular scenarios, so they need to be re-tuned following changes in the environment and the introduction of new network services. This need for frequent human intervention limits the benefit of the automation aspect [8]. Thus, prior to ZSM, mobile networks are still at the risk of overspending due to over-provisioning or violating the SLA as a result of under-provisioning.

To harness AI required for enabling ZSM, role of relevant data is pivotal, besides that of Machine Learning (ML) algorithms. One such important dataset is CDR. But, very few works in the literature examine the possibilities of enabling ZSM using AI-driven Call Detail Record (CDR) data analysis. It is so because, the CDR data is very difficult to obtain as it contains detailed personal information about the network users such as their phone numbers, whereabouts, social network, and the phone used. In addition, CDR data reveals the internal structure of a mobile network, connections, and routes. It can expose vulnerable nodes that may be maliciously targeted.

For the privacy and security concerns, mobile network operators commonly share only anonymous and aggregate CDR data Rather than the raw CDRs. For instance, the works in [9]–[12] use aggregated CDR dataset publicly shared by Telecom Italia.[1] Each of these works employs different AI techniques to detect anomalies in traffic surges per square grid, which is the highest resolution available in the dataset. However, this level of abstraction hinders the possibility of data-driven network fault diagnosis. Authors in [13] employ deep learning methods on the same dataset to detect sleeping cells, a notorious problem in mobile networks. Authors in [14] propose an automated mechanism to annotate CDR data collected locally between a radio site and the mobility management equipment. The dataset is then examined to extract the behavior of the users of this radio site. Nevertheless, due to its narrow scope, this study does not capture the relations and dependencies among various network layers and the dynamic traffic.

Recently, there has been a rise in various efforts and research in different areas which would ultimately contribute towards ZSM. For example, in [15], authors have proposed a knowledge plane-based Management and Orchestration (MANO) framework dedicated for zero-touch network slicing that exploits deep reinforcement learning model to minimize energy consumption and virtual network function (VNF) instantiating cost. Another recent study [16] presents a decentralized Federated Learning scheme for content offloading to user equipment or network edge. It takes advantage of the future user demand predicted by the ML models. This scheme can be used in caching and load balancing for ZSM. In contrast, our work examines another key stream in ZSM which relates to Network Performance Management (NPM) with a focus on automated fault diagnosis.

In this manuscript, we present the first effective NPM framework toward ZSM which automates both fault detection and root-cause analysis. We put forward an ML model that feeds on streaming CDR to identify network cells with sub-optimal performance. These are analyzed online to locate the cause of the problem based on a novel ML Algorithm. The algorithm proposed identifies the greatest common features of cells with sub-optimal performance: geographical description and network architecture. Our work is validated with a real CDR dataset collected at the core network of an African operator and annotated by domain experts to a level of details that is not available in any of the existing studies. Our approach overcomes the limitations faced by previous

[1] 2014, [online] Available: https://dandelion.eu

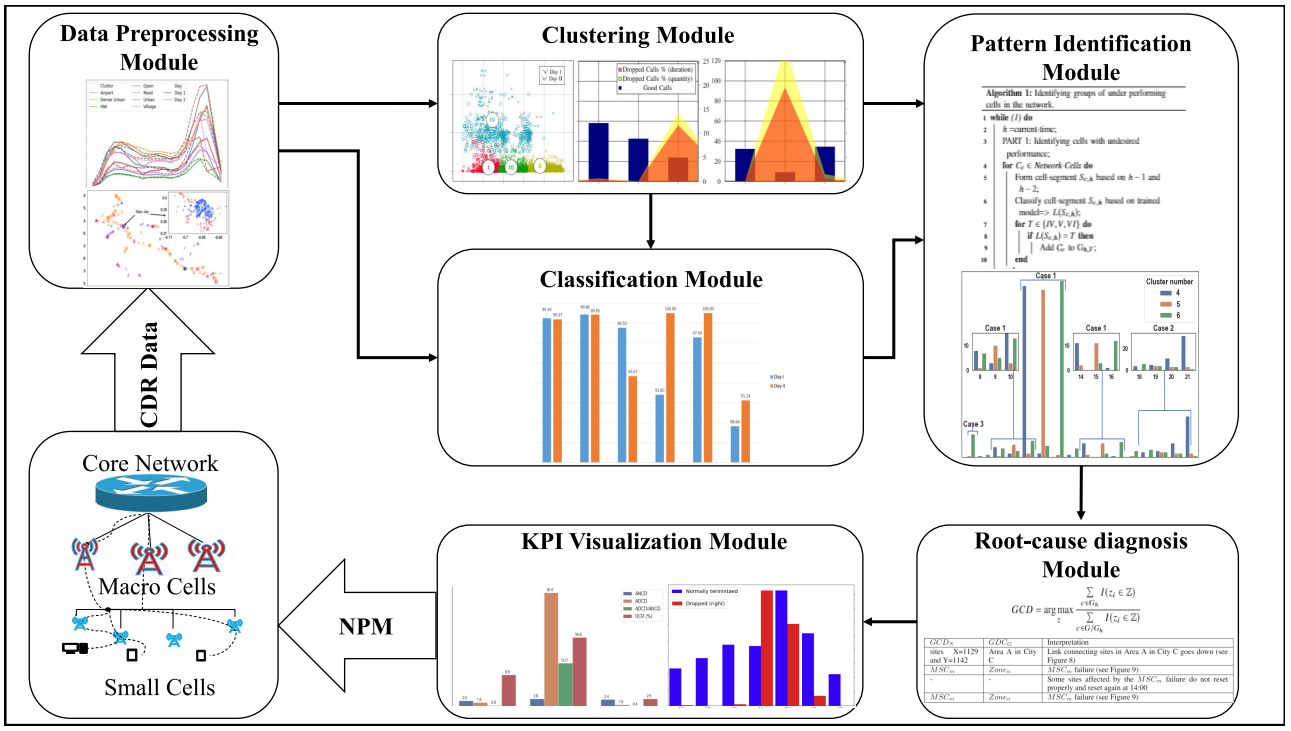


FIGURE 1. Framework for the detection and classification of cells based on their performance and diagnosis of the root cause of their sub-optimal performance.

works such as [9]–[11] owing to our access to detailed cell-level data and information about the network architecture, users' profiles, and detailed records of all dropped user connections. Moreover, this ML-based diagnosis method being the first such solution to address the need for real-time performance analysis is a steppingstone towards the design of ZSM-compliant proactive elastic mobile networks.

A. CONTRIBUTIONS

We summarize our contributions in the following points:

- We propose the first comprehensive framework shown in Figure 1 for the data-driven automation of mobile network performance analysis. We employ our previously designed classifier that identifies network cells with sub-optimal performance and the type of issue observed based on the streaming CDR data [17]. We present new results in which we validate the model in [17] on unseen data with encouraging accuracy of more than 97%.
- We present the first detailed and annotated network-wide CDR dataset which resulted from our processing of full raw CDR records. We make this dataset available to the public in the hope that it will motivate more research in this area. The data includes:
  - Hourly data per cell: Traffic volume, number of calls, cause of termination, the time before the drop, average call duration, number of SMS.
  - Each cell in the network is annotated based on the region it covers, the specific area within this region, and the type of land it covers also called clutter type (e.g., some clutter types identified are Urban, Village, Highway (HW), etc.).
  - Anonymous coordinates are associated with every cell while preserving the dominant features (e.g. relative positions).
  - Each site in the network is annotated based on the associated Location Area Code (LAC), a key identifier in the mobile network architecture.
- Based on an expert's approach towards network problem analysis, we propose and elaborate an ML-driven root-cause analysis algorithm that automates the detection and framing of network faults. Further to that it identifies the root-cause of that fault, where applicable, all in automated manner. In cases where the root-cause of the fault is not identified, a detailed KPI report is automatically generated to expedite the work of a network expert examining the problems. The outcome is an explicable report that presents the AI-based reasoning for the faults identified and allows an efficient, speedy, and cost-effective counter measures. The method is applied to unseen data and the outcome is validated by a domain expert.

Overall, the paper is structured as follows. A literature review conducted on NPM is presented in Section II. Section III introduces the CDR dataset and the annotation approach. Section IV describes the classification algorithm and results for the classification of cells based on their performance. This is followed by Section V in which we

present a domain expert's approach to interpreting the data by visual inspection. On the other hand, Section VI proposes an ML-based root-cause analysis algorithm that is validated by the domain expert's interpretation. The paper is concluded in Section VIII. Note: in this work, we interchangeably use the terms cells or radio sites to refer to a uniquely identifiable radio unit providing wireless coverage.

II. STATE-OF-THE-ART IN NPM FOR ZSM NETWORKS

The traditional NPM relied mostly on a silo-approach for the data analysis which focused on subsystems within the network or deep packet inspection of interfaces connecting them. Moreover, performance analysis relied heavily on manually engineered features to identify problematic network nodes; these are referred to as Key Performance Indicators (KPI). Networks were generally over-provisioned to account for peak traffic. When needed, resources were reshuffled (or turned off) based on fixed patterns such as traffic change between weekdays and weekends or between summer and winter seasons. Previously, daily or weekly reactive measures were sufficient to optimize the network performance and compete with other mobile network service providers. Such an approach is no longer valid in the 5G and beyond networks which are significantly more complex and diverse than legacy networks.

Intuitively, complexity in structure, due to heterogeneous and ultra-dense deployment of cells, comes with bigger challenges in optimization. On the other hand, this complexity also offers more flexibility and a higher degree of freedom in optimization. At the same time, in the emerging networks, the traffic demand changes much faster, spatially and temporally, than the traditional voice-centric networks. The potential gain of fast pro-active optimization that matches the speed of change in the traffic and benefits from the offered degrees of freedom in the network is significant. This gain is of pivotal importance to network operators as it enables a cost-effective network deployment as opposed to the crippling cost of traditional over-provisioning. In order to capitalize on the offered degrees of freedom and traffic diversity, a holistic NPM approach is essential to curtail the limitations of the silo-based analysis. Indeed, different sections of the network can no longer be looked at independently as separate subsystems since these have transformed into a fluid architecture that requires delicate tuning [18], [19]. To this end, operators need to leverage holistic data-driven AI methods to automate the process of NPM and unravel the gain of indispensable elasticity.

ZSM leverages the data-driven AI tools to automate the network functions for creating the much-needed adaptability in mobile networks. Data-driven approaches naturally adjust to the changes in the environment as they feed and learn from these data [20]–[22]. In addition, the flexibility in the implementation of AI allows for the automation of performance management and root-cause analysis whilst giving the operator the option to keep the human-in-the-middle. Moreover, recent solutions for explainable AI reinforce the operator's propensity to trust the AI algorithm and capitalize fully on the potential of automation [23], [24]. As much as AI is a pivotal enabler of ZSM, it still faces limitations and presents security risks that need to be addressed ahead of its full-scale deployment [2]. To this end, a distributed ledger technology is proposed in [25] to implement distributed security and trust in multi-tenant and multi-stakeholder environments. An augmented reality approach is proposed in [26] to allow network administrators to understand real-time automated tasks as a path to a true ZSM network. Few works examine the problem of resource-to-network-slice allocation and employ AI in network traffic forecasting, such as [27] and [28], in order to optimize the allocation mechanism. Authors in [29] examine a radio-signal dataset using AI to learn the mobility behavior of users and predict, accordingly, the traffic load.

On the other hand, authors in [30] integrate deep reinforcement learning and federated learning to enable the inter-node exchange of learning parameters to improve the model without overloading the network with control data. The model is evaluated for the optimization of caching and computation offloading tasks at the mobile edge. A context-aware reinforcement learning approach is presented in [31] in which the authors jointly optimize the connectivity and computational speed of the Internet of Things (IoT) network in a smart port to deliver the qualities required by each vertical. These works shed light on the many applications of AI in the mobile network domain. However, none addresses the issue of performance management and root-cause analysis.

In this work, we propose an NPM automation framework that embodies the first step toward ZSM networks. The proposed framework is composed of two stages. First, a CDR-driven ML algorithm is put forward to automate fault detection in near-real-time. Next, a novel ML-based method is presented which computes the greatest common features of cells with the sub-optimal performance from two perspectives: geographical description and network architecture to locate the root-cause of their behavior. This work is possible thanks to a real CDR dataset that is annotated by domain experts to a level of details which is not available to any of the existing studies.

III. DATA DESCRIPTION AND PREPROCESSING

Through this work, we offer an annotated CDR dataset with more network-related information to allow fault detection and diagnosis. In this section, we present an overview of the data collection followed by the steps that we have taken in order to anonymize and annotate the raw CDR dataset obtained from an African network operator.

A. NETWORK OVERVIEW

This work is based on real network data acquired from an African GSM (Global System for Mobile Communication) network operator. The analysis presented here is based on the data collected on two Fridays, one in September and the other one in November 2017. Figure 2 shows the distribution of the 741 radio access cells in the network form which the

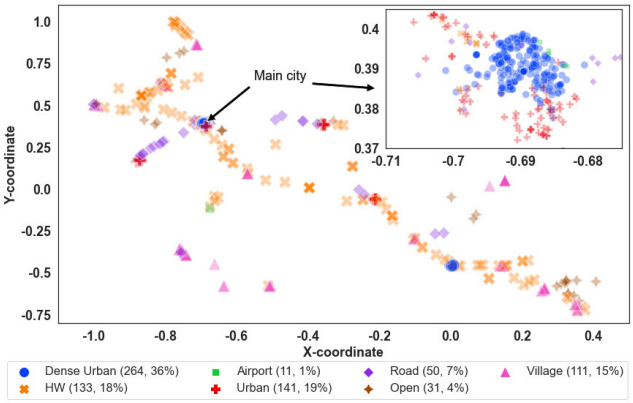


FIGURE 2. **Country-wide distribution of radio access cells in the designated network. The numbers between parenthesis in the legend represent the number of cells in each clutter category and the percentage of the total number of cells.**

data is collected. The marker for the each cell reflects the clutter type as shown in the Figure 2 and discussed in the Section III-D. As it can be seen, the network coverage is restricted to main cities and towns and connecting roads and highways. Many open areas seem to be without network coverage. The zoomed area in Figure 2 shows the distribution of 356 cells with the clutter types represented by different markers for a major metropolitan city. It further highlights that urban cells constitute more than half of the total number of cells in the network. The legend in the Figure 2 also presents the total number and the percentage of cells belonging to each clutter type in the network.

The hourly voice traffic load in the network on a Friday is presented in Figure 3 for each individual clutter type. From Figure 3, it is clear that the peak hour of the network occurs in the evenings between 19:00 and 21:00. It can also be noted that cells covering urban areas carry the highest volume of traffic, whereas those covering roads and highways carry the least. We have observed similar behavior in the voice traffic

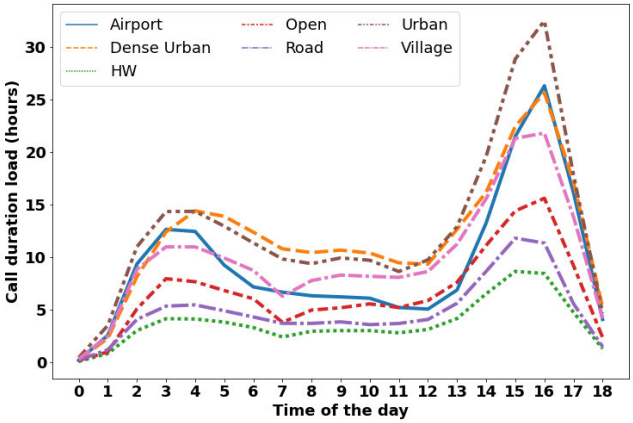


FIGURE 3. **Traffic trend during the day in the designated network per clutter type.**

on the other Friday. This confirms the existence of cell traffic patterns that depend on the time of the day and the clutter type of the covered area. For simplicity, here we present the traffic behavior of one day only.

B. CDR OVERVIEW

CDRs represent the metadata and the detail key information about the of mobile communication while excluding its content. CDRs are generated by telephone switches, mainly for billing purposes, whenever a subscriber consumes dedicated network resources [32]. For example, making (by Party A) or receiving (by Party B) a call, exchanging data (with another user or server), and conducting a location update, that all requiring dedicated resources. The non-content information includes details about Party A and Party B, their respective locations, time and duration of the call, the type of communication, the call routing, among other information. CDR generation is standardized by the 3GPP and more details can be found in [33].

In Table 1, we present the data included in the raw CDR dataset obtained from the African operator. From the Table 1 it can be seen, each record in the CDR provides complete information about the established connection, the mobile devices involved, the network nodes that serve it, and the label for the cause of termination. The value in the cause-of-termination field indicates whether the call was successfully terminated or forced to terminate. In the latter scenario, the value further specifies the dominant reason for undesired termination. Thus, the CDR information is ideal for the NPM of complex networks for two reasons. First, it offers a unique holistic perspective about the network end-to-end

TABLE 1. **CDR fields available in original dataset.**

Call type	Originating call, Terminating call, Short Message Service (SMS)
Date	
Start time	
Call duration	
IMSI of Party A	International Mobile Subscriber Identity
Number of Party A	
Number of Party B	
MCC	Mobile Country Code (unique per country)
MNC	Mobile Network Code (unique per network)
LAC	Location Area Code (limits paging area)
Cell ID	Identifies a cell in a site
ID of MSC	Mobile Switching Centre
ID of communication	Used to link Party A and Party B CDR records
Name of outgoing trunk	Outgoing trunk (or route)
Name of incoming trunk	Incoming trunk (or route)
IMEI of Party A	International Mobile station Equipment Identity. This includes information on the origin, model, and serial number of the device
IS_IMEI_TRACKED	Blacklist of illegal devices
Disconnecting party	Party A, Party B or Network
Cause of termination	Normal termination, No answer, congestion, timer expiry, destination not available, etc.
Location of termination	Where in the network the disconnection decision was made

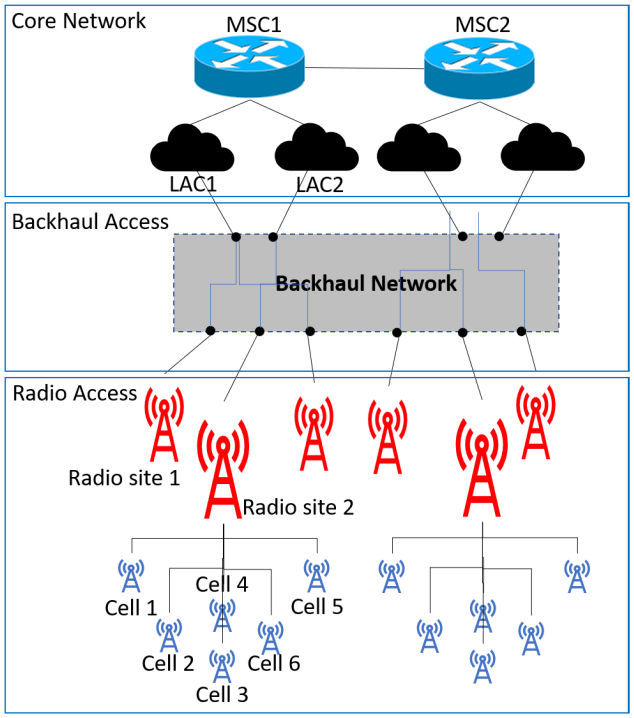


FIGURE 4. Network topology extracted from CDR dataset.

performance and, second, it is automatically labeled by the network nodes.

The network topology that can be extracted from the CDR dataset is shown in Figure 4. This figure shows the relation between cell, radio site, trunk, area defined by an LAC, and Mobile switching center (MSC), which will be of pivotal importance in the root-cause analysis. The geographical coverage of each of these network layers depends very much on the network's specificity, such as operating frequencies, tower height, backhaul network technologies, and core network planning. In Table 2, we extract the general coverage range of each network layer from our data.

TABLE 2. Mapping of network architecture to geographical span.

Network layer	Geographical span
MSC	Region in the country (includes multiple LACs)
LAC	A town, part of a city, group of highways, indoor area, etc.
Radio Site	An area of up to 50 sq.km (includes 1-6 co-located directional cells)
Cell	A sub area within the radio site (covers an angle of radio site coverage)

C. CDR GROUPING

The fields in Table 1 reveal private information about subscribers, their devices, and their whereabouts. In this pre-processing phase, we perform CDR grouping to remove any information related to subscribers while retaining the data that concerns network usage.

As highlighted in Section III-B, each entry in a CDR pertains to one leg of a given communication. For instance, a leg could be the outgoing call from Party A or the incoming call to Party B. Moreover, each of these entries includes a field labelled as *"Cause of termination"* (see Table 1) which qualifies this communication and may describe the immediate behavior of one or more of the network nodes that serve it. It is, thus, undesirable to take NPM actions on these nodes based on a single CDR's label. Therefore, a second phase of preprocessing is required in which streaming CDRs are aggregated based on different criteria like the network nodes involved, the subscribers types, and the devices used. For instance, it would be interesting to group together the CDRs labelled as *abnormal termination*,[2] the CDRs that include the same type of device, the CDRs of all roaming numbers, or all the CDRs that originate from the same cell or relate to a particular application, etc. Among these groups, some of them cannot be fixed immediately such as problems that relate to a particular mobile device type. The actionable groups are those that relate to network nodes, and a pertinent NPM mechanism would be able to build these groups while CDRs are streaming.

In this work, the pre-processing phase was conducted offline and the CDRs were grouped per hour. These are further sorted into groups that represent communications that either originated or terminated in each cell of the network. It should be noted that the unit of time (1 hour) used in this process is a design parameter that may be changed, depending on the type of data service. It is not recommended to reduce this aggregation period below 30 minutes for two reasons. Firstly, intermittent degradation in performance may be caused by factors external to the network; the root-cause analysis proposed in this manuscript aims to detect network issues that can be remedied. Second, taking frequent (less than half an hour apart) actions based on detected network problems may lead to network instability. In order to make the anonymous dataset available for the open research, the information in the CDRs was aggregated to a level as shown in Table 4. Where we have removed all data relating to mobile subscriptions and devices. The traffic data that was extracted from the CDRs was aggregated based on the associated cell. The traffic of each cell was divided into two groups depending on the value of the termination cause: *Normal termination* and *Abnormal termination*. The traffic in each group is characterized by two metrics, as described in Table 3.

The number of fields in Table 1 are 21 whereas the reduced features in Table 4 are only four. The decrease in the number of features comes with a reduction in the information about the network behavior, but, in our case, it is essential to ensure data protection and network anonymity. However, we would like to emphasize that the reduction ratio is a design parameter that may be adjusted according to the type of network service at hand. For instance, traffic volume and count are not enough

[2]Abnormal termination in this work is any Cause of termination that is not normal: No answer, Congestion, Timer expiry, destination not available, etc.

TABLE 3. Adopted metrics in traffic characterisation.

Metric	Description	Formula
$C_{\hat{c},h}$	the total number of calls by cell $\hat{c}$ that terminated normally during hour h	$C_{\hat{c},h} = \sum l$ where, l is a normally terminated call in cell $\hat{c}$ during hour h
$C'_{\hat{c},h}$	the total number of calls by cell $\hat{c}$ that terminated abnormally during hour h	$C'_{\hat{c},h} = \sum l'$ where, l' is an abnormally terminated call in cell $\hat{c}$ during hour h
$V_{\hat{c},h}$	the total number of seconds of all calls by cell $\hat{c}$ that terminated normally during hour h	$V_{\hat{c},h} = \sum_{l=1}^{n} \delta(l)$ where, $\delta(l)$ is the call duration of normally terminated call l in seconds in cell $\hat{c}$ during hour h, and n the total number of normally terminated calls
$V'_{\hat{c},h}$	the total number of seconds of all calls by cell $\hat{c}$ that terminated abnormally during hour h	$V'_{\hat{c},h} = \sum_{l'=1}^{C'_{\hat{c},h}} \delta(l')$ where, $\delta(l')$ is the call duration of abnormally terminated call l' in seconds in cell $\hat{c}$ during hour h, and n' the total number of abnormally terminated calls

TABLE 4. Reduced CDR data of a given cell $\hat{c}$ in the network.

		Normal termination		Abnormal termination		
Date	Time	Traffic volume	Call count	Traffic volume	Call count	Number of SMS
Day I	00:00	$V_{\hat{c},0}$	$C_{\hat{c},0}$	$V'_{\hat{c},0}$	$C'_{\hat{c},0}$	$S_{\hat{c},0}$
Day I	01:00	$V_{\hat{c},1}$	$C_{\hat{c},1}$	$V'_{\hat{c},1}$	$C'_{\hat{c},1}$	$S_{\hat{c},1}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Day II	23:00	$V_{\hat{c},23}$	$C_{\hat{c},23}$	$V'_{\hat{c},23}$	$C'_{\hat{c},23}$	$S_{\hat{c},23}$

to represent the quality of experience of data services, instead, these should be augmented by data rate, jitter, and delay information. Nonetheless, we present here a methodology for mining raw streaming CDR data and we demonstrate that, despite the high reduction ratio, the classification accuracy in Section IV is reliable and leads to sound root-cause analysis in Section VII. To the best of the authors' knowledge, none of the open-source CDR databases that describe data services has the same level of call details available in the dataset used in this work. None of those, therefore, has the same potential to be mined for root-cause analysis.

D. DATASET ANNOTATION

The steps taken in the CDR grouping in Section III-C have successfully removed any information related to subscribers, such as phone number, IMSI, IMEI, and whereabouts. The aggregated CDR presented in Table 4 aggregated exploiting the Cell Global Identifier (CGI) which is a compound number composed of the MCC, MNC, LAC, and the Cell ID (see Table 1 for the defiantion of these acronyms) [34]. The CGI, therefore, reveals the identity of the mobile operator, and jointly with a worldwide CGI database (https://www.opencellid.org/), reveals the location of each cell in the network. In fact, we have extracted the location of each cell by interrogating the worldwide CGI database. Next, we mapped coordinates of each location with the country open-source map (https://www.google.com/maps) and extracted information like the land type (it is commonly referred to as clutter type in mobile network planning and optimization) and the GPS (global positioning system) coordinates. The common clutter types identified are *Urban*, *Village*, *Road*, *Higway*, etc. as shown in Figure 2.

In order to conduct a context-aware NPM, it is sufficient to retain information about the placement of the cells relative to each other, their association with each LAC and MSC, their location in different areas, and their clutter type. Keeping the information mentioned above, the names of the country, operator, and cities were anonymized. In addition, the locations of the cells are anonymized using a key that preserves the respective distances. As a result, the unique identifier of the cell CGI is replaced with an anonymous identifier, and additional annotations are added. The Table 5 shows the saple representation of annotated information. The database used for determining the original location of the cells generated some erroneous locations where the cells were found to be in a different country or continent. From a total of 759 cells, relevant location information could be identified for 741 cells. For each of these cells, we associate a clutter type annotation. The other 18 cells remain in the database with their LAC information but do not have information about the *Area*, *Clutter Type*, or coordinates. It is worth mentioning that the location information of 741 cells mapped by the online method are not confirmed by any other source. The annotated data, therefore, is likely to include noise that may in the end affect the results of automated NPM algorithm.

TABLE 5. Anonymised network data with annotations.

Cell ID	Site	LAC	MSC	Town	Area	Clutter Type	X-coord	Y-coord
$\hat{c}_1$	S_{10}	L_6	M_M	$Town_A$	A_H	Urban	9	23
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\hat{c}_c$	S_s	L_l	M_m	$Town_t$	A_a	Road	x	y
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

IV. ML FOR CLASSIFYING NETWORK CELLS WITH SUB-OPTIMAL PERFORMANCE

In this section, we briefly discuss our ML-based scheme for the detection and classification of network performance issues. The proposed method first generates cell-instances data based on the hourly data of each cell (see Table 4). In this case, a cell-instance $S_{\hat{c},h}$ represent a three-hours sample of cell $\hat{c}$ that terminates at hour h. The aim of this method is to classify the performance of each cell-instance. The scheme developed, as presented in Figure 5, contains two main functions: clustering and classification. The first function is visited twice and corresponding results are labelled as Tier I and Tier II results. Where in both tiers the $K - means$ clustering is used for the detection and labelling of faults.

As shown in Figure 5, the feature extraction, a prerequisite step for the $K - means$ clustering, in this function, is conducted based on data analysis and exploration. The outcomes are then fed to a $K - means$ clustering. The goal of the $K - means$ clustering is to segregate cell-instances based on their performance over the three hours sliding window. Since a sliding widow is adopted in this analysis, the detection of cells with sub-optimal performance occurs on an hourly

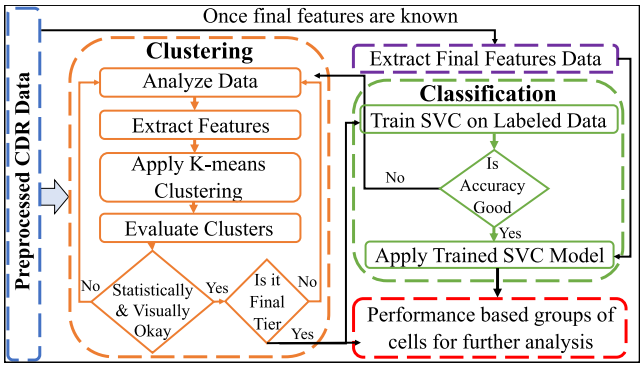
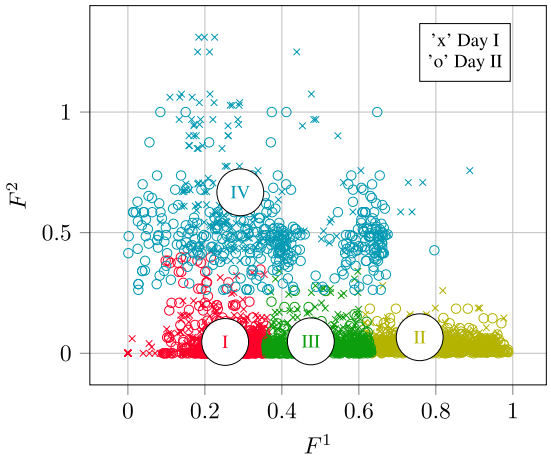
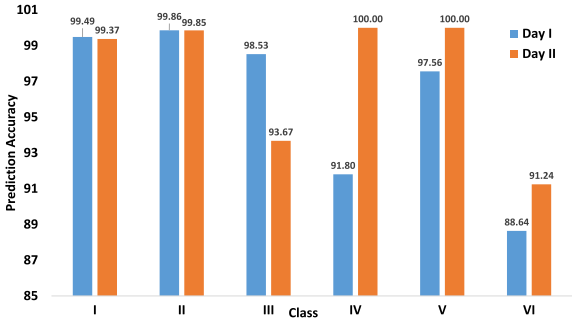


FIGURE 5. Classification scheme for the detection of cells with sub-optimal performance.



(i): Results of Tier 1 $K-means$ showing Type IV for cell-instances with sub-optimal performances based on F^1 and F^2 features. F^1 is the average duration of normally terminated calls and F^2 is the relative load of the maximum number of dropped calls, further details are available in [17]



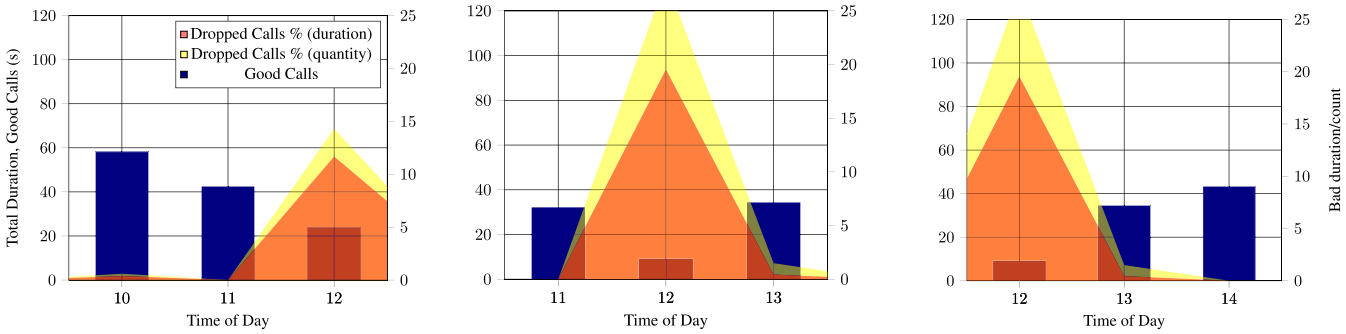
(ii): Prediction accuracy of each class for samples from Day I and Day II using SVM classifier. Types IV, V, and VI refer to different types of sub-optimal cell performances.

FIGURE 6. Results of cell performance classification.

basis. The sliding window approach is applied in the analysis to distinguish temporary performance degradation. In other words, a given cell $\hat{c}$ may be labeled as Cluster I at one time and as Cluster IV another time, during the same day. As discussed in Section III-C, it is not recommended to reduce the aggregation time unit below half an hour in order to avoid network instability.

Clusters produced by $K-means$ are evaluated statistically to measure how cell-instances within a cluster are distributed relative to each other. In essence, the objective of this step is to ensure that distinguished types of faults realted to the performance of cells could be identified. Metrics used to statistically evaluate the quality of the clusters are Root Mean Square Standard Deviation (RMSSTD) for compactness, R-squared (RS) for separation, Calinski-Harabasz index (CH), and Silhouette index (S) for compactness and separation. Besides that, domain experts visually inspected the faults identified. While we mainly rely on domain knowledge for the finalization of clusters, statistical metrics are essential to identify the optimal number of clusters.

The first $K-means$ clustering, referred to as Tier I clustering, could segregate samples, from cell-instances, into four groups as shown in Figure 6(i), with the main goal of segregating cells with the bad performance from the cells with good performance. For this purpose, the average duration of normally terminated calls, F^1, and the relative load of the maximum number of dropped calls, F^2, are used as features [17]. In Figure 6(i), we highlight three clusters: Type I with low traffic (red markers), Type III with medium traffic (green markers), and Type II with high traffic (light green markers). It can be seen that these three clusters have no significant network performance issues. On the other hand, Type IV which is presented with aqua markers is an aggregation of cell-instances with network issues. The Figure6(i) shows the Tier I clustering results of two different days (both Friday, around ten weeks apart); samples from Day I are presented with marker 'x', and samples from Day II are presented with a marker 'o'. It is clear that the clustering scheme at Tier I can group samples with the same behavior in these four prominent clusters. At the end of Tier I, the samples of interest are those gathered in Cluster IV, as they all represent cell-instances with poor performance. In Tier II clustering, these samples with poor performance are further segregated with the aim of identifying the different potential types of poor performance. Therefore, the $K-means$ clustering is called again on Cluster IV samples, referred to as Tier II clustering. It again involves the complete process of feature extraction, $K-means$ application and evaluation. But the objective here is to further segregate samples based on the type of issue causing the sub-optimal performance. The outcome of Tier II clustering is three distinct Types (IV, V, and VI), as shown in Figure 7. They depict the three common types of sub-optimal performances registered in this network. At the end of the two-Tier clustering in this study, we get six distinct clusters in total, representing the different types of traffic loads and performance behaviors. These six clusters are analyzed by a domain expert to label their respective performance type based on the traffic load and quality (see Table 6). Here it can be seen that the 701 unique cells have



(i): Type IV: Start of performance degradation

(ii): Type V: Unstable performance

(iii): Type VI: Recovering from performance degradation

FIGURE 7. Samples from three clusters representing three different types of bad performance in the cells over the three hours window of time.

TABLE 6. Distinct cell behavior trends.

Behavior	description	Number of cell-instances	Number of unique cells
Type I	Low traffic / Good QoS	4541	701
Type II	High traffic / Good QoS	2769	733
Type III	Medium traffic / Good QoS	4196	747
Type IV	QoS drop	225	176
Type V	QoS trough	176	162
Type VI	QoS rise	237	172

4541 times low traffic and overall good performance over the intervals of three hours in a day. Similarly the behaviour of cells in other classes listed the Table 6 can also be interpreted.

The second function shown in Figure 5, is the classification which uses a Support Vector Machine (SVM) based model. The model is trained based on the labeled data obtained as the result of clustering. In [17] the clustering and classification scheme is applied on the data of Day I only. In contrast, here we present results for the data of both Day I and Day II. Samples of the cells from two different days are differentiated by the marker style in Figure 6(i).

Results for the prediction accuracy for each type are presented in Figure 6(ii) with two different colors for two different days. The accuracy results reported in [17] are shown in blue bars in Figure 6(ii). These results are obtained by training the SVM classifier on 75% of Day I data and tested on 25% unseen Day I data. In this work, we have trained the SVM classifier on the complete data of Day I and tested it on the unseen complete data of Day II, shown with red bars in Figure 6(ii). We have used Linear kernel and penalty score C = 1000. These are the same parameters that yielded the best performance for Day I in cross-validation and testing [17].

The classifier trained on the data from Day I classifies the completely unseen samples from Day II with an accuracy of 97.69%. Some significant variation in the performance of the classifier can be seen in detecting Types III, and IV on two different days. However, the overall accuracy score and type-specific accuracy reflect that the model is very effective, even on unseen data of complete Day II. It also reasserts that the features extracted and used for clustering are very relevant and meaningful. One possible explanation of the observed differing model performance can be the variance in the number of cell-instances of poor performance on Day I and Day II. On Day I there are 638 cell-instances reported with poor performance. On Day II there are only 362 such cell-instances; almost half of those on Day I.

In the subsequent sections, we further investigate how the results of this hybrid scheme for classification can lead to root-cause analysis of the source of poor performance.

V. VISUAL DATA INTERROGATION AND INTERPRETATION

As discussed in Section IV, each cell-instance, $S_{\hat{c},h}$, is classified as having one of six possible behaviors listed in Table 6. In this notation, the index $\hat{c}$ refers to the anonymous cell ID and the index h indicates the most recent hour of this cell-instance. In this section, a root-cause analysis is conducted manually by examining the results of cell-instance classification conducted in Section IV.

We first examine the cell-instances with sub-optimal performance, particularly, those which are classified as, Type IV, Type V, and Type VI on Day I. The distribution of these cell-instances during the period of interest is shown in Figure 8. It is immediately evident that a major failure occurred between 10:00 and 12:00 and has affected around 130 cells. The corresponding cell-instances follow the same pattern of classification in the same chronological order: Type IV followed by Type V which precedes Type VI. We refer to this group behavior as *Case 1* and we provide a comprehensive analysis of such behavior in Section VI-A. A similar trend is observed later between 14:00 and 15:00, although fewer (only 17) cells are affected. This is another occurrence of *Case 1* group behavior and is further analyzed in Section VI-A. Moreover, by inspecting Figure 8, another sub-optimal performance trend is seen in the evening starting at 18:00 which is characterized by an exponential increase in the number of cell-instances classified with behavior Type IV. This trend is presented as *Case 2* and is further analyzed in

TABLE 7. Root-cause analysis pertaining to Case 1.

Day	SPG	No. of cells	Common features	GCD_N	GCD_G	Interpretation
1	$SPG^6_{10,VI}$	7	LAC, two sites, city, area, clutter	sites X=1129 and Y=1142	Area A in City C	Link connecting sites in Area A in City C goes down (see Figure 9)
1	$SPG^6_{12,VI}$	130	MSC, $Zone_n$	MSC_m	$Zone_n$	MSC_m failure (see Figure 10)
1	$SPG^6_{16,VI}$	13	$SPG^6_{12,VI}$	-	-	Some sites affected by the MSC_m failure do not reset properly and reset again at 14:00
2	$SPG^6_{21,VI}$	109	MSC, $Zone_n$	MSC_m	$Zone_n$	MSC_m failure (see Figure 10)

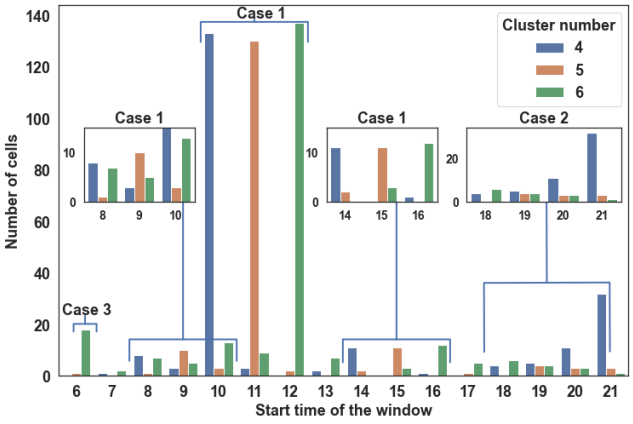


FIGURE 8. Distribution of cell-instances of Types IV, V, and VI on Day I.

Section VI-B. Another issue that can be spotted by visual inspection of Figure 8 occurs around 06:00 and is characterized by a peak in the number of cell-instances (corresponding to exactly 17 cells) suffering from Type IV. This trend is referred to as *Case 3* and is discussed in Section VI-C.

VI. ALGORITHMIC DATA INTERROGATION AND INTERPRETATION

For NPM, visual inspection is neither efficient nor cost-effective. Instead, we present Algorithm 1 which runs every time unit (an hour, in this example) with the aim of automatically extracting poor performance trends without visual inspection. The algorithm is divided into four parts. Part 1 (Lines 4-10) performs the classification of cell-instances created using a sliding window of three hours. Each cell-instance contains data of the last three hours, where hour h is the recent one, $h-1$ the last one and $h-2$ is the second last hour. All cell instances $S_{\hat{c},h}$ that are classified as sub-optimal performance type T, (i.e.issue type $T \in \{IV, V, VI\}$) at time h, are grouped into three separate sub-optimal performance group: $G_{h,IV}$, $G_{h,V}$, or $G_{h,VI}$, depending on the type of sub-optimal performance T. Some patterns are observed by analyzing the way some groups of cells have sub-optimal performance for consecutive hours or on a particular hour of the day. Such Sub-optimal Performance Groups, SPG, are identified in Part 2 (Lines 13-18), as defined below:

- $SPG^i_{h,T}$: These represent groups of cells with sub-optimal performance (i.e., $T \in \{IV, V, VI\}$), during the three hours of the sliding window and they follow the same pattern of sub-optimal performance. Patterns of such performances are described as sequence $\{T, T_1, T_2\}$ of type of issues faced at time h, $h-1$ and $h-2$ respectively. Thus, there are $3^3 = 27$ possible patterns or group formations in this category for each hour of the day. These are formed such that $\hat{c} \in SPG^i_{h,T}$ if $\hat{c} \in G_{h-2,T_2} \cap G_{h-1,T_1} \cap G_{h,T}$ and where $i \in \{1, \ldots, 27\}$ such that for $i = 1$, $T = T_1 = T_2 = IV$, for $i = 12$, $T = V, T_1 = IV, T_2 = VI$, and so on, for $i = 27$, $T = T_1 = T_2 = VI$.
- $SPG^j_{h,T}$: These represent cells that do not fit in any of the 27 groups $SPG^i_{h,T}$ at time h. There are three such groups $SPG^{28}_{h,T}$, $SPG^{29}_{h,T}$, and $SPG^{30}_{h,T}$ for T = IV, V, and VI, respectively. These are formed such that, $SPG^j_{h,T} = G_{h,T} - SPG^i_{h,T}$ for $T \in \{IV, V, VI\}$ and for all $i \in \{1, \ldots, 27\}$.

Part 3 (Lines 21-24) aims to identify a particular undesired behavior ($T \in \{IV, V, VI\}$) where the number of affected cells increases with time i.e., $|G_{h,T}| > |G_{h-1,T}| > |G_{h-2,T}|$, where $|G|$ indicated the size of the group. In other words, the actual cells that are affected by Type T may change with time, but the number of affected cells increases. This part is important to capture issues that cause a ripple effect in the network (e.g., congestion). A Focus Group $FG_{h,T} = G_{h,T} \cup G_{h-1,T} \cup G_{h-2,T}$ is formed to track the root-cause of the problem.

The last part (Lines 27-32) identifies the root-cause of the problem by inspecting both the network architecture features and the geographical/landscape characteristics of cells with sub-optimal performance. Two Greatest Common Descriptors (GCD) with respect to the network architecture (GCD_N) and the geographical distribution (GCD_G) are calculated to identify the largest common feature among the majority of cells in the group and to none (or few) outside the group, as follows:

$$GCD = \arg\max_{z} \frac{\sum_{\hat{c} \in G_h} I(z_i \in \mathbb{Z})}{\sum_{\hat{c} \in G/G_h} I(z_i \in \mathbb{Z})} \quad (1)$$

where z_i is the value of the *ith* feature from the potential feature set $\mathbb{Z}$ (see Table 5) contributing towards the bad performance found in Group G_h at time h of the day. I is an indicator function that takes 0 when a feature value is not present or 1 when it is present. Equation (1) identifies feature z with the highest frequency of its value in the bad performance group of cells G_h as compared to the rest of cells in G/G_h at time h. There are two types of features:

Algorithm 1: Identifying Groups of Cells in the Network With Sub-Optimal Performance

while *(1)* **do**
  h = current-time;
  PART 1: Identifying cells with undesired performance;
  for $\hat{c}_c \in$ *Network-Cells* **do**
    from cell-instance $S_{\hat{c},h}$ based on $h-1$ and $h-2$;
    Classify cell-instance $S_{\hat{c},h}$ based on trained model=> $L(S_{c,h})$;
    for $T \in \{IV, V, VI\}$ **do**
      if $L(S_{c,h}) = T$ **then**
        Add $\hat{c}$ to $G_{h,T}$;
    end for
  end for
  PART 2: Grouping of cells with similar sub-optimal performance;
  for $T \in \{IV, V, VI\}$ **do**
    for *All* $\hat{c}_c \in G_{h,T}$ **do**
      for $T_1 \in \{IV, V, VI\}$ **do**
        for $T_2 \in \{IV, V, VI\}$ **do**
          if $\hat{c}_c \in G_{h-2,T_2} \cap G_{h-1,T_1} \cap G_{h,T}$ **then**
            Calculate value of i;
            Add $\hat{c}_c$ to $\mathrm{SPG}^i_{h,T}$;
            Set $Member-of-SPG(C_c) = True$
        end for
      end for
      if $Member - of - SPG(C_c) = False$ **then**
        Calculate value of j;
        Add $\hat{c}_c$ to $\mathrm{SPG}^j_{h,T}$
    end for
  end for
  PART 3: Identifying network trends;
  for $T \in \{IV, V, VI\}$ **do**
    if $|G_{h,T}| > |G_{h-1,T}| > |G_{h-2,T}|$ **then**
      Find $\mathrm{FG}_{h,T} = \mathrm{G}_{h,T} \cup \mathrm{G}_{h-1,T} \cup \mathrm{G}_{h-2,T}$
  end for
  PART 4: Root-cause analysis and KPIs;
  for *All* $g \in SPG$ *where* $|SPG| \geq MinSize$ **do**
    Find $GCD_N(g)$ and $GCD_G(g)$ using Eq (1);
    if $GCD_N(g)$ *is empty and* $GCD_G(g)$ *is empty or* $|SPG| < MinSize$ **then**
      Generate cell specific KPIS for past 24h;
      ANCD, *ADCD*, *ADCD*/*ANCD*, *DCR*
  end for
end while

Network-related and Geography-related. The network-related features are ranked from greatest to smallest as follows: *Network* > *MSC* > *LAC* > *site* > *cell*. Whereas the geography-related features are ranked according to the size of the area covered by the problematic group *Countrywide* > *Region* > *Town* > *AreaInTown* > *Scattered*. In this case, *Scattered* refers to a feature that is common to multiple scattered areas, such as clutter type. For example, for a given LAC, if most of its cells belong to G/G_h and few or none have normal performance at time h, then LAC will be selected as network-related GCD (or GCD_N). From a geography-related perspective, if the most of the cells with sub -optimal performance are located in the same Town *TownA* and no/few cells in *TownA* exhibit normal behavior, it can be said that the GCD_G is *TownA*. In case no GCD is found from both network and geography perspectives, a detailed KPI report is generated to facilitate the diagnosis analysis by a domain expert.

Algorithm 1 successfully identifies all the issues that are visually observed by examining Figure 8. In addition, two previously undetected events are further identified on Day I. The first event follows the trend *Case 1*; it takes place at 8:00 and involves seven cells. The second occurs at around 13:00 where five cells follow a new pattern of sub-optimal performance: Type VI followed by Type IV, then Type V. This is discussed in Section VI-D with other cell-specific root-cause analysis. More importantly, the algorithmic approach guarantees that no other event of interest goes unnoticed during the period of interest.

A. ROOT-CAUSE ANALYSIS: CASE 1

There are three occurrences of this case in the CDR data of Day I and one occurrence in data of the Day II. This trend is followed by a group of cells that are classified as Type IV at time $h-2$, followed by Type V at $h-1$ and Type VI in the current hour h. These cases are detected by the algorithm (Part 2) as $\mathrm{SPG}^6_{h,VI}$ and also as $\mathrm{SPG}^{28}_{h-2,IV}$. In fact, the outcome of Algorithm 1 highlights only four such groups $\mathrm{SPG}^i_{h,T}$ (where $i \leq 27$) in which the size of $(\mathrm{SPG}^i_{h,T}) > 3$. The first three groups are in Day 1, $\mathrm{SPG}^6_{h,VI}$ at $h = \{10, 12, 16\}$.

For each of the groups $\mathrm{SPG}^6_{h,VI}$, we calculate GCD_N and GCD_G using Eq (1), as listed in Table 7. Based on the GCD_N and GCD_G findings for each SPG, we offer an interpretation validated by a domain expert, as shown in Table 7. The sub-optimal performance group captured in Day I $\mathrm{SPG}^6_{10,VI}$ is presented by Figure 9. Cells belonging to $\mathrm{SPG}^6_{10,VI}$ are highlighted in blue; these are mostly from two problematic sites in the same area. Thus, it can be seen that the GCD_G is Area A in City C and the $GCD_N = \{1129, 1142\}$ with the blue square and blue cross, respectively.

The sub-optimal performance groups $\mathrm{SPG}^6_{12,VI}$ in Day I and $\mathrm{SPG}^6_{21,VI}$ in Day II are visually shown in Figure 10. Cells belonging exclusively to Day I's $\mathrm{SPG}^6_{12,VI}$ are presented by left-directed orange triangles. Cells with bad performance in $\mathrm{SPG}^6_{21,VI}$ only on Day II are highlighted by right-directed green triangle markers. The first group shows that 98.5% of all cells in $Zone_n$ have issues on Day I and the second group represents 82.6% of these cells have the same issue on Day II. $Zone_n$ is served by MSC_m, hence in both days, the

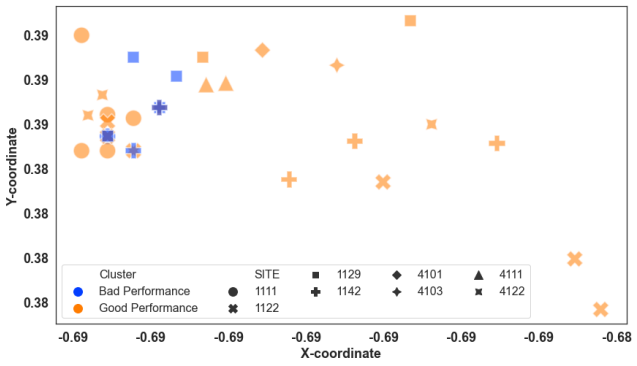


FIGURE 9. All cells in Area A of City C.

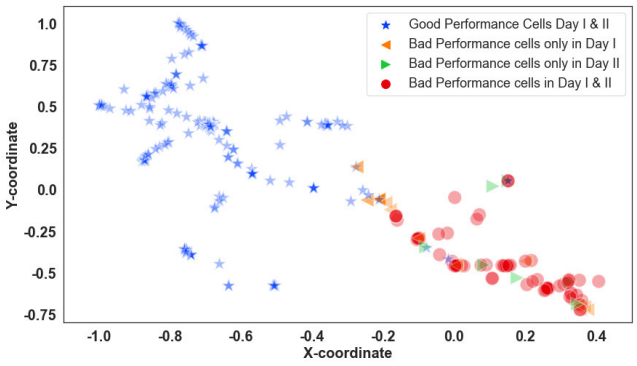


FIGURE 10. All cells from groups $SPG^{6}_{12,VI}$ in Day I and $SPG^{6}_{21,VI}$ in Day II in $Zone_n$ of the network.

$GCD_G = Zone_n$ and $GCD_N = MSC_m$ are visible greatest contributing descriptor for the faults present in these groups. Indeed, the failures, that caused these events would have likely generated an immediate alarm at the Operation and Maintenance Centre (OMC), alarm for link failure, or MSC failure for faults presented in Figures 9 and 10, respectively. These failures would have also triggered a ripple effect of subsequent alarms related to each link or node that is connected and affected. From the data available, it is evident that there is a serious problem with MSC_m which has caused disruptions of service to a whole region of the network ($Zone_n$ shown in Figure 10). However, it is not clear what is triggering the MSC failure; the cause does not seem to be correlated with the traffic volume. An immediate investigation should have been initiated on Day I and a redundancy plan should be put in place to avoid this problem from reoccurring. A domain expert at the OMC would have been able to identify the root-cause among all the beeping alarms and acted accordingly.

On the other hand, cells in $SPG^{6}_{16,VI}$ may not necessarily generate an alarm and would have the typical symptoms of a sleeping cell. Based on the common features listed in Table 7, these 13 cells seem to have suffered from the Case 1 disruption earlier at 12:00 and they never fully recovered. In this case, the CDR-driven NPM is able to identify such cells and hint to the possible cause in a timely fashion. The alternative expert-led detection would have taken days before it was noticed and diagnosed.

B. ROOT-CAUSE ANALYSIS: CASE 2

Only one occurrence of this trend is detected by the algorithm (Part 3), that is $FG_{21,IV}$ on Day I, and it matches the case reported in Section V. The number of cells classified as Type *IV* are 2 at $h = 19$ and increase to 5 cells at $h = 20$ and 21 cells at $h = 21$. A closer examination reveals that most of the cells in this group are in residential areas in various sections of the country. The remaining cells in $FG_{21,IV}$ cover the highways and roads leading out of the large cities. Thus, in this case, there is no network-related GCD_N but the underlying GCD_G is the clutter type *Residential*, although it does not apply to all cells in $FG_{21,IV}$. This case depicts the behavior of cells that suffer from congestion. Since these cells are dispersed and not within a specific area or city, the constricted resources are likely to be related to radio access. This case is only captured on Day I of the data which indicates that the congestion experienced may be related to a special event on this day. Indeed, Day II (which is also a Friday) does not manifest signs of congestion in these cells. In situations where *Case 2* is repeated, the affected cells should be diverted to the planning department with the recommendation of a capacity upgrade.

C. ROOT-CAUSE ANALYSIS: CASE 3

There are many occurrences of a group of cells having the same classification for a single hour, $SPG^{j}_{h,T}$, however, only one stands out on Day I, as the size of the group is larger than three ($MinSize = 3$ see Algorithm 1) and the occurrence is neither covered by *Case 1* nor *Case 2*. This is labeled as $SPG^{30}_{6,VI}$, detected by Algorithm 1 Part 2 and can be seen visually at 6:00 am in Figure 8. Locations of the cells in this group are varied: 13 are in the same city and neighboring areas, 2 are in other cities, and 2 are in open space/road. Thus, Part 4 in Algorithm 1 yields empty GCD_N and GCD_G and triggers the generation of a detailed KPI report. We examine the average normal call duration *ANCD* (for calls that terminate normally), the average dropped call duration *ADCD*, and the drop call rate *DCR* for each of these cells during the hour 6:00 and 7:00, and further group these cells as shown in Table 8.

In order to validate the interpretations in Table 8, a comparison with historical data would be very useful, in particular for *Case 3-A* and *Case 3-B*. If the cause of quality deterioration were interference, then it is likely to be detected at similar times in the past. In this case, an in-depth investigation can be conducted using drive tests or Minimisation Drive Test (MDT as in [35]) to locate the exact location of the quality degradation. Once the planning department is provided with this information, it should be possible to solve the problem with controlled frequency planning and antenna fine-tuning. As for *Case 3-C*, there seems to be a correlation between the duration of calls and the likelihood of a drop, as seen in

TABLE 8. Root-cause analysis pertaining to Case 3.

Trend	No of cells	$ANCD$ (sec)	$ADCD$ (sec)	$\frac{ADCD}{ANCD}$ (%)	DCR (%)	Interpretation
Case 3-A	5	137.39	58.87	51%	16.56%	A possible cause could be very poor quality related to interference
Case 3-B	6	114.23	159.63	140%	13%	A possible cause could be intermittent interference
Case 3-C	6	153.088	1836.76	1200%	18.76%	A possible cause could be related to handover as the dropped calls are 30 minutes long, on average

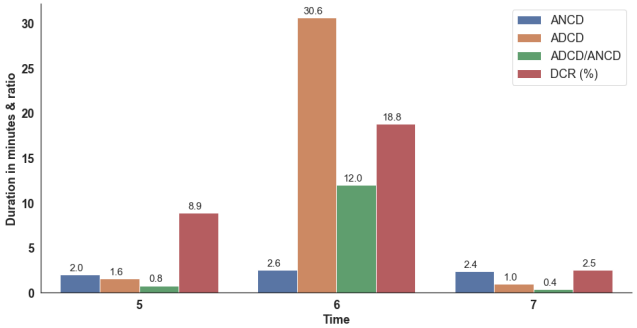


FIGURE 11. This figure represents the average performance of cells in Case 3-C of $SPG^{30}_{6,VI}$. It is clear that around 6 AM, the increase in Dropped Call Rate (DCR) (i.e., the percentage of dropped calls from the total number of calls) is correlated with the Average Dropped Call Duration (ADCD). In other words, calls that go on for longer than half an hour are likely to drop. This may be due to handover issues, but the data available is not sufficient to confirm.

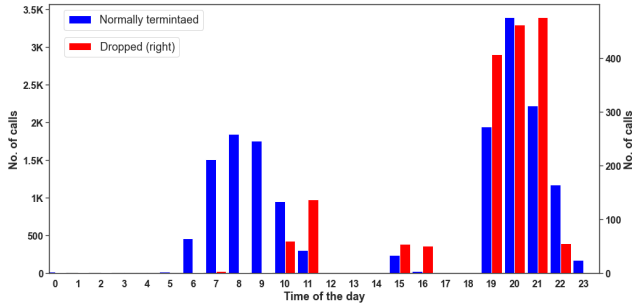


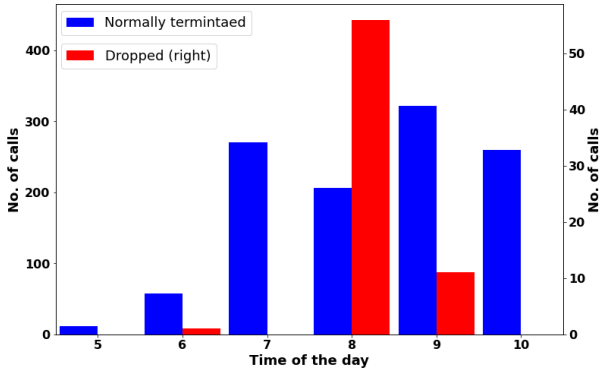
FIGURE 12. This figure represents the aggregate number of normally terminated and dropped calls in Site $S1$ which consists of three cells. The site is seen to go off (not carry any traffic) over two periods during Day I: at 11:00 and at 16:00.

Figure 11. Therefore, a closer examination of the handover metrics and of the timing advance parameters of the concerned cells would shed light on the cause of this problem. Similarly, a drive test along the roads concerned or an MDT would help recreate the problem and capture the detailed steps leading to the failure.

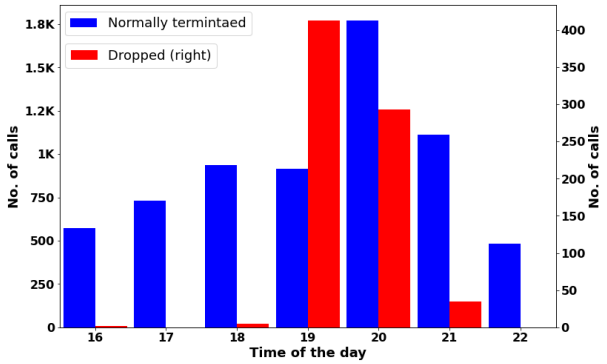
D. ROOT-CAUSE ANALYSIS: CELL-SPECIFIC USING KPIS

The algorithmic approach in 1 identifies a group $SPG^{20}_{15,V}$ which includes 5 cells with pattern VI, IV, V. These cells are captured in multiple groups during Day I including $SPG^{6}_{12,VI}$ discussed in Section VI-A and $SPG^{6}_{16,VI}$ and $SPG^{18}_{17,VI}$. These cells seem to never recover completely. The GCD_G shows that these cells belong to two distinct radio sites that are geographically very distant however they both serve highways in the proximity of small towns. The GCD_N shows that these radio sites do not serve other cells except the affected ones. Moreover, the GCD_N shows that each of these radio sites is associated with an exclusive LAC. It is likely that the dedicated trunk connecting each of these two LACs fails repetitively during the afternoon (possibly caused by the major incident reported in $SPG^{6}_{12,VI}$) and causes the performance degradation. One particular radio site, $S1$, is out of the geographical boundary of the MSC_m for which the failure is reported in Section VI-A on Day I. Nonetheless, all cells in this site seem to go completely off two times during the day and remain so for a duration of three hours each time, as seen in Figure 12. This first occurs between 11:00 and 12:00 and the second time between 16:00 and 17:00. Systematically, two hours prior to each of these episodes, a jump in dropped calls is noticed in all three cells. This indicates that the problem is related to the routing of calls within the LAC, i.e., to the connected trunk. A trunk failure may cause some processes in the site to restart or perhaps the restart action is taken by an operator at the OMC to clear some issues. In both cases, a reset normally results in no traffic on the site for a few minutes. In this particular case, however, it remains off for much longer. The site, thus, goes to sleep until a general reset is applied. These symptoms may be the result of multiple failures in the connected trunk as well as the site's hardware in the base station.

Another peculiar cell-specific behavior is seen on Day II and consists of a group of cells that remain as Type VI for three consecutive cell-instances. The first occurs between 6:00 and 8:00. It includes three cells of the same Site $S2$. The corresponding generated KPI report is presented in Figure 13(i) which shows the number of normally terminated and dropped calls for the three cells in $S2$. The values presented are the aggregated values for the three cells in that site. From Figure 13(i), it can be seen that the overall traffic (calls entertained) drops on-site $S2$ at 8:00am and the number of dropped calls increase. But it can be seen as a temporary issue as the traffic gets normalized afterward, indicating that it can be a temporary hardware failure issue. The second such cells specific behavior occurs between 18:00 and 20:00 for the two cells of the same Site $S3$. The corresponding generated KPI report is shown in Figure 13(ii). Here it can be seen that the traffic load increases from 19:00 to 20:00 and the call drop rate is also high, it can likely be a traffic congestion case.



(i): Aggregate number of normally terminated and dropped calls in Site $S2$ which consists of three cells. The traffic carried by this site drops significantly at 8:00 and picks up again indicating a temporary hardware failure affecting all three cells.



(ii): Aggregate number of normally terminated and dropped calls in Site $S3$ which consists of two cells. There seems to be a correlation between the peak in traffic at 20:00 and the increase in dropped calls around this time.

FIGURE 13. Cell-specific analysis of *S2* and *S3* on Day II.

VII. ANALYSIS OF PROPOSED AI-DRIVEN CDR-BASED ROOT-CAUSE ANALYSIS

We have presented a framework that analyses streaming CDR records to identify cells with sub-optimal performance and, where possible, locate the root-cause of the problem. This framework is validated using a real dataset and the obtained points of interest are compared to the visual inspection conducted by a domain expert in Section V. The visual inspection is carried off-line by the end of the day after all data has become available. It allows spotting patterns of sub-optimal performance, as shown in Figure 8 and focusing the attention of the domain expert to investigate and interpret the detected problems after 24 hours. In contrast, the proposed framework depicted in Algorithm 1 identifies each of these detected problems within a three-hours window (a design parameter that may be reduced) and offers an automated root-cause analysis that matched the expert's conclusion. Thus, the first gain of the proposed framework is the online capability joined with the overarching view of the network from the CDR perspective that allows the immediate location of the root-cause. Furthermore, the framework ensures that all cells with sub-optimal performance are detected, even those which may not be seen by visual inspection such as Case 1 occurrence on Day I at 10 AM ($SPG^{6}_{10,VI}$). Thus, the second gain of the proposed framework is the guaranteed detection of all problems and the identification of their root-cause. The algorithm is designed to detect groups of cells with sub-optimal performance that are greater than or equal to a threshold (3 in our implementation) and that match three predefined types of sub-optimal performance ($T \in \{IV, V, VI\}$ shown in Figure 7). With these settings, a network expert is asked to examine every hour a maximum of 30 possible groups of cells with sub-optimal performance that have been pre-processed and their GCDs are identified. In practice, there is never a time where more than three groups are identified within a single hour, based on our dataset. As such, the third gain of the proposed framework is the automated processing of online data that reduces the information that requires an expert's attention from the total number of cells in the network to a maximum of 30 cases.

The framework can be tuned and adjusted to different objectives by tuning the unit of time for CDR aggregation, sliding window size, the threshold for the size of sub-performing groups of cells, the network and geography features, and the detailed KPI report. In this work, we demonstrate the advantages of automating the cell performance classification and root-cause analysis in precision, time-efficiency, and usage of domain expert's resources.

VIII. CONCLUSION

In this paper, we have proposed an AI framework for the automation of the Network Performance Management (NPM) process towards a Zero-touch network and Service Management (ZSM). The proposed framework not only identifies and classifies cells based on their performance computed through hourly aggregated CDR data but also identifies performance-related issues present in the cells of the network. In addition, the distinctive feature of this framework, which is missing in state of the art, is CDR-driven algorithm that automates the diagnosis of the root-cause of any performance-related issues identified. As part of the AI framework, our SVM-based model classifies cells with poor performance into respective types of sub-optimal performance groups with an accuracy of 97.69%. Besides, the diagnosis algorithm in the proposed framework offers an accurate root-cause analysis that is verified by domain experts using visual graphs and statistical summaries. Furthermore, it guarantees that no network fault goes unnoticed. Our proposed framework significantly reduces an expert's job of detecting and analyzing networks faults and subsequently the network performance management cost. For example, in this study, the manual inspection of the performance of network cells is reduced to a maximum of thirty cases as opposed to analyzing 759 cells, at any time. The proposed AI framework is a stepping stone towards the realization of ZSM networks where the NPM is AI-driven and used for the automation of optimization process.

ACKNOWLEDGMENT

The statements made herein are solely the responsibility of the authors.

REFERENCES

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

[2] C. Benzaid and T. Taleb, "AI-driven zero touch network and service management in 5G and beyond: Challenges and research directions," *IEEE Netw.*, vol. 34, no. 2, pp. 186–194, Mar. 2020.

[3] *Technical Specification Group Services and System Aspects; Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services; Release 17*, document Rec. TS 23.288, 3GPP, Version 17.2.0, Sep. 2021. [Online]. Available: https://portal.3gpp.org/

[4] B. Balasubramanian, E. S. Daniels, M. Hiltunen, R. Jana, K. Joshi, R. Sivaraj, T. X. Tran, and C. Wang, "RIC: A RAN intelligent controller platform for AI-enabled cellular networks," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 7–17, Mar. 2021.

[5] A. Oi, R. Sato, Y. Suto, K. Sakata, M. Nakajima, and T. Furukawa, "A study on automation of network maintenance in telecom carriers for zero-touch operations," in *Proc. 21st Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2020, pp. 1–6.

[6] ETSI. *Zero Touch Network & Service Management (ZSM)*. Accessed: Nov. 24, 2021. [Online]. Available: https://www.etsi.org/technologies/zero-touch-network-service-management?highlight=YTozOntpOjA7czoxOiJuZnYiO2k6MTtzOjQ6IiduZnYiO2k6MjtzOjU6Im5mdidzIjt9

[7] L. Jorguseski, A. Pais, F. Gunnarsson, A. Centonza, and C. Willcock, "Self-organizing networks in 3GPP: Standardization and future trends," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 28–34, Dec. 2014.

[8] Y. Ouyang, Z. Li, L. Su, W. Lu, and Z. Lin, "Application behaviors driven self-organizing network (SON) for 4G LTE networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 3–14, Jan. 2020.

[9] M. S. Parwez, D. Rawat, and M. Garuba, "Big data analytics for user-activity analysis and user-anomaly detection in mobile wireless network," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2058–2065, Aug. 2017.

[10] R. Sharifi, M. M. Majdabadi, and V. T. Vakili, "Mobile user-activity prediction utilizing LSTM recurrent neural network," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process. (PACRIM)*, Aug. 2019, pp. 1–7.

[11] S. Jaffry, S. T. Shah, and S. F. Hasan, "Data-driven semi-supervised anomaly detection using real-world call data record," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Apr. 2020, pp. 1–6.

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

[13] B. Hussain, Q. Du, and P. Ren, "Deep learning-based big data-assisted anomaly detection in cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[14] A. Tauriainen, "Can you hear me now? A call detail record based end-to-end diagnostics system for mobile networks," in *Proc. 15th Int. Conf. Netw. Service Manage. (CNSM)*, Oct. 2019, pp. 1–7.

[15] F. Rezazadeh, H. Chergui, L. Christofi, and C. Verikoukis, "Actor-critic-based learning for zero-touch joint resource and energy control in network slicing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

[16] V. Balasubramanian, M. Aloqaily, and M. Reisslein, "FedCo: A federated learning controller for content management in multi-party edge systems," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–9.

[17] A. Rizwan, J. P. B. Nadas, M. A. Imran, and M. Jaber, "Performance based cells classification in cellular network using CDR data," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[18] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "On the joint optimisation of radio access and backhaul networks," in *Proc. Int. Conf. Innov. Electr. Eng. Comput. Technol. (ICIEECT)*, Apr. 2017, pp. 1–5.

[19] M. Jaber, O. Onireti, and M. A. Imran, "Backhaul-aware and context-aware user-cell association approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.

[20] D. Mulvey, C. H. Foh, M. A. Imran, and R. Tafazolli, "Cell fault management using machine learning techniques," *IEEE Access*, vol. 7, pp. 124514–124539, 2019.

[21] G. Cao, Z. Lu, X. Wen, T. Lei, and Z. Hu, "AIF: An artificial intelligence framework for smart wireless network management," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 400–403, Feb. 2018.

[22] B. Ma, W. Guo, and J. Zhang, "A survey of online data-driven proactive 5G network optimisation using machine learning," *IEEE Access*, vol. 8, pp. 35606–35637, 2020.

[23] 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.

[24] W. Guo, "Explainable artificial intelligence for 6G: Improving trust between human and machine," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 39–45, Jun. 2020.

[25] G. Carrozzo, M. S. Siddiqui, A. Betzler, J. Bonnet, G. M. Perez, A. Ramos, and T. Subramanya, "AI-driven zero-touch operations, security and trust in multi-operator 5G networks: A conceptual architecture," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, 2020, pp. 254–258.

[26] I. Sanchez-Navarro, P. Salva-Garcia, Q. Wang, and J. M. A. Calero, "New immersive interface for zero-touch management in 5G networks," in *Proc. IEEE 3rd 5G World Forum (5GWF)*, Sep. 2020, pp. 145–150.

[27] D. Bega, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "DeepCog: Optimizing resource provisioning in network slicing with AI-based capacity forecasting," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 361–376, Feb. 2020.

[28] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, "Network slicing meets artificial intelligence: An AI-based framework for slice management," *IEEE Commun. Mag.*, vol. 58, no. 6, pp. 32–38, Jun. 2020.

[29] U. R. Mughal, M. Ahmed Khan, A. Beg, and G. Q. Mughal, "AI enabled resource allocation in future mobile networks," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2020, pp. 1–6.

[30] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[31] M. Ozturk, M. Jaber, and M. A. Imran, "Energy-aware smart connectivity for IoT networks: Enabling smart ports," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–11, Jun. 2018.

[32] C. Swenson, C. Adams, A. Whitledge, and S. Shenoi, "On the legality of analyzing telephone call records," in *Advances in Digital Forensics III* (The International Federation for Information Processing), vol. 242, P. Craiger and S. Shenoi, Eds. New York, NY, USA: Springer, 2007, doi: 10.1007/978-0-387-73742-3_2.

[33] *Telecommunication Management; Charging Management; Charging Data Record (CDR) File Format and Transfer*, document Rec. TS 32.297, Version 16.0.0, 3GPP, Sep. 2019. [Online]. Available: https://portal.3gpp.org/

[34] *Technical Specification Group Core Network and Terminals; Numbering, Addressing and Identification*, document Rec. TS 23.003, Version 16.3.0, 3GPP, Jun. 2020. [Online]. Available: https://portal.3gpp.org/

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

ALI RIZWAN received the bachelor's degree in applied and theoretical math and the M.B.A.-I.T. degree from Bahauddin Zakariya University, Pakistan, in 2006 and 2008, respectively, the M.Sc. degree in big data science from the Queen Mary University of London, London, U.K., in 2016, and the Ph.D. degree from the University of Glasgow, Glasgow, U.K., in 2021. He is currently working as an AI Research Scientist at the Qatar Mobility Innovations Center (QMIC), Qatar University. His research interest includes artificial intelligence for self-organizing wireless networks.

MONA JABER (Senior Member, IEEE) received the B.E. degree in computer and communications engineering and the M.E. degree in electrical and computer engineering from the American University of Beirut, Lebanon, in 1996 and 2014, respectively, and the Ph.D. degree from the 5G Innovation Centre, University of Surrey, in 2017. Her Ph.D. research was on 5G backhaul innovations. She was a Telecommunication Consultant in various international firms with a focus on the radio design of cellular networks, including GSM, GPRS, 3G, and 4G. She has led the IoT Research Group, Fujitsu Laboratories, Europe, from 2017 to 2019, where she researched the IoT-driven solutions for the automotive industry. She is currently a Lecturer in IoT with the School of Electronic Engineering and Computer Science, Queen Mary University of London. Her research interests include zero-touch networks, the IoT-driven digital twins, and AI/ML innovation for the IoT data mining in the context of smart mobility.

FETHI FILALI (Senior Member, IEEE) received the Ph.D. degree in computer science and the Habilitation degree from the University of Nice Sophia Antipolis, France, in 2002 and 2008, respectively. He was with the Mobile Communications Department, EURECOM, France, as an Assistant Professor and then an Associate Professor for eight years. He is currently the Director of technology and research with the Qatar Mobility Innovations Center (QMIC), Qatar University. He is also leading the technology development of QMIC's solutions in the areas of smart cities, the Internet of Things, intelligent transportation systems, and connected and automated vehicles solutions. He has invented technologies and developed algorithms that have been shipped in many QMIC products, including Masarak, Labeeb, and WaveTraf, creating commercial impact in the order of millions of dollars. His research grants include 15 competitive awards from several funding agencies, including the European Commission, the French National Research Agency, and the Qatar National Research Fund. He was the Ph.D. Director for over ten Ph.D. students in the areas of intelligent transportation, wireless sensor and mesh networks, vehicular communications, big data analytics, the Internet of Things, and mobility management. He has coauthored over 130 research papers in international peer-reviewed conferences and journals. He holds over ten patent applications.

ALI IMRAN (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2005, and the M.Sc. degree (Hons.) in mobile and satellite communications and the Ph.D. degree from the University of Surrey, Guildford, U.K., in 2007 and 2011, respectively. He is currently a Presidential Associate Professor in ECE and the Founding Director with the Artificial Intelligence (AI) for Networks (AI4Networks) Research Center and the TurboRAN Testbed for 5G and Beyond, The University of Oklahoma. His research interests include AI and its applications in wireless networks and healthcare. His work on these topics has resulted in several patents and over 100 peer-reviewed articles, including some of the most influential papers in domain of wireless network automation. On these topics, he has led numerous multinational projects, given invited talks/keynotes and tutorials at international forums and advised major public and private stakeholders and cofounded multiple start-ups. He is an Associate Fellow of the Higher Education Academy, U.K. He is also a member of the Advisory Board to the Special Technical Community on Big Data and the IEEE Computer Society.

ADNAN ABU-DAYYA (Senior Member, IEEE) received the Ph.D. degree in digital mobile communications (electrical engineering) from Queen's University, Kingston, ON, Canada, in 1992. He worked with AT&T Wireless, Seattle, USA, for ten years, where he served in a number of senior management positions covering product innovations, emerging technologies, systems engineering, product realization, and intellectual property management. He is the Executive Director with the Qatar Mobility Innovations Center, Doha, Qatar. He led the establishment of the Qatar Mobility Innovations Center, in 2009. It is one of the first independent innovations institution in the Middle East focused on translating research and development and technology innovations into scalable digital platforms and solutions in the field of intelligent mobility and smart cities. He worked as a Senior Manager with the Advanced Technology Group, Nortel Networks, Canada, and as a Senior Consultant with the Communications Research Centre, Ottawa, ON, Canada. He has ten issued patents, and about 100 refereed publications. He serves as the Chairperson of the Advisory Board of the Electrical and Computer Engineering Department, Texas A&M University, Qatar, where he is a member of the Steering Committee of the Smart Grid Research Center.