Intelligent Network Data Analytics Function in 5G Cellular Networks using Machine Learning

Salih Sevgican, Meriç Turan, Kerim Gökarslan, H. Birkan Yilmaz, and Tuna Tugcu

Abstract: **5G cellular networks come with many new features compared to the legacy cellular networks, such as network data analytics function (NWDAF), which enables the network operators to either implement their own machine learning (ML) based data analytics methodologies or integrate third-party solutions to their networks. In this paper, the structure and the protocols of NWDAF that are defined in the 3rd Generation Partnership Project (3GPP) standard documents are first described. Then, cell-based synthetic data set for 5G networks based on the fields defined by the 3GPP specifications is generated. Further, some anomalies are added to this data set (e.g., suddenly increasing traffic in a particular cell), and then these anomalies within each cell, subscriber category, and user equipment are classified. Afterward, three ML models, namely, linear regression, long-short term memory, and recursive neural networks are implemented to study behaviour information estimation (e.g., anomalies in the network traffic) and network load prediction capabilities of NWDAF. For the prediction of network load, three different models are used to minimize the mean absolute error, which is calculated by subtracting the actual generated data from the model prediction value. For the classification of anomalies, two ML models are used to increase the area under the receiver operating characteristics curve, namely, logistic regression and extreme gradient boosting. According to the simulation results, neural network algorithms outperform linear regression in network load prediction, whereas the tree-based gradient boosting algorithm outperforms logistic regression in anomaly detection. These estimations are expected to increase the performance of the 5G network through NWDAF.**

Index Terms: **Handover, machine learning, NWDAF, 5G networks.**

I. INTRODUCTION

TECHNOLOGICAL developments in wireless cellular networking are expected to increase the number of users and end-points tremendously, resulting in the creation of very complex and busy networks. The standardization of the fourth-generation (4G) cellular systems by the 3rd Generation Partnership Project (3GPP) enabled users to reach a couple of hundred Mbps, thus allowing users to access the applications requiring high data rates such as high-definition TV [1], [2]. Yet, 4G is incapable of addressing exponentially increasing user demands. In particular, Cisco's Visual Networking Index forecasts that the number of mobile-connected devices will reach to 12.3 billion, and the average smartphones will generate 132 GB traffic annually. [3]. Moreover, the rapid emergence of applications requiring machine-to-machine (M2M) type communication (e.g., Internet of things (IoT)) has brought new requirements that are not addressed by the previous cellular technologies designed for human-to-human (H2H) communications [4]. As such, the fifth-generation (5G) cellular networks have emerged in the last decade. 3GPP first released the specifications for the non-standalone (NSA) mode of 5G access that is based on the existing 4G infrastructure, thus supporting interoperability between the existing cellular technologies.

Researchers anticipate that 5G will be a major paradigm shift rather than an incremental advancement on 4G [5]. To this end, 3GPP published another series of specifications for standalone (SA) 5G that introduces a new cloud-native 5G core independent from the 4G cellular infrastructure. 5G SA brings much simplified Radio Access Network (RAN) and device architecture as it is a targeted 5G architecture option, which facilitates a wider range of use cases for new device types, and device communication patterns such as M2M communication [6]. More specifically, 5G SA introduces an evolved packet core considering the newer use cases including IoT [6]. Therefore, some of the already existing network functions (NF) in the service-based architecture (SBA) of legacy cellular networks (i.e., 4G or below) are replaced with new NFs in 5G SBA. For example, instead of policy and charging rules function (PCRF) in 4G, 5G has a policy control function (PCF). Similarly, instead of charging data record (CDR), 5G has charging function (CHF).

One such function is related to the data analytics. Data analytics has become vital in 5G SA as it is designed to support data rates that can reach a gigabit. Network data analytics function (NWDAF) is one of the newly proposed data analytics functions for 5G networks, and it provides network analysis to other NFs [7]–[10]. To do so, NWDAF can use any machine learning (ML) or artificial intelligence (AI) algorithm based on the requirements (e.g., time constraints) of the consuming NF [11]. According to the 3GPP consortium, NWDAF is expected to have several capabilities, where we investigate three of them deeply in this paper. We focus on the prediction of: (1) Abnormal behaviour information for a group of user equipment (UE), (2) expected behaviour information for a group of UE, (3) network load performance in an area of interest. The reason for elaborating on the specific capabilities of NWDAF is to keep the focus of the study clear. Considering the sake of the paper, we pick three capabilities of NWDAF for investigation. We believe that the selected capabilities are crucial for network

Manuscript received December 19, 2019; revised June 3, 2020; approved for publication by Changhee Joo, Division III Editor, June 3, 2020.

This work is supported by the Turkish Directorate of Strategy and Budget under the TAM Project number 2007K12-873.

S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz, and T. Tugcu are with the Department of Computer Engineering, NETLAB, Bogazici University, Istanbul 34342, Turkey, email: {salih.sevgican, meric.turan, kerim.gokarslan, birkan.yilmaz, tugcu}@boun.edu.tr.

M. Turan is the corresponding author.

Digital Object Identifier: 10.1109/JCN.2020.000019

1229-2370/19/$10.00 © 2020 KICS

sustainability and quality of service (QoS).

Even before 3GPP introduced NWDAF for 5G cellular networks, AI/ML models were frequently used in wireless networking as well as in many other areas. Yet, with the need for ultra-reliable and low latency communication (URLLC) [12], [13] and unprecedented data traffic that increases exponentially, the use of AI/ML models in cellular networks has become a serious necessity. 3GPP could not overlook this requirement, and consequently introduced NWDAF to fulfill this requirement [8].

The first issue that comes to mind is the data set while studying ML topics. To be fair, the best option is to use a data set gathered from an actual network setup. An alternative approach could be using Generative Adversarial Networks (GAN) [14]. Unfortunately, we could not come up with the former one for the NWDAF scenarios studied in this paper due to the lack in the literature, which is elaborated in Section II. Similarly, one requirement to use the latter one is to train GAN by using a sample from a real data set, which does not exist as stated. In that manner, we firstly generate a publicly available 5G data set [15] inspiring from 3GPP specifications for 5G networks to fill this gap in the literature [16]–[19]. The generated 5G data set includes a topology with a fixed number of cells for a fixed number of subscriber categories, where different types of devices that have different traffic patterns (e.g., cell phone, vehicle) can connect to the network. We model each cell using a set of features that we retrieve from other NFs: Bytes transmitted during the monitoring time, list of categories associated with the subscriber, personal equipment ID, and network area information (i.e., RRU cell ID). Also, in order to make our synthetic 5G data set more realistic, we include anomalies such as unexpectedly increasing data traffic through a particular RRU cell. Then, we present a novel system to perform network data analytics for 5G networks using state-of-the-art ML models in two parts. In the first part, we perform network load performance prediction by using linear regression, long short term memory (LSTM), and recursive neural network (RNN) models. Then, in the second part, using the anomalies integrated into the 5G data set, we perform classification on the current status of a network cell in order to detect the existing anomalies by using logistic regression and a widely used tree-based ML algorithm, named extreme gradient boosting (XGBoost) [20]. We, then, train different ML models using thoughtfully labeled data that we have generated according to the 3GPP specifications. Thus, the main contributions of this paper are threefold:

- We generate a cell-based synthetic data set for 5G networks using the fields defined by the 5G standard documents.
- We introduce ML approaches, which are compatible with the NWDAF system, for different 5G cellular network topologies. While one of the subsystems estimates the load performance in the network, the other classifies network status to predict whether there is an anomaly in the network.
- We present simulation results of the proposed system with a fixed topology and compare the merits of different ML approaches for 5G network data analytics.

This paper is organized as follows. Section II investigates the related works in the literature. Section III presents the network data analytics function. Section IV describes our system model and topology we used in our simulations. Section V introduces our method for data generation to evaluate ML-based 5G network data analytic systems. Section VI discusses the set of features that we pick for 5G cells and how we use them with several state-of-the-art ML methods. Section VII gives the simulation results based on different scenarios using the proposed system. Finally, Section VIII concludes the paper and discusses several future directions for ML-based 5G data analytics approaches.

Throughout the paper, we use several acronyms. To increase the readability of the paper, we share these acronyms in Table 1.

Table 1. List of acronyms.

3GPP	3rd generation partnership project
AF	Application function
AI	Artificial intelligence
AMF	Access and mobility management
ANN	Artificial neural network
AUC	Area under curve
AUC-ROC	Area under receiver operating characteristics
CDR	Charging data record
CHF	Charging function
DL	Deep learning
GAN	Generative adversarial networks
H2H	Human-to-human
IoT	Internet of things
LogReg	Logistic regression
LR	Linear regression
LSTM	Long-short term memory
LTE	Long term evolution
M2M	Machine-to-machine
MAC	Mobile access control
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MIMO	Multiple input multiple output
ML	Machine learning
mmWave	Millimeter wave
NEF	Network exposure function
NF	Network function
NSA	Non standalone
NSSF	Network slice selection function
NR	New radio
NWDAF	Network data analytics function
OAM	Operation administration and maintenance
PDCP	Packet data convergence protocol
PCF	Policy control function
PCRF	Policy and charging rules function
QoS	Quality of service
RAN	Radio access network
RNN	Recursive neural network
ROC	Receiver operation characteristics
RRC	Radio resource control
RRU	Remote radio unit
SA	Standalone
SBA	Service-based architecture
SBI	Service-based interface
SMF	Session management function
SubsCat	Subscriber category
TCP/IP	Transmission control protocol internet protocol
UDM	Unified data management
UDR	Unified data repository
UE	User equipment
URLLC	Ultra reliable and low latency communication
WiMAX	Worldwide interoperability for microwave access
xG	xth-generation
XGBoost	eXtreme gradient boosting

II. RELATED WORK

Due to the freshness of NWDAF in 5G cellular networks, the literature about this topic is not as comprehensive as it should be, and this paper institutes a novel work that considers NWDAF compatible implementation. In [21], the authors introduce mobility and network slicing capabilities of NWDAF. In [11], the capabilities of NWDAF related to the network slicing are investigated. In terms of traffic data for 5G, unfortunately, there is no comprehensive selection of data sets to be utilized for the NWDAF scenarios studied in this paper. In [22], a ray-tracing simulation is performed in order to produce data for 5G multiple-input and multiple-output (MIMO) study. In [23], researchers create a data set by monitoring one eNodeB and one user equipment. The data set includes Packet Data Convergence Protocol (PDCP), Radio Resource Control (RRC), and Mobile Access Control (MAC) data. Considering the coverage and singularity of the monitoring, this data set also is not suitable for ML purposes. So, to the best of our knowledge, there is no user traffic data gathered from gNodeB based on 5G SA implementation and suitable for NWDAF scenarios of this paper in the literature.

Intelligent cellular networks based on the state-of-the-art AI/ML techniques, on the other hand, have been studied thoroughly in the last decade. In [24], the authors address the significance and the necessity of using AI for the next-generation cellular networks (i.e., 5G and 6G), and they specify some of the challenges and the roadmap. In [25], Jiang *et al.* argue that the next-generation wireless technologies, including 5G, would require the support of extremely high data rates; thus, decision making in the new radio systems can benefit from ML techniques. They propose different ML techniques for different tasks in 5G networks including supervised learning-based methods for MIMO channel controlling, unsupervised learning for anomaly detection, and reinforcement learning for decision making under unknown network conditions. Yet, they do not investigate how ML would be used for network analytics in 5G. Casellas *et al.* also emphasize the necessity of enabling AI/ML techniques to control, manage, and orchestrate components of 5G networks without mentioning about NWDAF [26]. In [27], ML is proposed as a way to manage self-organizing 5G networks. Self-organization in cellular networks comes into prominence as 5G technology leans towards millimeter wave (mmWave) radio models that would require massive network densification. To this end, the authors of [27] give a detailed network management system using different ML techniques together. In [28] and [29], the authors investigate the existing studies using ANN/ML techniques related to wireless networks. Fang *et al.* propose to use ML techniques for 5G and beyond wireless networks due to the security concerns, especially considering authentication issues [30]. In [31], the authors argue that due to the unprecedented user demands on mobile and wireless networking, real-time extraction of fine-grained analytics and agile management of network resources become more and more critical. The authors investigate deep learning studies in the literature since they find it as a solution to meet the unprecedented user demands. There are also various studies for more specific tasks based on 5G cellular networks. [32]–[34] investigate the use cases of ML in vehicular networks based on 5G networks. Moreover, [35]–[37] study ML in mmWave massive MIMO to control the radio including beamforming.

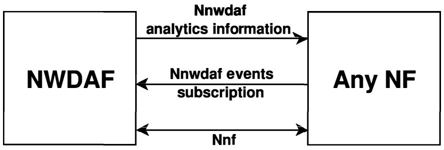


Fig. 1. Data collection and network data analytics exposure architecture.

In summary, several works in the literature study ML, AI, and deep learning (DL) concepts for 5G cellular networks with or without considering NWDAF. Yet, these works are currently immature. Moreover, even some sections in the 3GPP specifications regarding NWDAF [7]–[10] are left blank at the present due to the novelty of NWDAF. Furthermore, the existing data sets are not meeting the NWDAF scenario requirements of this paper. We, therefore, generate our synthetic data set based on 3GPP specifications, as we discuss in Section V.

III. NETWORK DATA ANALYTICS FUNCTION

NWDAF is a newly defined data analytics function in 5G cellular networks that provides network analysis upon request from other NFs [8]. As its data source, NWDAF can use any other NF. Therefore, there is a two-way relation between NWDAF and NFs as depicted in Fig. 1. Note that Nnwdaf represents the service-based interface of NWDAF, and Nnf represents the service-based interface of any NF (e.g., Npcf represents the service-based interface of PCF) [8]. As seen in the figure, NWDAF can either provide network analysis data to other NFs (i.e., analytics information) or NFs can request subscription from NWDAF for data delivery (i.e., events subscription) by using Nnwdaf interface. Also, NWDAF fetches data from other NFs using Nnf interface.

There are two Nnwdaf services, namely, events subscription (i.e., analytics subscription), and analytics information [7]–[10]. Nnwdaf events subscription service allows NF consumers to subscribe to and unsubscribe from different analytics events, and notifies NF consumers with a corresponding subscription about observed events. On the other hand, Nnwdaf analytics information enables NF consumers to request and get a different type of analytic event information from NWDAF. Therefore, one can consider events subscription service as more of a charging service, whereas analytics information service is the one that requires the AI/ML techniques to apply. Considering these, our main focus in this paper is on Nnwdaf analytics information service. With Nnwdaf analytics information, the following events can be observed:

- Abnormal behaviour information for a group of UE or a specific UE,
- Expected behaviour information for a group of UE or a specific UE,
- Network load performance in an area of interest,
- Load level of network slice instance,
- NF load analytics information for a specific NF,
- Communication pattern for a specific UE,
- Congestion information of user data in a specific location,

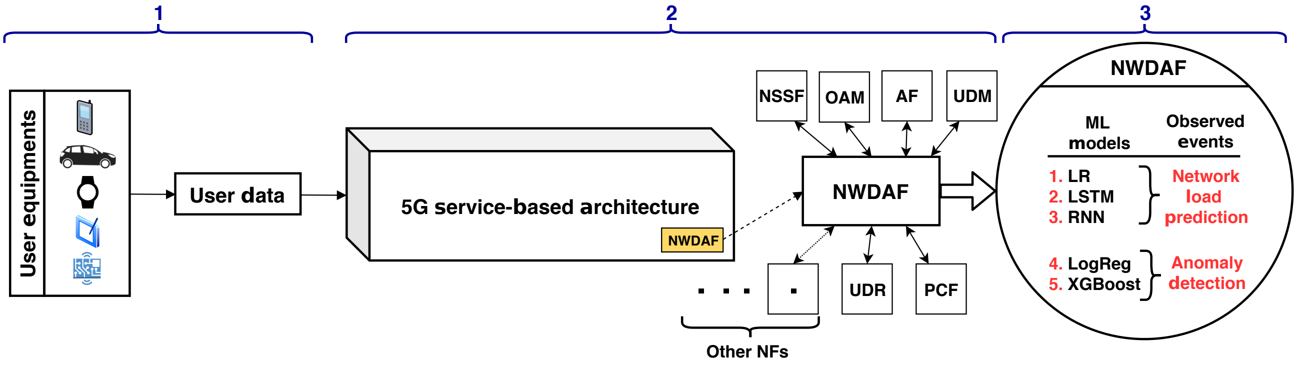


Fig. 2. The high-level workflow of the proposed system. The workflow is threefold: (1) data is generated by UE and sent to 5G SBA, (2) NWDAF gathers this information (i.e., the ones sent from UE) from related NFs, and (3) ML functions within NWDAF produce network analytics information.

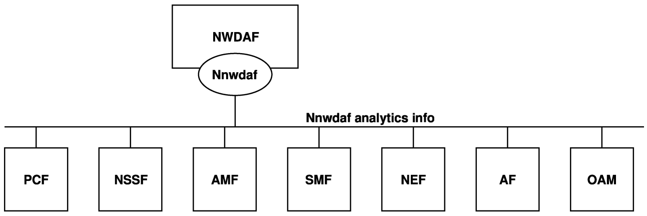


Fig. 3. Service-based architecture of the Nnwdaf analytics info service [9].

- Mobility related information for a group of UE or a specific UE,
- QoS change statistics and potential QoS change in a certain area,
- Service experience for an application or for a network slice.

In this work, we focus on the first three events and classify the behaviour information for a group of UE while estimating the network load performance. Commonly known consumers of the NWDAF analytics information service can be seen in Fig. 3. Any NF in Fig. 3 can be the consumer of the events NWDAF analyzes upon subscription.

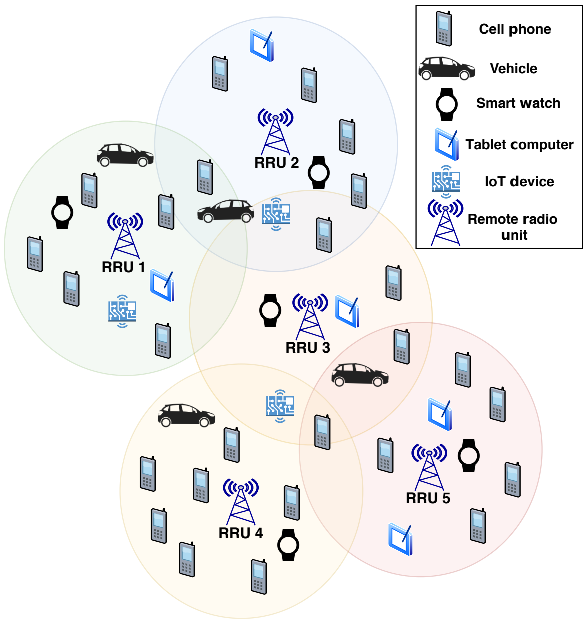


Fig. 4. Sample network topology representation.

IV. SYSTEM MODEL

In this section, we first introduce the high-level workflow of our system. We then describe the type of network topology and the traffic conditions that we consider in this paper. We study a system workflow that consists of UE, user data, 5G SBA, ML models, NWDAF, and other NFs.

A. Workflow

As depicted in Fig. 2, data obtained from UE is transferred into the 5G SBA, where NWDAF and other NFs stand. NWDAF is connected to other NFs via the service-based interface (SBI), and NWDAF and other NFs mutually make data transfer between each other. NWDAF then gathers the information from different NFs and fits several ML models to both predict the network load performance and detect the network load anomalies. Given a fixed topology, our system uses labeled data to train; therefore, it picks the finest ML model depending on the characteristic of the topology.

B. Topology

We use a fixed cellular topology, which consists of a fixed set of RRU cells, a fixed set of subscriber categories, and a fixed set of personal equipment (i.e., device) types. For the sake of simplicity, even though our system model can support topologies that consist of a large number of RRU cells, subscriber categories, and personal equipment types, we consider a network topology that consists of five RRU cells in our simulations. In each of these cells, there are three subscriber categories, where these subscriber categories represent platinum, gold, and silver subscriptions. The logic behind these three subscriptions is to make the generated data more realistic because mobile service providers sell these kinds of subscriptions in the real world. Also, within each subscriber category, there are five different types of personal equipment (i.e., user equipment), namely, IoT device, vehicle, cell phone, smartwatch, and tablet computer. A sample representation of our network topology can be seen in Fig. 4.

Table 2. Mean handover ratios per hour.

Time of day	IoT device	Vehicle	Cell phone	Smart watch	Tablet computer
00:00–06:00	1%	10%	2.5%	2.5%	1%
06:00–07:00	1%	18%	4.5%	4.5%	1.8%
07:00–09:30	1%	12%	3%	3%	1%
09:30–11:00	1%	14%	3.5%	3.5%	1.2%
11:00–13:00	1%	16%	4%	4%	1.5%
13:00–16:00	1%	14%	3.5%	3.5%	1.2%
16:00–20:00	1%	12%	3%	3%	1%
20:00–22:00	1%	18%	4.5%	4.5%	1.8%
22:00–00:00	1%	10%	2.5%	2.5%	1%

C. *Traffic*

According to the proposed model, each personal equipment within each subscriber category and RRU cell includes a predetermined amount of traffic load at the beginning of the simulation. Therefore, we can say that network traffic is saturated from the beginning to the end of the simulation. Then, for every simulation time step (Δt), some portion of the load handovers from one cell (i.e., source) to another cell (i.e., target), which is adjacent to the source cell. The handover process also occurs based on predetermined ratios. To make the traffic more realistic, we change mean handover ratios according to the time of the day. Between 22:00–06:00, since people are expected to move less, the mean handover ratio is also expected to be smaller. Between 06:00–07:00 and 20:00–22:00, since roads will most likely be traffic-free, people are expected to move faster from one place to another, and the mean handover ratio is expected to be higher. Between 11:00–13:00, traffic will be slightly more, and people will move a bit slower compared to 06:00–07:00 and 20:00–22:00, and consequently the mean handover ratio is expected to be smaller. Between 09:30–11:00 and 16:00–20:00, the mean handover ratio is expected to be smaller than 11:00–13:00 due to the increasing traffic. Since 07:00–09:30 and 16:00–20:00 are rush hours, the mean handover ratio is expected to be the smallest excluding the night time. Also, for IoT devices, no major difference for the mean handover ratios is expected for the time of the day because these devices are not expected to move as much as other personal equipment we consider due to their nature. On the other hand, mean handover ratios are higher for vehicles due to their mobility. We give the detailed mean handover ratios in Table 2. Also, note that the handover ratio values in Table 2 are mean values, which imply that there are also variance values. With these carefully tuned statistical parameters by canonical approximations, we aim to achieve a real-like user traffic. The details of the distribution used for the handover ratio determination process is explained in Section V.

V. DATA GENERATION

There are different types of AI/ML models throughout computer science history. Among these many AI/ML models, the algorithms behind these models work differently. In general, one can say that ML algorithms can be categorized under three different parts, namely, supervised, unsupervised, and reinforcement learning. Since we consider supervised ML algorithms in this paper, labeled data becomes crucial in this context. As a result, the data generation part becomes an important aspect of this paper. Considering all these, we generate a labeled data set for 5G cellular networks. While selecting the fields of the proposed data set, we are inspired by the 5G specifications published by the 3GPP consortium [16]–[19]. The selected fields are as follows:

- **Data rate:** Amount of transmitted data in bytes for a certain period of time.
- **Network area information:** Cell information of a group of UE that is connected to.
- **Subscription categories:** The policy of a group of UE is subscribing.
- **Personal equipment ID:** Device type information of a UE (e.g., cell phone, smart watch).

While generating the labeled data set, we come up with the following predefined parameters as the input fields of the data generation simulation, which are

- RRU cell number, which represents the number of RRU cells in the topology.
- Category names and IDs.
- Personal equipment IDs and names of the device types.
- Initial load configurations, which represent the initial load per group of personal equipment.
- Adjacency cell configurations.
- Handover percentage for a group of personal equipment.
- Mean and variance ratios for the handover process.
- Simulation time step, which represents the period of data fetching process from NFs.
- Simulation time, which represents the total length of the simulation.

For the beginning of the data generation simulation, regarding the subscriber category and the personal equipment type, initial load configurations for each RRU cell can be seen in Table 3. As seen, different loads are assigned for each personal equipment type as well as each subscriber category. The logic behind these load values is that it is more likely for a user to subscribe to the platinum category rather than gold and silver categories for the cell phone subscription. On the other hand, it is less likely to subscribe to a higher category for a tablet computer, vehicle, and IoT device. Also, for the RRU cell with four adjacent RRU cells in Fig. 4, the mean handover ratios in Table 2 are doubled in order to keep the balance in the network. Moreover, as aforementioned, the handover ratios in Table 2 are mean values. We assume that the handover events exhibit Gaussian distribution and the mean and the variance parameters are given as follows

$$\Delta H_{\text{ratio}} \sim \mathscr{N}(\mu, \frac{\mu}{8}), \tag{1}$$

Table 3. Initial loads.

Subscriber category (Number)	IoT device	Vehicle	Cell phone	Smart watch	Tablet computer
Platinum (1)	3 Gbps	20 Gbps	90 Gbps	1 Gbps	6 Gbps
Gold (2)	4 Gbps	18 Gbps	72 Gbps	1 Gbps	5 Gbps
Silver (3)	5 Gbps	16 Gbps	53 Gbps	1 Gbps	5 Gbps

where ΔH_{ratio} is the resulting handover ratio, and $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian random variable with mean μ and variance σ^2.

Then, we generate six months long network traffic data, which consist of a network snapshot in each 15 minutes interval (Δt). In each of these intervals, UE may handover between adjacent cells.

In order to make our data set more realistic, we add anomalies to the generated network traffic data throughout the simulation. We describe anomalies as unexpectedly generated large amounts of network traffic compared to the average network traffic, which fades and stabilizes in time. While creating these kinds of anomalies, we are inspired by our daily lives, where constantly some videos go viral or some breaking news occurs, which affects the network traffic data in an increasing manner.

While describing anomalies, we also label the time points, which include abnormal traffic loads. This is required in order to generate the ground truth and to extract behaviour information from the 5G network data set.

We implement data generation methodology in Python programming language. Many ML algorithms are implemented as a library for this language since it is widely used, easy to read, and code. Alongside implemented ML libraries, data processing libraries are also available for Python, which makes this programming language very convenient for our purposes.

In the implementation, we create models for device type, subscriber category, and RRU cell. The cell object model contains information of adjacent cells as mentioned in Section IV.B. The subscriber category object model keeps the information for each device type including their loads and statistics. The device type object model handles handover operations and keeps load information at the current time point. After deciding all three models for the data generation process, we create the proposed system model by using a predefined adjacency matrix for RRU cells, handover parameters for device types, simulation duration, and percentage of anomaly occurrences. After defining all these, data generation simulation is ready to start. At each time point t (i.e., 0, Δt, $2 \times \Delta t, \cdots, t$), each device type determines how much load will be transferred for handover, then handover operation occurs by transferring the network load for the particular device type from the source RRU cell to the target RRU cell. Moreover, the starting and ending time of anomalies are randomly chosen before the data generation simulation starts. Throughout the simulation, when an anomaly period starts, a predefined percentage of the network load, which increases exponentially during the anomaly period, is added to each device. With the completion of the data generation process, all network load data are exported in order to analyze and generate appropriate features, which is discussed in Section VI.A. In Fig 5, aggregated data rates of each cell are represented.

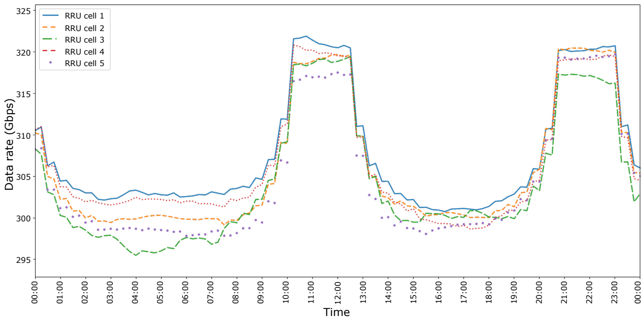


Fig. 5. Aggregated data rate per RRU cell for a sample day.

VI. MACHINE LEARNING MODELS

With the recent advancements in the technology, and consequently the huge amount of data available and required to be processed, ML algorithms become very popular and are seen as a potential solution for many different types of problems. Out of past data, ML models are used to produce certain required information. There are various types of ML algorithms, where each of these algorithms are designed to solve a certain kind of problem. As mentioned in Section V, ML models can be categorized under three parts, and the ones we use in this paper fall under the supervised learning algorithms category. Since the generated 5G data set is properly labeled, we can benefit from the merits of supervised learning algorithms. In the paper, we focus on two different problems. For both of these problems, we use different solution specific ML models and compare the performance of these models.

A. Feature Extraction

ML models need to be trained in order to produce accurate results. Data should be processed and fitted in a form that the ML model can understand. In order to train ML models, some of the certain features are required to be extracted from the data set. With the help of human insight and elaborative data analysis, these features become very helpful to an ML algorithm throughout the training process, and consequently resulting in better prediction results. On the other hand, DL models are generally capable of learning the patterns of the data without the help of these insights and features due to their neural network nature. Yet, for a fair comparison, we feed both ML and DL models with the same input.

In the feature extraction process, we produce features by considering the data rates in the previous time slots. Extracted features are specifically as follows

- **last2_mean**: Average data rate of the last two Δt.
- **last4_mean**: Average data rate of the last four Δt.
- **last8_mean**: Average data rate of the last eight Δt.
- **per_change_last2**: Percentage of the data rate difference

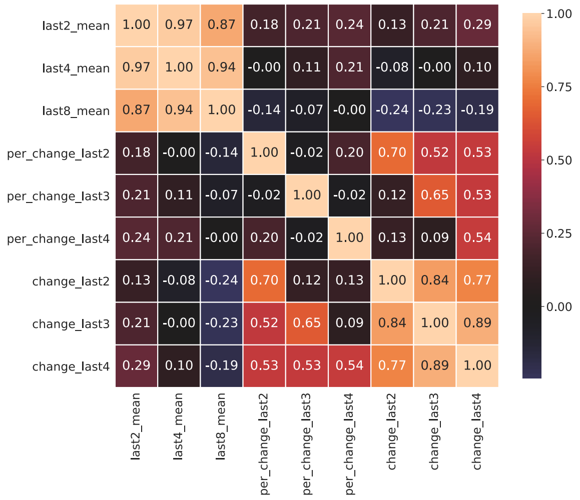


Fig. 6. Correlation matrix of the extracted features.

between the last two Δt.

- **per_change_last3**: Percentage of the data rate difference between the $t - \Delta t$ and $t - 3 \times \Delta t$.
- **per_change_last4**: Percentage of the data rate difference between the $t - \Delta t$ and $t - 4 \times \Delta t$.
- **change_last2**: Data rate difference between the last two Δt.
- **change_last3**: Data rate difference between the $t - \Delta t$ and $t - 3 \times \Delta t$.
- **change_last4**: Data rate difference between the $t - \Delta t$ and $t - 4 \times \Delta t$.

Among all considered features, feature importance and correlation tests are performed. These two tests give information about the quality of features, which affects the quality of the trained models. Feature importance test shows that last2_mean, last4_mean, and last8_mean are the most important features. These features are followed by per_change_last2, per_change_last3, and per_change_last4, which are followed by per_change_last2, per_change_last3, and per_change_last4. However, as can be seen in Fig. 6, data rate averages and change in data rates have higher correlations among their feature subgroup, compared to the percentage change in data rates feature. Considering the feature importance test, we select the most important feature from data rate average features, which is last2_mean and all percentage change in data rates features. Change in data rates features is not selected due to their low scores compared to other features in feature importance test.

B. Network Load Performance Prediction

As depicted in Section III, one of the main focus of this paper is to estimate network the load performance. We define this problem as a time series problem, and then use three different ML models, namely LR, RNN, and LSTM, where LSTM is a slightly modified version of RNN. While training these ML models, we both use the labeled data set (i.e., data rate of each time slot is the label) and the extracted features in Section VI.A.

B.1 Linear Regression

LR fits a linear relation between the given features of many observations and labels. Since LR is one of the most commonly used ML models for predictions and forecasting problems, we choose this model as a base model for comparison purposes.

B.2 Recursive Neural Networks

Neural networks and deep learning provide effective solutions to many problems. The nature of neural networks enables the algorithm to learn complex relations in the features and produces highly accurate results. RNN is a specialized version of the neural network algorithm, enabling to carry data from previous neurons and increasing accuracy for prediction and forecasting problems. We use RNN for our network load prediction purposes. The model consists of one simple RNN layer, four hidden layers, and one output layer. The loss function of RNN is mean absolute error (MAE), which is also one of the performance metrics we consider.

B.3 Long-Short Term Memory

LSTM is a modified version of RNN, where neuron structure is modified compared to RNN. While RNN keeps information from previous neurons, LSTM's complex structure helps to keep information from very past data unlike simple RNNs. In this way, even there is a time gap between an observed pattern, LSTM is able to make accurate predictions. In the LSTM model, we use one LSTM layer, four hidden layers, and one output layer. The loss function of LSTM is also MAE similar to the RNN model.

C. Anomaly Detection

As aforementioned, the contribution of this paper from the NWDAF perspective is twofold, where the one we investigate in this subsection is about gaining insights of behaviour information for a group of UE. To do so, network load anomalies are added as well as labeling each time slot as normal or abnormal while creating the network traffic data set. This behaviour information is classified by using two different ML models, namely, logistic regression, and extreme gradient boosting.

C.1 Logistic Regression

Logistic regression is a commonly used model for various classification problems. Logistic regression uses a logistic function in order to estimate the labels of the given data. We choose logistic regression as the base model for the anomaly detection problem. Since the produced data has unequal numbers of the anomaly and normal states throughout the simulation, we set class weight parameters in order to overcome the issues that will be caused due to the imbalanced number of label types (i.e., anomaly and normal).

C.2 Extreme Gradient Boosting

For classification problems, there are many algorithms that use tree-based approaches. Tree-based algorithms make a decision based on given features while reducing the loss function. XGBoost is a state-of-the-art implementation of gradient boosted decision trees designed for speed and performance.

Considering it is a widely used algorithm for classification problems, we also use XGBoost model in this paper. As discussed in the logistic regression subsection, the produced data has an imbalanced number of anomalies. In order to eliminate the problems caused by the imbalanced number of labels, we also tune XGBoost model by reducing the effectiveness of the dominating label. This methodology helps the model to achieve a higher score in terms of the AUC-ROC performance metric.

VII. SIMULATION RESULTS

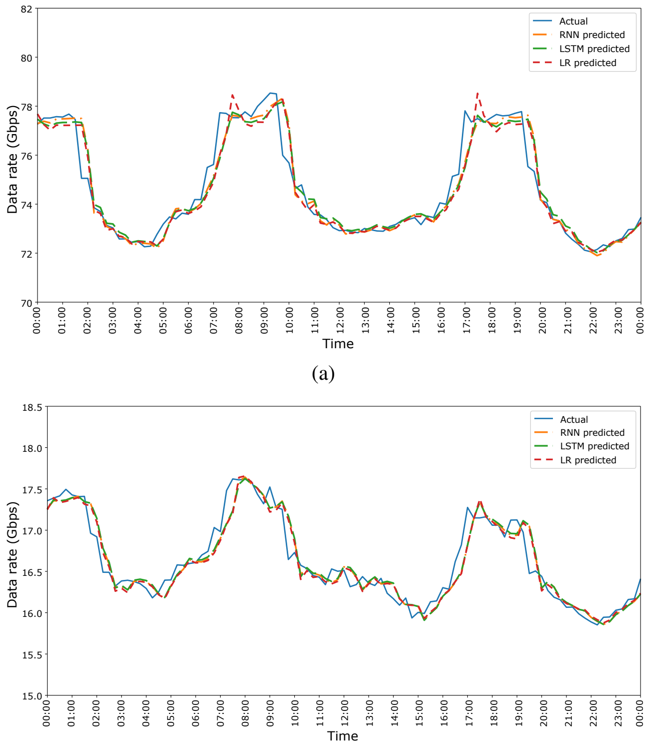
Considering the scenarios described in Section IV, and generated the data by following the procedures in Section V, the following results are obtained. Results are considered in two folds. In the first one, we investigate the network load performance using LR, LSTM, and RNN models. In the latter one, we investigate the anomalies in the network throughout the simulation using logistic regression and XGBoost models.

Table 4. Results for load performance predictions.

Metric name	(Cell - SubsCat) ID	LR	LSTM	RNN
MAPE	1 - 1	0.577	0.504	0.512
	1 - 2	0.575	0.512	0.573
	1 - 3	0.579	0.511	0.498
	2 - 1	0.578	0.510	0.499
	2 - 2	0.576	0.510	0.521
	2 - 3	0.585	0.504	0.519
	3 - 1	0.761	0.680	0.735
	3 - 2	0.757	0.688	0.754
	3 - 3	0.750	0.696	0.735
	4 - 1	0.581	0.507	0.487
	4 - 2	0.576	0.505	0.501
	4 - 3	0.581	0.505	0.499
	5 - 1	0.578	0.506	0.515
	5 - 2	0.581	0.511	0.539
	5 - 3	0.583	0.509	0.500
	Average	**0.615**	**0.544**	**0.560**
MAE	1 - 1	0.185	0.157	0.148
	1 - 2	0.237	0.205	0.233
	1 - 3	0.290	0.251	0.217
	2 - 1	0.184	0.158	0.139
	2 - 2	0.238	0.202	0.192
	2 - 3	0.291	0.242	0.219
	3 - 1	0.223	0.196	0.204
	3 - 2	0.282	0.252	0.269
	3 - 3	0.339	0.312	0.307
	4 - 1	0.185	0.157	0.135
	4 - 2	0.238	0.205	0.180
	4 - 3	0.289	0.243	0.218
	5 - 1	0.185	0.159	0.147
	5 - 2	0.238	0.204	0.197
	5 - 3	0.289	0.243	0.214
	Average	**0.247**	**0.213**	**0.202**

A. Network Load Performance Prediction

Using the aforementioned three AI/ML models, we perform the simulation for each user equipment within each subscriber category and RRU cell. Then, we calculate the mean average percentage error (MAPE) and mean absolute error (MAE) for each of these scenarios. The results using these metrics can be seen in Table 4. Note that cell ID represents the RRU cell number as depicted in Fig. 4, and SubsCat represents the subscriber categories, platinum, gold, and silver from one to three, respectively. In Table 4, it can be seen that LSTM and RNN perform better than LR in all scenarios as expected. For almost half of the scenarios, RNN outperforms LSTM. The reason for this competition is caused due to the nature of neural networks, which includes significant randomization factor. However, in the average, LSTM also outperforms RNN in terms of MAPE metric. Yet, as seen, RNN outperforms LSTM in terms of MAE metric. This is because RNN is more successful in detecting unexpected conditions (i.e., unsteady data rates). In other words, LSTM is more successful in detecting the steady data rates compared to RNN. Even though the score of RNN according to the MAE metric is lower, since unsteady data rates are higher than the steady ones, according to the MAPE metric, the score of LSTM is lower compared to RNN due to the ratio between the numerator and the denominator.



(a)

(b)

Fig. 7. Time versus data rate for different AI/ML models: (a) RRU cell number is three, subscriber category is gold, and UE type is cell phone and (b) RRU cell number is four, subscriber category is platinum, and UE type is vehicle.

In Fig. 7, data rate predictions of RNN, LSTM, and LR models can be seen as well as actual results throughout the time, which is randomly captured one full day within the six months long data set. Both in Fig. 7(a), where the results captured from RRU cell number three, subscriber category gold, and cell phone device type, and in Fig. 7(b), where the results captured from RRU cell number four, subscriber category platinum, and vehicle device type, we come up with two distinct conclusions. Firstly, ANN models outperform LR showing that they are better at predicting sudden changes (i.e., sharp slopes). Since LR does not provide a complex formula for predictions, these re-

Table 5. Average results for anomaly predictions (averaging is done over device types).

		Logistic regression			XGBoost		
Cell ID	SubsCat	AUC-ROC	Accuracy	Precision	AUC-ROC	Accuracy	Precision
1	Platinum	88.0%	55.4%	77.8%	91.5%	63.4%	77.5%
1	Gold	87.4%	56.0%	77.4%	91.5%	63.5%	77.9%
1	Silver	87.6%	55.3%	77.4%	91.7%	63.6%	78.0%
2	Platinum	87.3%	55.5%	77.0%	91.4%	63.3%	77.6%
2	Gold	87.6%	55.7%	77.5%	91.2%	63.1%	77.6%
2	Silver	87.5%	56.1%	77.5%	91.9%	63.7%	78.0%
3	Platinum	84.9%	55.6%	75.2%	87.7%	60.1%	75.5%
3	Gold	85.4%	55.8%	75.7%	88.5%	89.8%	76.4%
3	Silver	84.5%	54.9%	75.1%	87.9%	59.4%	76.1%
4	Platinum	88.0%	56.3%	77.5%	91.6%	63.5%	77.7%
4	Gold	87.4%	55.7%	77.2%	91.4%	62.9%	77.6%
4	Silver	87.9%	55.6%	76.9%	91.8%	63.8%	77.8%
5	Platinum	87.2%	55.5%	77.0%	91.0%	63.1%	77.2%
5	Gold	87.5%	55.7%	77.2%	91.0%	63.0%	77.3%
5	Silver	87.4%	55.5%	77.1%	91.0%	63.0%	77.6%
Average		**87.0%**	**55.6%**	**76.9%**	**90.7%**	**62.6%**	**77.3%**

sults were expected. Secondly, for steady data rates, it can be seen that LSTM predictions are closer to the actual values followed by RNN predictions and later LR predictions. However, for unsteady data rates, RNN predictions are closer to the actual values followed by LSTM predictions that are also followed by LR predictions. One can clearly see these two observations by looking between 07:00 and 10:00 for unsteady data rates in Fig. 7(a), and between 00:00 and 01:30 in Fig. 7(b).

Lastly, even though LR performs reasonable with our synthetically generated data set, we would like to point out that regression models might perform worse, and consequently are likely to provide lower accuracy values in actual deployments. We have generated the data set with high standard variation and randomization parameters to make it as realistic as possible. Even though the base idea behind training ML models is the same, it can be expected to have slightly different results in basic ML models. On the other hand, complex ML models are expected to provide similar performance compared to the performance with the proposed data set.

B. Anomaly Detection

For the detection of the anomalies in the 5G network traffic data, we use logistic regression and XGBoost models. In order to measure the performance of these two models, we use AUC-ROC as the performance metric. Moreover, we visualize the results using receiver operation characteristics (ROC) curves. ROC curve has two axes, namely, true positive rate (i.e., probability of detection -$\mathbf{P_d}$), and false positive rate (i.e., false alarm rate -$\mathbf{P_f}$). $\mathbf{P_d}$ is the ratio of true positives versus actual true values. It represents the accuracy of the model for the detection of the anomalies (i.e., percentage of the successful predictions among the network states having anomaly). The other axis, $\mathbf{P_f}$, is the ratio of the false positives versus actual negative values. It represents the probability of false decisions when the actual network state is not exhibiting anomaly (i.e., percentage of the false predictions for the non-anomaly states). The desired ROC curve shape is similar to an elbow going towards the upper side of the figure.

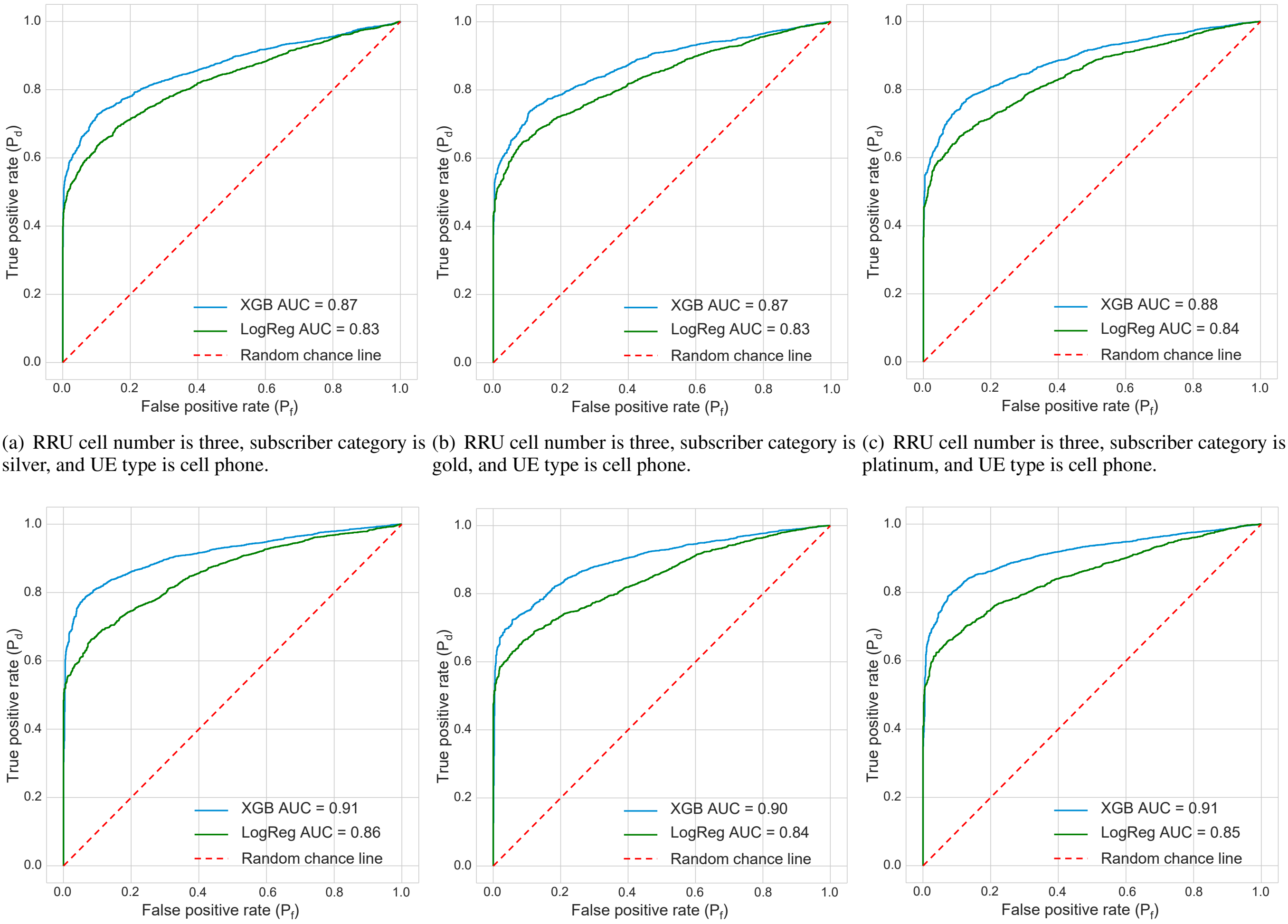
AUC-ROC, accuracy, and precision metrics are summarized for XGBoost and logistic regression models in Table 5. As depicted in Table 5, the AUC-ROC score of XGBoost model is significantly higher than the logistic regression model for each subscriber category and RRU cell. Consequently, as also seen in the average AUC-ROC, accuracy, and precision scores, XGBoost model predicts anomalies much better than the logistic regression model. While precision is not improved significantly, we observe a significant improvement in accuracy.

In Fig. 8 we compare the results of cell phones' different subscriber categories that is under RRU cell number three and four. Comparing Fig. 8(a) with Fig. 8(d), Fig. 8(b) with Fig. 8(e), and Fig. 8(c) with Fig. 8(f), it can be seen that the area under ROC curve difference between XGBoost and logistic regression is higher in favor of XGBoost model. Overall, among RRU cells three and four, logistic regression is not able to increase $\mathbf{P_d}$ as XGBoost does for a fixed false alarm rate. If we consider the network topology in Section IV.B, where RRU cell three has more neighbors than RRU cell four, RRU cell three is more volatile due to having higher handover ratio compared to the other RRU cells. Since logistic regression is a less complex ML model than XGBoost, logistic regression is worse with the predictions in both cells and is not able to improve the performance score adequately.

Among subscriber categories, as depicted in subfigures of Fig. 8, the ROC curves of the models are not varying much. We understand the fact that subscriber categories do not affect model performance significantly. On the other hand, one can see that when two ROC curves (i.e., curves of logistic regression and XGBoost models) are compared in all cases, the logistic regression curve stays below XGBoost curve in each particular case, which means that XGBoost achieves higher $\mathbf{P_d}$ for a given $\mathbf{P_f}$.

VIII. CONCLUSION AND FUTURE WORK

This paper presents a novel system to achieve intelligent network analytics for 5G cellular networks. To this end, we first describe NWDAF in the service-based architecture of 5G cellular networks, and then employ several ML techniques to overcome two major problems. In the first problem, network load is predicted using time series analysis, by specifically using linear



(a) RRU cell number is three, subscriber category is silver, and UE type is cell phone.

(b) RRU cell number is three, subscriber category is gold, and UE type is cell phone.

(c) RRU cell number is three, subscriber category is platinum, and UE type is cell phone.

(d) RRU cell number is four, subscriber category is silver, and UE type is cell phone.

(e) RRU cell number is four, subscriber category is gold, and UE type is cell phone.

(f) RRU cell number is four, subscriber category is platinum, and UE type is cell phone.

Fig. 8. False positive rate versus true positive rate for logistic regression and XGBoost models (AUC-ROC).

regression, LSTM, and RNN models. In the second problem, anomalies in the network are classified by using a state-of-the-art tree-based gradient boosting technique, XGBoost, and logistic regression models. Moreover, we introduce a systematical generation of a cell-based data set to evaluate network data analytics for 5G cellular networks using the fields defined by the 5G standard document. In the experiments, we see that neural network models outperform the linear regression model for the correct prediction of the network load. Similarly, tree-based XGBoost outperforms logistic regression while classifying the anomalies in the network. In conclusion, we show a very practical usage of NWDAF by using popular and common ML models.

Due to the novelty of NWDAF, there are also many open issues in the literature. One such issue is the multi-label classification for the detection of anomalies. It is because an anomaly event can be at different levels, and classifying this event with its corresponding level is the ultimate goal for the network operators. Another issue is enriching generated data with the necessary fields by inspiring from 3GPP specifications in order to observe and gather analytics on network slice level base. Additionally, based on the generated data set, our future work includes communication pattern detection using exploratory data analysis. Last but not the least, our work can be extended by using different AI/ML models while focusing on other capabilities of NWDAF.

REFERENCES

[1] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, July 2015.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys & Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart. 2016.

[3] "Mobile visual networking index (VNI)," accessed: 2019-12-05. [Online]. Available: https://www.cisco.com/c/m/en_us/solutions/service-provider/visual-networking-index.html

[4] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sept. 2015.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Selected Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, June 2014.

[6] "Non-standalone and standalone: Two standards-based paths to 5G," accessed: 2019-12-05. [Online]. Available: https://www.ericsson.com/en/blog/2019/7/standalone-and-non-standalone-5g-nr-two-5g-tracks

[7] 3GPP, "Architecture enhancements for 5G System (5GS) to support network data analytics services," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 23.288), Sept. 2019, version 16.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3579

[8] 3GPP, "System architecture for the 5G System (5GS)," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 23.501), Sept. 2019, version 16.2.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144

[9] 3GPP, "5G System; Network data analytics services; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.520), September 2019, version 16.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3355

[10] 3GPP, "5G System; Unified data management services; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.503), Sept. 2019, version 16.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3342

[11] C. Hernandez-Chulde and C. Cervello-Pastor, "Intelligent optimization and machine learning for 5G network control and management," in *Proc. PAAMS*, June 2019, pp. 339–342.

[12] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[13] 3GPP, "Study on latency reduction techniques for LTE," 3rd Generation Partnership Project (3GPP), Technical Report (TR 36.881), July 2016, version 14.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2901

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[15] S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz, and T. Tugcu, "Synthetic 5G cellular network data for NWDAF," 2019. [Online]. Available: https://github.com/sevgicansalih/nwdaf_data

[16] 3GPP, "5G System; Usage of the unified data repository service for policy data, application data and structured data for exposure; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.519), Oct. 2019, version 16.2.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3357

[17] 3GPP, "5G System; Network exposure function northbound APIs; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.522), Sept. 2019, version 16.2.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3437

[18] 3GPP, "5G System; Background data transfer policy control service; Stage 3," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.554), Sept. 2019, version 16.2.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3438

[19] 3GPP, "T8 reference point for Northbound APIs," 3rd Generation Partnership Project (3GPP), Technical Specification (TS 29.122), Sept. 2019, version 16.4.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3239

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.

[21] S. Barmpounakis, P. Magdalinos, N. Alonistioti, A. Kaloxylos, P. Spapis, and C. Zhou, "Data analytics for 5G networks: A complete framework for network access selection and traffic steering," *International J. Advances Telecommunications*, vol. 11, no. 3, pp. 101–114, 2018.

[22] A. Klautau, P. Batista, N. Gonzalez-Prelcic, Y. Wang, and R. W. Heath Jr., "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. ITA*, 2018, pp. 1–1. [Online]. Available: http://ita.ucsd.edu/workshop/18/files/paper/paper_3313.pdf

[23] B. Koksal, R. Schmidt, X. Vasilakos, and N. Nikaien, "CRAWDAD dataset eurecom/elasticmon5g2019 (v. 2019-08-28)," https://crawdad.org/eurecom/elasticmon5G2019/20190828, Aug. 2019.

[24] R. Shafin, L. Liu, V. Chandrasekhar, H. Chen, J. Reed, and J. C. Zhang, "Artificial intelligence-enabled cellular networks: A critical path to beyond-5G and 6G," *arxiv preprint arxiv:1907.07862*, 2019.

[25] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, 2016.

[26] R. Casellas, R. Martínez, L. Velasco, R. Vilalta, P. Pavón, D. King, and R. Muñoz, "Enabling data analytics and machine learning for 5G services within disaggregated multi-layer transport networks," in *Proc. IEEE ICTON*, July 2018, pp. 1–4.

[27] J. Moysen and L. Giupponi, "From 4G to 5G: Self-organized network management meets machine learning," *Computer Commun.*, vol. 129, pp. 248–268, 2018.

[28] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys & Tuts.*, vol. 21, no. 4, pp. 3039–3071, 2019.

[29] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *arxiv preprint arxiv:1809.08707*, 2019.

[30] H. Fang, X. Wang, and S. Tomasin, "Machine learning for intelligent authentication in 5G and beyond wireless networks," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 55–61, Oct. 2019.

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

[32] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. IEEE INFOCOM*, 2018, pp. 1961–1969.

[33] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu, "Machine learning for vehicular networks," *arXiv preprint arXiv:1712.07143*, 2017.

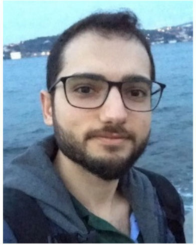
[34] N. Cheng, F. Lyu, J. Chen, W. Xu, H. Zhou, S. Zhang, and X. S. Shen, "Big data driven vehicular networks," *IEEE Netw.*, vol. 32, no. 6, pp. 160–167, 2018.

[35] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. IEEE ITA*, 2018, pp. 1–9.

[36] X. Gao, L. Dai, Y. Sun, S. Han, and I. Chih-Lin, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE ICC*, 2017, pp. 1–6.

[37] L. Bai, C.-X. Wang, J. Huang, Q. Xu, Y. Yang, G. Goussetis, J. Sun, and W. Zhang, "Predicting wireless MmWave massive MIMO channel characteristics using machine learning algorithms," *Hindawi Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–12, Aug. 2018.

Salih Sevgican received B.Sc. degree on Computer Engineering from Bogazici University, Turkey on June 2018. He continued to his Master's degree in Computer Engineering from Bogazici University. He is currently pursuing his exchange term in University of Oulu, Finland. His research interest includes next-generation cellular networks, software-defined networking, cloud computing and artificial intelligence.



Meriç Turan received his B.Sc. degree in Computer Engineering from Istanbul Technical University, Istanbul, Turkey in 2017, and his M.Sc. degree in Computer Engineering from Bogazici University, Istanbul, Turkey in 2019. He is currently a Ph.D. candidate and teaching assistant at the Department of Computer Engineering, Bogazici University. He is also a recipient of TUBITAK 2211 Ph.D. Fellowship. His research interest includes next-generation cellular networks, wireless networks, and molecular communications.

Kerim Gökarslan received B.S., degrees in Computer Engineering and Mathematics from Bogazici University, Turkey in June 2017 with high honors. He received his M.Sc. degree in Computer Science from Yale University, New Haven, Connecticut, USA in May 2020. He is currently a graduate researcher at Bogazici University. His research interest includes cellular networks, software-defined networking (SDN), network function virtualization (NFV), cloud computing, wireless communication, distributed systems, and network verification. He was a recipient of numerous NSF student grants, Cisco Systems CPSG Q3-FY18 Innovation Award, and TUBITAK 2205 Science Fellowship.

H. Birkan Yilmaz is currently an Assistant Professor in the Department of Computer Engineering at Bogazici University, Istanbul, Turkey. He received his B.S. degree in Mathematics in 2002, and the M.Sc. and Ph.D. degrees in Computer Engineering from Bogazici University in 2006 and 2012, respectively. He worked as a post-doctoral researcher at Yonsei Institute of Convergence Technology, Yonsei University, South Korea for four years and at Universitat Politecnica de Catalunya, Spain (via Beatriu de Pinos fellowship) for two years. He was awarded TUBITAK National Ph.D. Scholarship during his Ph.D. studies and the Marie Sklodowska-Curie Actions Seal of Excellence in 2016. He was the co-recipient of the best demo award in IEEE INFOCOM (2015) and best paper award in AICT (2010) and ISCC (2012). His research interests include cognitive radio, spectrum sensing, molecular communications, and detection and estimation theory. He is currently in the editorial board of IEEE Wireless Communications Letters and IEEE Transactions on Molecular, Biological, and Multi-Scale Communications. He is awarded as Exemplary Reviewers of the IEEE Wireless Communications Letters in 2014, 2017, and 2019. He has also served as a TPC Member for many IEEE conferences, such as IEEE GLOBECOM and ICC. He is a member of IEEE and TMD (Turkish Mathematical Society).

Tuna Tugcu received the B.S. and Ph.D. degrees in Computer Engineering from Bogazici University, Istanbul, Turkey, in 1993 and 2001, respectively, and the M.S. degree in Computer and Information Science from the New Jersey Institute of Technology, Newark, NJ, USA, in 1994. He was previously a postdoctoral fellow and a visiting professor with Georgia Institute of Technology, USA. He is currently a professor in the Department of Computer Engineering, Bogazici University. His research interests include nanonetworking, molecular communications, wireless networks, and IoT. Prof. Tugcu has served with the North Atlantic Treaty Organization science and technology groups and the IEEE standards groups. He is an associate editor in IEEE Transactions on Molecular Biological and Multi-scale Communications journal.