NETWORK DATA ANALYTICS FUNCTION IN 5G NETWORKS

by

Salih Sevgican

B.S., Computer Engineering, Boğaziçi University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2022

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor Prof. Tuna Tuğcu for his endless support, patience and confidence, his invaluable guidance, encouragement and coaching for the last five years.

I am grateful to Assoc. Prof. Ali Emre Pusane and Assist. Prof. Ahmet Teoman Naskali for their participation in my thesis committee and for their precious feedback.

I would like to offer my deepest thanks to the NETLAB members and all the participants of our research group; Meriç, Kerim, Yunus and Prof. Birkan for their hard work, support, creative ideas and contributions to my research.

My partners in business: Kutay, Özer and Buğra, deserve a mention for their unlimited support and trust. I would like to thank them for never doubting me for finishing my graduate education.

I would like to mention Prof. Ari Pouttu and Idris Badmus for their guidance and teachings during my Erasmus term in Oulu, Finland.

I cannot begin to express my thanks to my family for their unconditional love, care and support. I would like to thank my grandmother specifically for her wishes, prayers and motivational speeches.

And finally, I am deeply thankful to my dear Anelya, for her trust, support, encouragement and for her love. Thank you for being with me every step of the way.

ABSTRACT

NETWORK DATA ANALYTICS FUNCTION IN 5G NETWORKS

Wireless cellular networking in the world goes through a tremendous structural change where many advances in technology find an opportunity to present themselves for assistance. 5G cellular network, the most recent generation wireless network currently undergoing implementation, welcomes artificial intelligence with the novel network data analytics function (NWDAF). NWDAF is a data analytics mechanism where other components of 5G can request information from in order to utilize their operations. In this thesis, the structure and protocols of NWDAF are described. A 5G network data set is generated by using the fields obtained from the technical specification documents provided by 3rd Generation Partnership Project (3GPP). To bring the generated data set closer to reality, randomly created anomalies are added. Several machine learning (ML) algorithms are trained to study two aspects of NWDAF, namely network load prediction and anomaly detection. Linear regression (LR), recurrent neural network (RNN) and long-short term memory (LSTM) algorithms are implemented and trained using the generated data set and a data set obtained from a real enterprise network for network load prediction [1, 2]. Mean absolute error and mean absolute percentage error performance metrics indicate that RNN and LSTM outperform LR in both generated and real life data sets. LSTM is the best performing algorithm for the real life data set. Logistic regression and a tree-based classifier, XGBoost are implemented for anomaly detection, and trained using the generated data set to maximize the area under receiver operating characteristics curve. The results indicate that tree-based classifier XGBoost outperforms logistic regression. These predictions are expected to assist 5G service-based architecture through NWDAF to increase its performance.

ÖZET

5G AĞLARINDA AĞ VERİ ANALİTİK İŞLEVİ

Dünyadaki kablosuz hücresel ağ, teknolojideki birçok ilerlemenin yardım için kendilerini sunma fırsatı bulduğu muazzam bir yapısal değişimden geçiyor. Şu anda uygulanmakta olan en yeni nesil kablosuz ağ olan 5G hücresel ağ, yeni ağ veri analitiği işlevi (NWDAF) ile yapay zekayı memnuniyetle karşılıyor. NWDAF, 5G'nin diğer bileşenlerinin kendi operasyonlarını iyileştirmek için bilgi talep edebileceği bir veri analizi mekanizmasıdır. Bu tezde, NWDAF'ın yapısı ve protokolleri anlatılmaktadır. 3. Nesil Ortaklık Projesi (3GPP) tarafından sağlanan teknik şartname dokümanlarından elde edilen alanlar kullanılarak 5G ağ veri seti oluşturulmuştur. Yapay veri setini gerçeğe yaklaştırmak için rastgele oluşturulmuş anomaliler eklenir. Birkaç makine öğrenimi (ML) algoritması, NWDAF'nin iki yönünü, yani ağ yükü tahmini ve anormallik algılamayı incelemek için eğitilmiştir. Doğrusal regresyon (LR), tekrarlayan sinir ağı (RNN) ve uzun kısa süreli bellek (LSTM) algoritmaları, ağ yükü tahmini için yapay veri seti ve gerçek bir kurumsal ağdan elde edilen bir veri seti kullanılarak uygulanmıştır ve eğitilmiştir [1,2]. Ortalama mutlak hata ve yüzdesel ortalama mutlak hata performans ölçümleri, RNN ve LSTM'nin hem oluşturulan hem de gerçek hayattan toplanan veri setlerinde LR'den daha iyi performans gösterdiğini göstermiştir. LSTM, gerçek hayattan toplanan veri seti için en iyi performans gösteren algoritmadır. Lojistik regresyon ve ağaç tabanlı bir sınıflandırıcı olan XGBoost, anormallik tespiti için uygulanır ve alıcı işletim karakteristikleri eğrisi altındaki alanı en üst düzeye çıkarmak için yapay veri seti kullanılarak eğitilmiştir. Sonuçlar, ağaç tabanlı sınıflandırıcı XGBoost'un lojistik regresyondan daha iyi performans gösterdiğini ortaya çıkarmıştır. Bu tahminlerin, performansını artırmak için NWDAF aracılığıyla 5G hizmet tabanlı mimariye yardımcı olması bekleniyor.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a	Bias term for logistic regression equation
b_i	Weight coefficient of i^{th} value for logistic regression equation
e	Euler's number
FN	Number of false negatives in classification results
FP	Number of false positives in classification results
N	Number of predictions
P_d	True Positive Rate
P_r	False Positive Rate
t	A single time-frame in simulation
TN	Number of true negatives in classification results
TP	Number of true positives in classification results
y	Output variable, prediction
y_i	i^{th} output variable, prediction
$\hat{y}_i$	i^{th} actual variable, that is compared to i^{th} prediction
x	Independent input variable
x_i	i^{th} independent input variable
α	Interceptor coefficient for linear regression equation, bias
β	Weight coefficients for linear regression equation
β_i	Weight coefficient of i^{th} value for linear regression equation
μ	Mean
σ^2	Variance
ΔH_{ratio}	Handover ratio
$\mathcal{N}$	Gaussian distribution
Δt	Time step interval in simulation

LIST OF ACRONYMS/ABBREVIATIONS

3GPP	3rd Generation Partnership Project
4G	Fourth Generation Wireless Cellular Networks
5G	Fifth Generation Wireless Cellular Networks
6G	Sixth Generation Wireless Cellular Networks
5GC	5G Core Network
AF	Application Function
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AP	Access Point
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristics
CDR	Charging Data Record
CHF	Charging Function
DL	Deep Learning
IoT	Internet of Things
LogReg	Logistic Regression
LR	Linear Regression
LSTM	Long-Short Term Memory
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
NFV	Network Function Virtualization

NSA	Non Standalone
NSSF	Network Slice Selection Function
NRF	Network Function Repository Function
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	Recurrent 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	Service Management Function
SubsCat	Subscriber Category
UDM	Unified Data Management
UDR	Unified Data Repository
UE	User Equipment
URLLC	Ultra Reliable and Low Latency Communication
XGBoost	eXtreme Gradient Boosting

1. INTRODUCTION

Next generation wireless cellular networks are designed and expected to handle a huge number of users and devices. To meet expectations, network architectures are becoming more complex and advanced with the usage of recent technological developments such as micro-service architectures and artificial intelligence (AI). The fourth-generation wireless cellular network (4G) systems, which are standardized by the 3rd Generation Partnership Project (3GPP), can provide speeds limited to a a couple of hundred Mbps to the end user devices. The data rates around this number are sufficient for applications such as high-definition TV streaming to work well [6,7]. However, considering the tremendously increasing number of devices and applications requiring higher data rates, 4G wireless network architecture is not designed to support these demands. Particularly, Cisco's Annual Internet Report [8] estimates that the number of mobile-connected devices will reach 13.1 billion and the number of machine-to-machine (M2M) type connections will be 14.7 billion by 2023. Thanks to the rapidly changing technology, Internet of Things (IoT) and M2M type communications require higher technical specifications from current wireless network technology designed for human-to-human communications [9]. In order to address the new-coming technology requirements and to update system with recent technological developments, the fifth-generation wireless cellular networks (5G) have come to our lives. Furthermore, the sixth-generation wireless cellular network (6G) specifications have become the subject of researchers' discussions [10–12]. In [13], the authors discuss groundbreaking intelligent technologies for 6G which also contains the subjects we study in this thesis. The outcomes of this study can be applied not only for 5G but for beyond 5G networks. However, for the sake of simplicity, the study is explained over the 5G.

The transition is already underway from the current 4G technology to 5G. 3GPP released the non-standalone (NSA) version of 5G to insure a smooth transition and a backward compatibility with existing cellular networks.

On top of 5G NSA version, 3GPP also released standalone specifications, which introduce 5G core architecture based on cloud technology. In [14], it is anticipated that 5G will be a major shift in the familiar model of cellular networks rather than an incremental advancement on 4G cellular network [14]. In terms of radio architecture, a cleaner version of Radio Access Network (RAN) where the control and data planes are separated, is specified. It addresses the need for different device communication protocols such as M2M communication [15]. Likewise, network functions (NF) introduced in 5G service-based architecture (SBA), are going to replace the ones from previous cellular network architectures. For example, 4G uses policy and charging rules function (PCRF), which is replaced by policy control function (PCF) in 5G. Similarly, 4G has charging data record (CDR) function whereas 5G improves this function with charging function (CHF).

There is another network function introduced in 5G specifications related to data analytics. With tremendously high data rates and a large number of devices in 5G network, data analytics is becoming a crucial component since it can unequivocally help to improve network features such as resource optimization. Network data analytics function (NWDAF) one of the newly proposed NFs for 5G networks, exposes network data analytics information and analytics event information to other NFs [3,4,16,17]. To provide such information, any AI or machine learning (ML) algorithm can be used as long as these algorithms satisfy the requirements of incoming data analysis requests [18]. NWDAF has several capabilities defined in the specifications for analytics information exposure and this thesis analyzes two of them, more specifically, the prediction of abnormal behaviour information and network load prediction. Because this way keeps the focus of the study clear. The studied capabilities are the prediction of abnormal behaviour information and the prediction of network load performance. The analytics information of both capabilities are essential to keep the network running with high performance and to satisfy the requirements of quality of service (QoS).

Before the introduction of NWDAF, there were studies in the literature about the usage of AI/ML in various areas of wireless networking. However, with next generation wireless networks, standards are set high with the need for ultra-reliable and low latency communication (URLLC), and an increase in the network data traffic is expected. Thus, assistance of AI/ML models has become a crucial necessity [19, 20]. A way to integrate such models into 5G SBA is to train an intelligent decision maker for every NF requires AI/ML utilization. Such way causes each NF to have independent models, which leads to recurrency and unnecessary data transfer among NFs. A centralized mechanism for intelligent decision making is the optimized way of implementing AI/ML models into the network. 3GPP acknowledged this necessity, and therefore introduced NWDAF to fulfill this requirement [16].

To perform a study in the area of ML, the most essential requirement is the data set. The best option is to work with a data set which is gathered from a deployed 5G network. However, to the best of our knowledge, the only publicly available data set that has the information suited for NWDAF research is published by the authors in [2]. This data set contains two months long recordings from access points (AP) in the campus of University of Oulu, Finland. It includes transmitted byte information from various APs and the number of users connected to the AP at the time of recording. Yet, the data set is relatively small and has no information about the state of the network in terms of abnormalities. To overcome the problem of finding a comprehensive data set, a publicly available data set [1] is generated inspired by 3GPP specifications of 5G networks. This generated data set includes a topology with a fixed number of cells, device types and device type features such as subscriber category. Various device types have different traffic and mobility patterns that create loads on each cell they are connected to. Each cell is modelled by using a set of features obtained from other NFs in the network. These features include the amount of transmitted bytes, list of categories associated with the subscription of the user, device ID and connected cell ID information at the time of monitoring. At last, randomly generated data load spikes, representing anomalies in the network are added to the generated data set in order to make it more realistic.

1.1. Contribution of Thesis

Given the data set gathered from a real life network [2], and carefully generated data set as explained in detail in Section 4.4, a novel system to perform network data analytics using AI/ML models is presented. The system is composed of two subsystems, where one performs network load performance prediction by using linear regression, long short term memory (LSTM), and recursive neural network (RNN) algorithms and the other performs classification on the current network status for a cell area by using logistic regression and a tree-based classification algorithm, namely extreme gradient boosting (XGBoost) [21]. The latter subsystem is studied only with the generated data set due to insufficient information of network status in the real life data set. Then, ML models under these subsystems are trained, given two different sets of data.

The key contributions of this thesis are summarized as follows:

- A synthetic data set for 5G networks is generated by following the definitions specified by 3GPP consortium for 5G cellular network.
- Various ML algorithms are proposed, which are compatible with the NWDAF system. Then, the proposed approaches are implemented and trained by using generated and real life data sets. One of the subsystems is responsible for network load performance prediction while the other is responsible for network status classification to understand whether there is an anomaly in the network.
- Simulation results based on the proposed system and real life network is presented. Effectiveness and performance of different ML algorithms used for NWDAF system are compared.

2. RELATED WORK

NWDAF is a relatively new component in the cellular network architecture. Due to this fact, the works related to NWDAF subject are not comprehensive. In [22], the authors show that NWDAF is a key enabler for data analytics in traffic steering and resource management. In [18], the use of ML and NWDAF for network slicing with network function virtualization (NFV) is explained.

In the literature, there is no data set available for the NWDAF scenarios studied in this thesis. In [23], the authors create a simulation in order to produce data for 5G multiple-input and multiple-output (MIMO) study. In [24], the authors monitor one eNodeB and one user equipment(UE) by using ElasticMON framework. Packet Data Convergence Protocol (PDCP), Radio Resource Control (RRC), and Mobile Access Control (MAC) are the fields of the data observed in this framework. Considering the coverage and singularity of the monitoring, this data set is also not suitable for ML purposes. In [25], the authors managed to create their own real traffic data of an enterprise network architecture in their campus and showed the analysis of network data and methods. They focus on the behaviour of APs in the network to understand the traffic data. Thus, to the best of our knowledge, there is no user traffic data in the literature, gathered from gNodeB based on 5G SA implementation and also suitable for NWDAF scenarios studied in this thesis.

AI/ML techniques in the next-generation of cellular networks have been extensively studied in the literature during the last decade. In [26], the authors address the importance of AI in the next-generation cellular networks and its necessity. Challenges and possible solutions are explained to implement intelligent networks based on AI. In [27], Jiang *et al.* explain that the intelligent decision making in the new radio systems is crucially beneficial for supporting high data rate specification in 5G and beyond networks.

Different ML algorithms and methods including supervised and unsupervised learning, are implemented and discussed for MIMO channel controlling and anomaly detection problems. However, the way these intelligent algorithms could be used in network analytics in 5G is not investigated. In [28], Casellas *et al.* study the AI/ML techniques to orchestrate and manage 5G network components without touching the subject of network data analytics. In [29], self-organizing 5G network paradigm is explained and incorporation of ML is discussed. While 5G technology specifications are required to be in millimeter wave (mmWave), radio network coverage is supposed to be more dense compared to previous generations. The researchers study network management by using different ML techniques. In [30] and [31], the authors discuss ML and DL techniques to study wireless networks. In [32], Fang *et al.* lean on the security concerns in the next-generation wireless networks and propose AI/ML in order to overcome the issue. In [33], the authors discuss that data analytics should be specifically focused on various tasks in the network to manage and utilize resources as swiftly as possible, since user demands become unparalleled. The authors consider and investigate numerous DL algorithms to handle resource management for such high demands. The new technologies of next generation wireless networks can bring ML into play for more field specific problems in the literature. In [34–36], vehicular networks based on 5G networks are investigated, and the role of ML in such networks is discussed. Moreover, [37–39] explain ML techniques in mmWave massive MIMO to manage the radio and beamforming.

In the end, ML and DL are the concepts that are new to the traditional mobile wireless networking, but researchers show that these AI techniques are essential for intelligent and efficient next-generation networking. NWDAF is not considered in many of these studies. The existing work and data set in the literature related to 5G do not meet the NWDAF scenario requirements of this thesis. In the Section 4.4 we generate our synthetic data set based on 3GPP specifications [3, 4, 16, 17].

3. 5G ARCHITECTURE

3.1. Service Based Architecture Overview

In 3GPP standards, 5G cellular networks have ultra-high throughput and ultra-low latency specifications [40]. 5G architecture handles these requirements by relying on massive radio deployment and mmWave communications. Orchestration of this massive architecture is becoming essential with the increase in the scale of the 5G network. 5G SBA is proposed by 3GPP [16]. In 5G SBA, there are NFs implemented as micro-services that are connected to each other on Service-Based Interface (SBI). Micro-service architecture of NFs enables operators to implement or update NF functionalities to their 5G network as needed.

The internal management of NFs is operated by NF repository function (NRF) that is responsible for available NFs and their services presented to network. NRF helps other NFs to discover available supported services. The traditional operations of a cellular network such as handover, and user access are managed by the access and management function (AMF) and session management function (SMF). Policy controls are handled by PCF. The 5G infrastructure, in addition to the traditional functions, contains newly introduced functions such as network slice selection function (NSSF) and NWDAF. With the progress of 5G implementation and updates on the 3GPP specifications, additional NFs can be included to 5G architecture.

3.2. Network Data Analytics Function

NWDAF is a network data analytics function defined in 5G cellular network specifications provided by 3GPP [4]. NWDAF is a function that provides analysis to other NFs when requested [16]. In order to provide network insights and analysis, NWDAF uses other NFs and subscribes information and data. The fusion of gathered data and the power of AI enables the data analytics function.

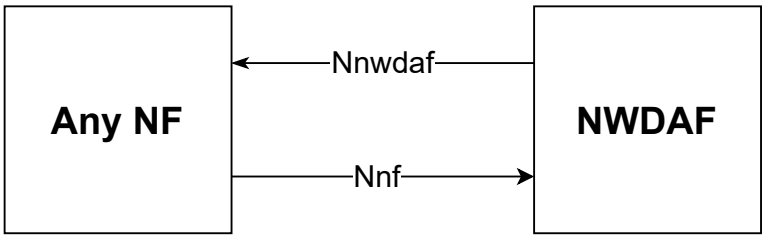


Figure 3.1. Data Collection from any 5GC NF and Network Data Analytics Exposure architecture (Inspired from [3]).

The relationship between NWDAF and other NFs are depicted in Figure 3.1. Data collection from other NFs is called the Nnf interface and analytics exposure to other NFs is known as Nnwdaf interface. As shown in the figure, NWDAF can either distribute network analysis data (i.e., analytics information) or notify analytics events (i.e., events subscription) to any NF which subscribes Nnwdaf interface. Moreover, NWDAF can request to collect data from other NFs using Nnf interface.

NWDAF provides two different services, namely, analytics information and events subscription [4]. The NFs in 5G Core (5GC) network, can use both of the NWDAF's services as required for NF operations. Analytics information service provides statistical information of the past events, or predictive information regarding the current network. Other Nfs can request specific analytics information from NWDAF in order to optimize their functional operation. After requesting analytics, NWDAF evaluates and responds to the NF requested analytics in a limited amount of time. Events Subscription service provides notifications to the other NFs about the analytics operations that are done in the NWDAF. NFs can subscribe or unsubscribe to these notifications, and if required the analytics information requested by one NF can also be shared with another NF in the network.

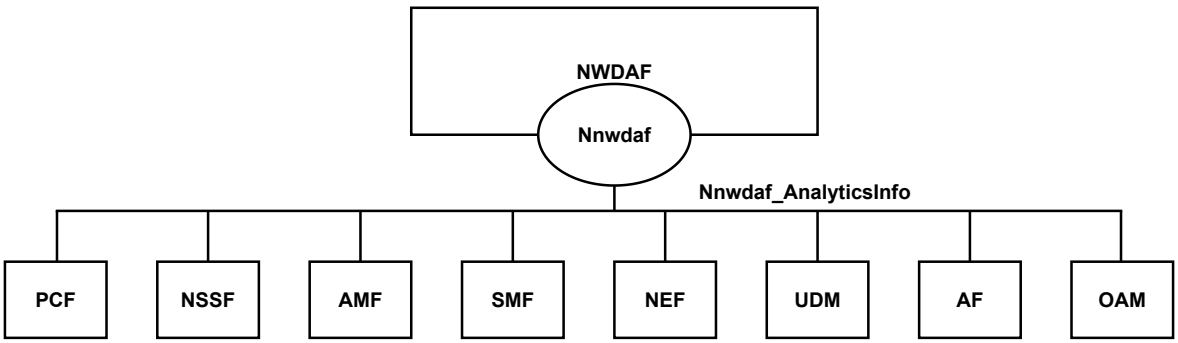


Figure 3.2. SBI architecture of the Nnwdaf analytics info service (Inspired from [4]).

The following events can be requested from analytics information service and can be observed from events subscription service:

- Abnormal behaviour information
- Network load performance
- Load level of network slice instance
- Load analytics information for a specific NF
- Communication properties for UE
- Congestion information of user data
- Mobility related information for a group of UE or a specific UE
- QoS sustainability performance and potential QoS change in a certain area
- Service experience for an application or for a network slice

Overall, events subscription service can be considered as notification hub and analytics observatory. On the other hand, analytics information service is the part where the required data analytics occurs, AI and ML models are trained and results are generated. Thus, considering the nature of NWDAF, our focus in this thesis is on NWDAF's analytics information service.

In this thesis, two events of network data analytics exposure are mainly investigated. Specifically, these are network load performance prediction and detection of abnormal behaviours in the network as anomaly. These analytics information events are served to other NFs using the Nnwdaf interface as depicted in Figure 3.2.

4. PROPOSED MODEL

As discussed in Section 3.1, 5G SBA consists of NFs where each service has a different role in the architecture, and as described in Section 3.2, NWDAF analyzes the network data obtained from other NFs. A system based on AI/ML is proposed to show the certain capabilities of NWDAF. A data set is a crucial necessity to implement and evaluate model performances of such intelligent system. We have generated a synthetic 5G network data set [1], by following the specifications of 3GPP in order to meet data set requirements. Additionally, another data set is found in the literature [2], and used to verify our NWDAF implementation.

In this chapter, the model representation of this proposed system is defined, and the data transfer workflow is explained in Section 4.1. Then, the algorithms considered to implement NWDAF are outlined in Section 4.2. Moreover, topology of studied network system is discussed in Section 4.3. Lastly, data generation process based on the depicted system workflow is explained in Section 4.4.

4.1. Model Representation and Workflow

Figure 4.1 is the high-level workflow explaining the system architecture that is used in this thesis. As shown, the data is obtained from UE and transferred to 5G SBA. Each NF uses the data from UE as required. For example, AMF manages the mobility based on the UE data, and UDR manages the storage of user information. NWDAF is connected to other NFs through service-based interface (SBI), which is the interface enabling the communication among all NFs. While the 5G network is operational, NWDAF gathers the data and information required from different NFs to make analysis and predictions. Network load prediction and anomaly detection capabilities of NWDAF require trained ML models, where NWDAF handles training by using several ML models and picks the one that performs best depending on the characteristics of the topology.

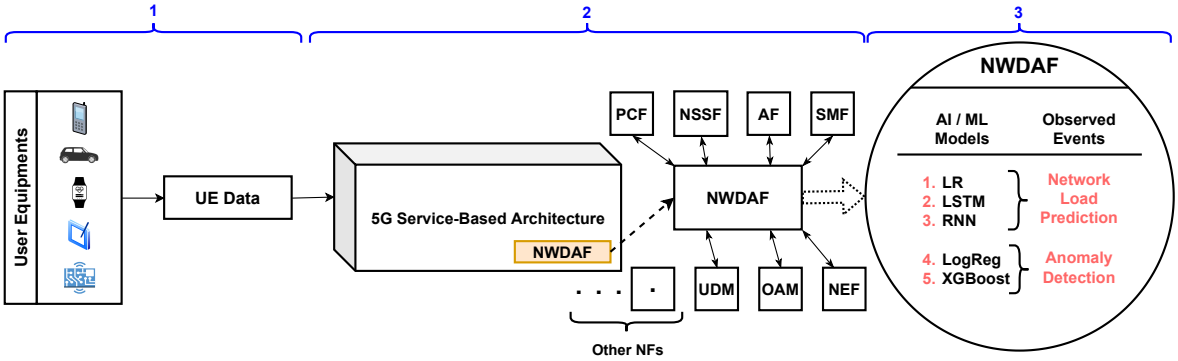


Figure 4.1. The workflow of the proposed system. The workflow consists of three stages: (1) UE data is generated and sent to 5G SBA, (2) NWDAF extracts data sent by UEs from related NFs, (3) NWDAF provides network analytics information with AI/ML models (Inspired from [5]).

4.2. Machine Learning Algorithms

While technology wildly progresses, the cumulative amount of data generated by smart devices increased enormously. Question of storing and managing a huge amount of data has brought challenges to computer scientists. A variety of data analysis tools are developed and many different algorithms are created, which can be trained with the data and used to make classifications, predictions, recognitions, and so forth. These algorithms are called machine learning algorithms and are used to produce new information from the data. As an analytics provider function, NWDAF holds several ML algorithms to make predictions, classifications and other tasks explained in Section 3.2.

In this thesis, two aspects of NWDAF will be focused on, namely anomaly detection and network traffic load prediction. The former is a type of classification problem, where as the latter one is a time series estimation problem, both of which require ML algorithms under supervised learning category and labeled data set. The generated 5G network data set [1] is also labeled for model training and testing purposes. An example of a label in the data set for prediction is the column name that represents the amount of traffic load. As for the classification, the information column stating whether the current state of the network is abnormal is an example of a label.

For these prediction and classification problems, specific ML models are used and their performances are compared. Before training ML models, the labeled data should be enriched with the feature extraction process, which is explained in Section 4.4.2, to achieve better accuracy in the results.

4.2.1. Network Load Performance Prediction

One of the essential contributions of both this thesis and NWDAF is to estimate network traffic load performance as discussed in Section 3.2. While describing the ML models in Section 4.2, it is also mentioned that traffic load performance prediction is a time series problem. In order to solve this time series problem, three different models will be used and compared. These models are Linear Regression, Recursive Neural Networks and Long-Short Term Memory. The data-set generated in Section 4.4 and feature set extracted in Section 4.4.2 are fed into these algorithms for training purposes.

4.2.1.1. Linear Regression. Linear regression is a statistical ML algorithm based on the mathematical formula of

$$y = \alpha + \beta x \,, \tag{4.1}$$

which draws a linear line on coordinate system. The coefficients A and B are fitted based on the input data x and output y. For multiple variables like the case in our study, the formula extends to

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 ... + \beta_n x_n \,, \tag{4.2}$$

where n is the number of features to explain the target in the data set. This is called multiple linear regression that aims to find the best coefficients to calculate the resulting output for every variable. Due to its simplistic nature and basic implementation, LR is generally used as a basis for comparison with other prediction algorithms. In this thesis, it is also used as a baseline model to compare the results of RNN and LSTM.

4.2.1.2. Recursive Neural Networks. Deep learning (DL) is a sub-field of ML where the concept of neural network architecture is brought to life. Neural network algorithms are capable of combining a large amount of inputs and creating a formulation by using complex relations of neurons to find out the most challenging problems with high accuracy. Combinations of these neurons in the neural network architecture are brought together to solve different sets of problems. For example, feed-forward neural network is an algorithm that learns non-linear relations between input and output. Another example is convolutional neural network, which learns spatial relationship of given input and is used mostly in computer vision. On the other hand, RNN is a neural network algorithm that can take previous data into account during its training phase, enabling it to be a powerful algorithm for forecasting and prediction problems. The RNN architecture in this study contains six layers. One starting RNN layer, four dense layers and one output layer. As the loss function, mean absolute error (MAE) is used since it is one of the performance metrics that aids evaluating algorithms in Chapter 5.

4.2.1.3. Long-Short Term Memory. LSTM is another version of neural network architecture based on RNN structure. Neural structure of LSTM enables short term previous neurons and long term previous neurons to carry information to current training phase. In other words, during training, LSTM is capable of considering the values from the beginning of the data in addition to the recent past data. The patterns in the data set can be observed by LSTM, meanwhile RNN considers only recent data. The LSTM architecture in this study contains seven layers: One starting LSTM layer, four dense layers, one dropout layer to prevent over-fitting and one output layer. Like RNN model, the loss function is MAE due to the same factors.

4.2.2. Anomaly Detection

Understanding the behaviour of the network and the UEs is another problem that requires different approaches and algorithms to solve. Providing NFs abnormal behaviour information is another task of NWDAF as discussed in Section 3.2.

One of the contributions of this thesis is to gain insights of the network and detect anomalies in the perspective of NWDAF. As depicted in Section 4.4, network load anomalies are added to generated data in order to create abnormal behaviour in the network, and the time period of spikes and extra loads are labeled as abnormal while the rest of the data is labeled as normal. To differentiate the characteristics of the network load at time t and classify them into two sections, namely, normal and abnormal, two different ML models are going to be implemented and compared. These models use logistic regression and extreme gradient boosting algorithms.

4.2.2.1. Logistic Regression. Logistic regression is a statistical ML model, which uses the sigmoid function to separate data points and make classifications. It is capable of handling multi-classifications where data is separated into more than two classes. However, in this thesis, binary logistic regression is used, since the classes are normal and abnormal.

Logistic regression has similarities with linear regression, where the coefficients of the equations are fit in order to find the best values that are close to actual ground truth. The basic formula behind logistic regression is

$$y = \frac{e^{(a+b_1x_1+b_2x_2+...+b_nx_n)}}{(1+e^{(a+b_1x_1+b_2x_2+...+b_nx_n)})}, \tag{4.3}$$

where y is predicted output, a is the bias, and b_n is the coefficients of variables that are feature inputs.

Logistic regression algorithm is used as a base model for the classification problem in this study. In addition to its basic implementation, class weight parameters are also added in order to neutralize the unfair bias of many normal data points and relatively less abnormal data points. The imbalances in the labels cause the algorithm to detect normal network traffic load condition better than abnormal traffic load condition.

4.2.2.2. Extreme Gradient Boosting. For classification problems, the first approach is to make a tree-based decision, based on the input values. There are different tree-based ML classification algorithms such as random forest algorithm. These algorithms make classifications by optimizing their loss function like DL models.

Extreme gradient boosting, namely XGBoost is an implementation of decision tree algorithm with gradient boost with performance optimization [21]. This algorithm is widely used in academic studies and in professional life, since the algorithm over-performs other traditional tree-based classification algorithms in complex problems.

As it is done for logistic regression, the class parameters for XGBoost are tuned in order to overcome the unfair bias created by the imbalanced number of labels in the data-set. By tuning this hyper-parameter, the model is set to achieve more accurate classification scores.

4.3. Topology

The proposed model is evaluated by means of the simulator we have developed. A realistic synthetic 5G data is also generated for analysis. Although the proposed model is independent of parameters such as the number of cells, user types, etc., certain parameters will be specified while defining the model for the sake of clarity and comprehensibility.

The topology consists of a set of remote radio unit (RRU) cells, a set of subscriber categories and a set of device types as UEs. The number of components and RRU cells in the topology is fixed for the sake of simplicity. The system model proposed in this thesis can support topologies with the scaled number of cells, device types and subscriber categories.

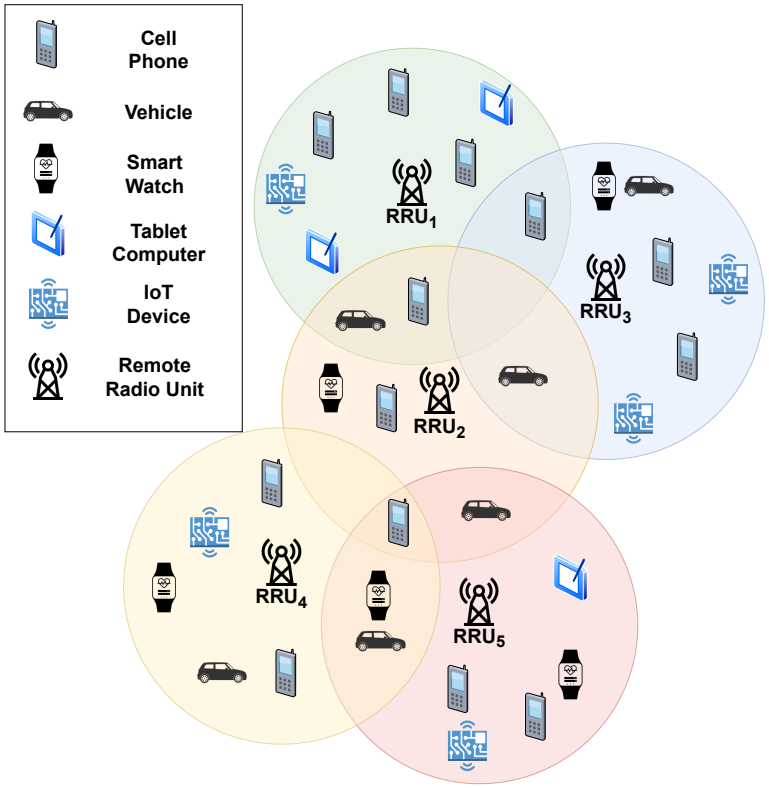


Figure 4.2. The network topology under consideration (Inspired from [5]).

We assume the network topology has five RRU cells. In each of these cells, there are users belonging to different subscriber categories, namely platinum, gold, and silver subscriptions that represent the level of subscription. The reasoning behind these three subscriptions is to make the generated data more genuine, as mobile service providers sell similar subscriptions in the actual world. In addition, there are five different types of user equipment within each subscriber category, namely, IoT device, vehicle, cell phone, smartwatch, and tablet computer. The network topology under consideration is shown in Figure 4.2.

4.4. Data Generation

In general, ML algorithms can be categorized under three different parts, namely, supervised, unsupervised, and reinforcement learning. In unsupervised learning, the machine learning algorithm is given unlabeled data, while in reinforcement learning the algorithm is based on a reward mechanism.

On top of these learning mechanisms, there is the supervised learning category where training data is also labeled with meaningful categorization. In this thesis, NWDAF in the proposed system applies supervised ML algorithms where labeled data plays an essential role. The challenging task of creating such data makes this section of the thesis important.

Overall, labeled data set is created for 5G cellular networks, and during data generation, 5G specifications defined by 3GPP [41–44] are followed to determine fields of this data set. The fields that are considered to be helpful for creating NWDAF are as follows:

- *Network area information:* Information about cells in the service field,
- *Personal equipment ID:* Device type information of each UE (e.g., cell phone, smart watch),
- *Subscription categories:* The policy of the groups subscribed to by the UEs,
- *Data rate:* Amount of transmitted data in bytes for a certain period of time.

In the process of labeled data set generation, the following parameters are defined to be used as input fields for the data generation simulation. These parameters are:

- Number of RRU cells in the topology,
- Subscriber category ID and name which are platinum, gold, and silver,
- Device Type IDs and respective names,
- Initial load parameters for each device type and subscriber category,
- Adjacency matrix of RRU cells,
- Mean handover ratios per hour for each device type group,
- Mean and variance parameters for the handover operation,
- Time step Δt, which determines the interval of data gathering duration,
- Total simulation time.

4.4.1. Network Traffic Load

According to the proposed model, each subscriber category and RRU cell contain a predefined amount of traffic load at the start of the simulation. Consequently, it can be stated that network traffic is saturated from the start to the end of the simulation.

Adjacent cells in the topology, at each time step (t) subject some percentage of their load to handover action towards a neighbour cell. In order to enable handover process, the system model has a predefined set of handover ratios, which differs depending on the time of the day, to make generated traffic more realistic.

There are some assumptions for deciding mean handover ratios for each device type. The day can be divided to several categories. Night time is the first category, and minimal mobility of devices is expected during night time. Rush hours constitute another category; low mobility of devices is expected since it would be challenging to move due to traffic jam. Lunch time is the third category since many individuals would like to go somewhere nearby to eat, causing a higher handover ratio. Before and after rush hours are the rest of the remaining categories. Before the rush hour starts and after the rush hour ends, the highest handover ratio is expected due to lower density of traffic. For the remaining categories, device types are expected to move with an average handover ratio. Furthermore, IoT devices, due to their nature, are not likely to move as often as other personal equipment we consider. Consequently, no significant difference in mean handover ratios is anticipated for the time of day.

The detailed version of anticipated handover ratios are given in Table 4.1 as mean values. As an additional note to these mean values in Table 4.1, there are also variance values. A realistic user traffic generation is desired by carefully calibrating statistical parameters using canonical approximations. In Section 4.4, the mathematical background for determining handover ratios are explained in detail.

Table 4.1. Mean Handover Ratios per Hour © 2020 IEEE [5].

Time of Day	Cell Phone	Vehicle	Tablet Computer	IoT Device	Smart Watch
00:00-06:00	2.5%	10%	1%	1%	2.5%
06:00-07:00	4.5%	18%	1.8%	1%	4.5%
07:00-09:30	3%	12%	1%	1%	3%
09:30-11:00	3.5%	14%	1.2%	1%	3.5%
11:00-13:00	4%	16%	1.5%	1%	4%
13:00-16:00	3.5%	14%	1.2%	1%	3.5%
16:00-20:00	3%	12%	1%	1%	3%
20:00-22:00	4.5%	18%	1.8%	1%	4.5%
22:00-00:00	2.5%	10%	1%	1%	2.5%

Table 4.2 shows the starting load settings for each RRU cell at the start of the data generation simulation, based on a subscriber category and a personal equipment type. As can be seen, various amounts of loads are assigned to each subscriber category group and personal equipment type. The idea behind these values is that a cell phone user is more likely to get the highest subscription while IoT and vehicle device types are less likely to have a subscription to the premium category since they would not demand high performing network connection all the time. Additionally, the RRU cell which is adjacent to all other cells as seen in Figure 4.2, has the mean handover ratios twice the values in Table 4.1 to preserve network balance. The handover ratios in Table 4.1 are mean values, as previously stated. The mean and variance parameters are given as follows, assuming that the handover events have a Gaussian distribution:

$$\Delta H_{\text{ratio}} \sim \mathcal{N}(\mu, \frac{\mu}{8}) \,, \tag{4.4}$$

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

Network traffic data is generated for six months period. The data contains a network snapshot taken every $\Delta t = 15$ minutes. A UE may handover between adjacent cells during each of these Δt intervals.

Table 4.2. Initial Traffic Loads for Device Types © 2020 IEEE [5].

Subscriber Category (ID)	Cell Phone	Vehicle	Tablet Computer	IoT Device	Smart Watch
Platinum (1)	90 Gbps	20 Gbps	6 Gbps	3 Gbps	1 Gbps
Gold (2)	72 Gbps	18 Gbps	5 Gbps	4 Gbps	1 Gbps
Silver (3)	53 Gbps	16 Gbps	5 Gbps	5 Gbps	1 Gbps

Anomalies are added to the generated network traffic data throughout the simulation to make our data set more realistic. Anomalies are defined in this study as large amounts of network traffic diverging from the average network behaviour, where network traffic load peaks, then, fades and stabilizes over time. We are inspired to create these kinds of anomalies by our daily lives, where videos go viral on a regular basis or breaking news occurs. In fact, both of these factors have an increasing impact on network traffic data.

While generating the anomalies, it is important to label the time period where the unexpected traffic load is presented to network. In order to know the ground truth for machine learning testing and understand the behaviour of anomalies by analysis, labeling anomaly periods is an important necessary task.

The coding implementation steps of the data generation are taken according to the object oriented paradigm. There are model objects defining RRU cell, device type, and subscriber category. Each device type's information, including loads and statistics, is stored in the subscriber category object model. Handover actions are handled by the device type object model, which also retains load information at the present time point. Lastly, RRU cell object model contains adjacent cell matrix and network topology information as mentioned in Section 4.3.

After creating all the models for data generation process, the proposed system model is generated by using required predefined values, which are total simulation time, handover parameters and ratios for each device types, adjacency matrix for RRU cells, and percentage of abnormalities in the network traffic load.

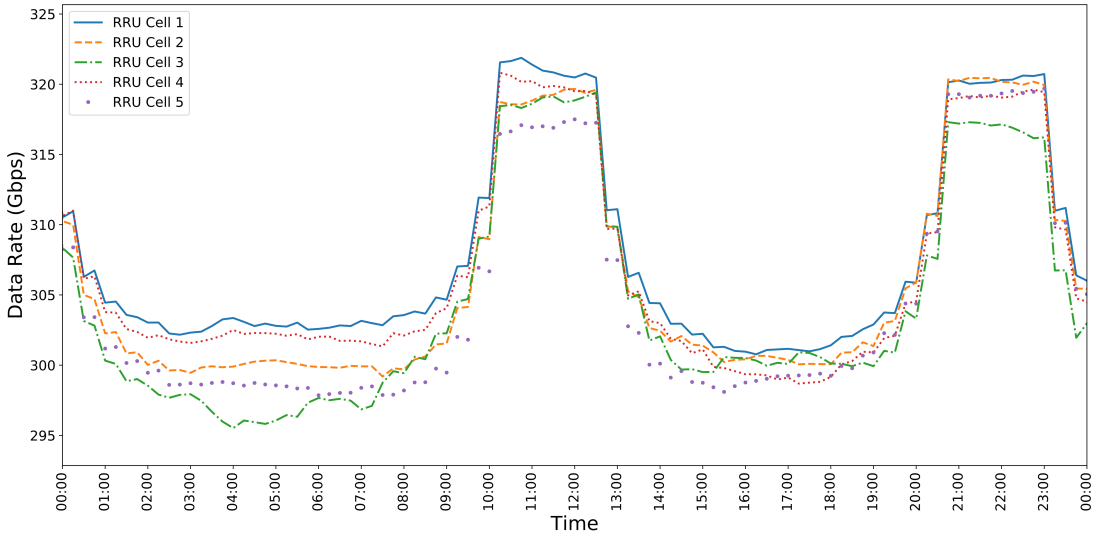


Figure 4.3. Data rate per cell for a sample day-Generated Data Set © 2020 IEEE [5].

The following events happen at each simulation time step t (i.e., 0, Δt, $2 \times \Delta t$, ..., t). Each device type model indicates how much load will be transferred for handover. The handover process occurs by transferring the network load for the particular device type from the source RRU cell to the target RRU cell. Furthermore, the start and end times of anomalies are determined at random before the data production simulation begins. When an anomaly period begins in the simulation, a preset percentage of the network load, which increases exponentially, is added to each device. Afterwards, the same amount of preset percentage of the network load is removed from each device with exponential fading. After the data generation procedure is completed, all network load data is exported in order to evaluate and generate appropriate features, as stated in Section 4.4.2. The aggregated data rates of each cell are depicted in Figure 4.3.

4.4.2. Feature Extraction

A good ML model should give accurate predictions when compared with the ground truth. In order to enable ML algorithm to cover the different cases of data set, for example, with high and low traffic load, some indicators of the situation should be added as an input to the algorithm. The detecting and creating indicator process is called feature extraction.

Temporal and spatial data analysis, evaluating the plots of data and some human thoughts are the ways of determining new features from data where it would boost an ML algorithm during its training phase. The sub-field of ML algorithms, DL algorithms, are more flexible in terms of feature. The neural network node architecture in a DL algorithm helps itself by detecting correlations and using them as input to its neural layers. However, due to DL's randomized nature, it is a safe choice to feed all DL algorithms with extracted feature set together with the input data set. Additionally, in this study all ML algorithms are fed with the same feature set in order to establish fair comparison for results evaluation in Section 5.

During data generation, as depicted in Section 4.4, not only traffic load but different features such as subscriber category are added as well. In addition to these basic features, it is required to extract features depending on traffic load in order to understand the previous time step during time step Δt. The generated features are as follows:

- *last2_mean*: Rolling average of the data rate during last two Δt steps.
- *last4_mean*: Rolling average of the data rate during last four Δt steps.
- *last8_mean*: Rolling average of the data rate during last eight Δt steps.
- *per_change_last2*: Percentage of the data rate difference between last two Δt.
- *per_change_last3*: Percentage of the data rate difference between $t - \Delta t$ and $t - 3 \times \Delta t$.
- *per_change_last4*: Percentage of the data rate difference between $t - \Delta t$ and $t - 4 \times \Delta t$.
- *change_last2*: Data rate difference between last two Δt.
- *change_last3*: Data rate difference between $t - \Delta t$ and $t - 3 \times \Delta t$.
- *change_last4*: Data rate difference between $t - \Delta t$ and $t - 4 \times \Delta t$.

The next step after feature generation is to check the importance of and correlation between features. This phase is called feature elimination. Feature importance test and correlation test are performed to eliminate highly correlated features and highlight the important features in order to prevent unnecessary training time for ML models and avoid overfitting by presenting the correlated features as a double input to ML algorithm.

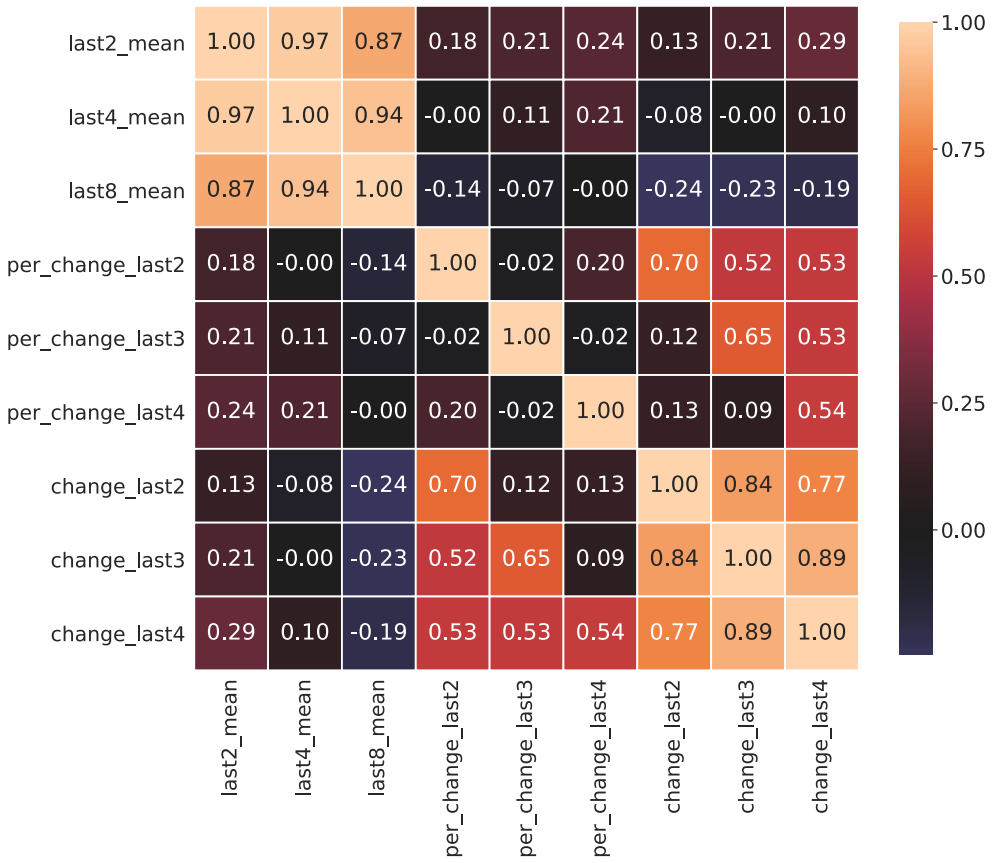


Figure 4.4. Correlation matrix of the extracted features © 2020 IEEE [5].

In this study feature importance score is the coefficients of a linear regression model, which is fitted to our data-set including extracted features. Linear regression fitting is performed for each feature that are present in data set separately to find out the feature which is best at explaining traffic load.

This test indicates that the most important features are *last2_mean*, *last4_mean*, and *last8_mean* fields, afterwards *per_change_last4*, *per_change_last3*, *per_change_last2* comes, then finally *change_last2*, *change_last3* and *change_last4*.

The correlation test results are shown in Figure 4.4. Percentage change features with different rolling averages have the lowest correlation compared to standard rolling average features and data rate difference features. Considering the results of both feature selection tests, only the *last2_mean* and all percentage change features with different rolling averages are selected as an input for ML algorithms, on the other hand, data rate difference features are eliminated due to low feature importance scores.

5. EXPERIMENTS AND RESULTS

5.1. Experiments

In this study, NWDAF and certain events that NFs can request from this function are focused on. In order to validate the capability and performance of this function, a realistic data set is generated. The algorithms depicted in Section 4.2 are implemented and trained to produce the results as NWDAF definition requires. Moreover, another data set [2] created by Sone *et al.* is used for comparison with the generated network traffic. The data set consists of the measurements from a local enterprise network established in the campus of University of Oulu, Finland. About two months of recorded data is collected from 470 different APs containing information of a number of users of time t, received and transmitted bytes. In [25], the authors used the data from four different APs, which have relatively high traffic compared to others. The aggregated data rate for these APs are shown in Figure 5.1 The network topology explained in Section 4.3 has 5 RRUs; similarly applying the ML approaches in NWDAF to these four APs makes a fair comparison.

The experiments conducted in this thesis are two fold. The first one is network traffic load prediction using LR, RNN and LSTM, using both the generated data set and real life data set. The second one is the anomaly detection where the abnormalities in the data are investigated and predicted as normal or abnormal. For anomaly detection experiment, only the generated data set can be used, since labeled data is required to perform ML training.

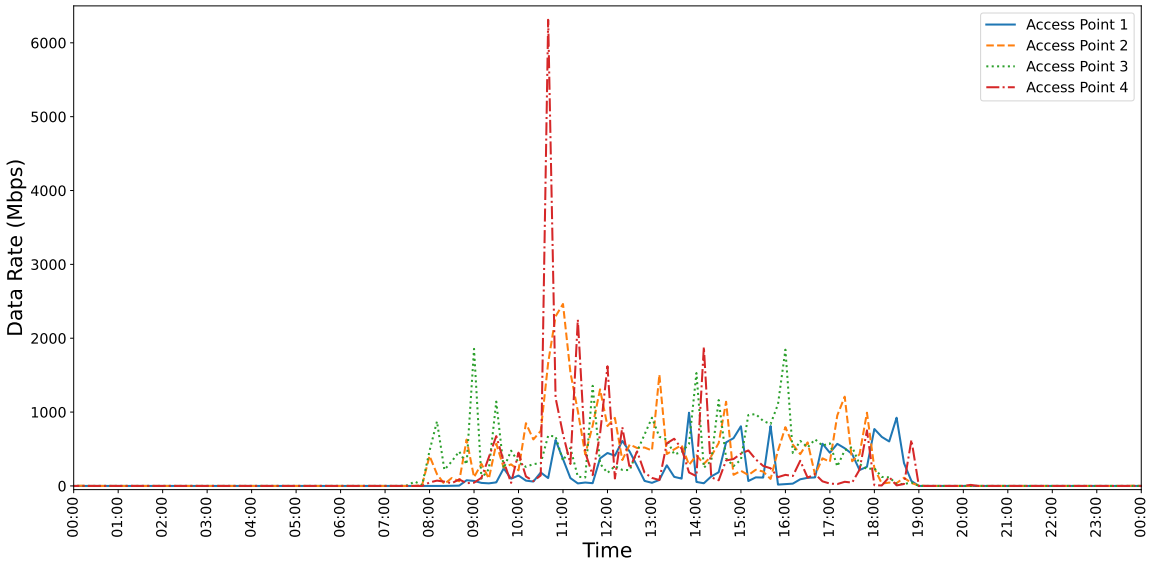


Figure 5.1. Data rate per AP for a sample day - Real Life Data Set [2].

5.2. Results

The experiments detailed in Section 5.1 are conducted using the two data sets, one generated as described in Section 4.4 and the other from [2]. The results are obtained and explained in two sections where in the first one, network load prediction performance is discussed with two different data sets being used to conduct the experiment and in the second, network anomaly detection is investigated by using the generated data set together with ML classifier algorithms.

5.2.1. Performance Metrics of Network Load Prediction

For network traffic load prediction experiment, two performance metrics are used, namely mean absolute percentage error (MAPE) and mean absolute error (MAE). MAE indicates how much the model prediction results actually differ from the true values in terms of the prediction unit. When the amount of data rate is smaller, it is easier to evaluate model performance, since the formula gives a direct relationship between the actual value and the predicted value.

In this study, MAE formula is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{\mathbf{y}}_i - \mathbf{y}_i| \,, \tag{5.1}$$

where N is the number of predictions, $\hat{y}$ is the real value and y is the model prediction.

MAPE is a performance metric based on MAE, where the predicted values can be correlated with actual values in such a way that the margin of error is represented by the percentage of actual value. This metric allows one to understand the error in a scaled way when the actual values consist of big numbers. In this study, MAPE formula is defined as

$$\mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N} |\frac{\mathbf{y}_i - \hat{\mathbf{y}}_i}{\mathbf{y}_i}| \,, \tag{5.2}$$

where N is the number of predictions, y is the real value and $\hat{y}$ is the model prediction.

These performance metrics explain the accuracy of model predictions. The experiment done with the real life data set is evaluated only with MAE metric whereas the experiment with the generated data is evaluated by both performance metrics. The underlying reason of this discrepancy is that in the real life data set, there are many data points where there is almost no data transmission since the university campus is closed to most of the students during non-working hours. MAPE is a metric, which does not work with zero data points, since it contains a division with the actual value. In addition to this reason, MAPE is a misleading metric when it comes to evaluating the margin of error with small numbers since it shows the error as magnitude of actual value. Due to these facts, only MAE performance is calculated.

5.2.2. Network Traffic Load Prediction Performance

Several ML models are generated by using the algorithms detailed in Section 4.2 for time series estimation. For all subscriber categories and each cell in the generated data set, a model is specifically trained. Similarly, for each access point in the real life data set, four more models are trained in order for NWDAF to make accurate assessments over the network. The performance results for generated data set are given in Tables 5.1 and the results from real life data set training are given in Table 5.2.

In Table 5.1, cell ID stands for the RRU cell number as shown in Figure 4.2 and SubsCat stands for the subscriber categories as enumerated in Table 4.2 and named platinum, gold and silver, respectively.

When MAE and MAPE metrics are compared in Table 5.1, it can be seen that LR has poor performance over LSTM and RNN. Additionally, in most of the cases, RNN performs slightly better than LSTM, which can be explained by the random initialization of deep neural networks. Average results of LSTM and RNN in terms of MAPE and MAE indicate different winners. MAPE results of LSTM outperforms RNN's results, yet MAE results of RNN outperforms LSTM's results. The underlying reason for this difference is the fact that LSTM and RNN work with different logics behind the curtain while taking the historical information into account. Since RNN disregards the seasonal structure of the time series data, it can perform better when it comes to detecting abnormal and unexpected network conditions. On the other hand, LSTM considers the historical information and performs well when it comes to detecting steady and seasonal parts of the time series. The abnormal data rates are presented as high data rate spikes in the generated data, the error of RNN is higher compared to LSTM, according to MAPE, since in the MAPE calculations, high value in the denominator yields low error score even if the absolute error between the predicted value and actual value is high.

Table 5.1. Network Load Prediction Performance - Generated Data Set © 2020 IEEE [5].

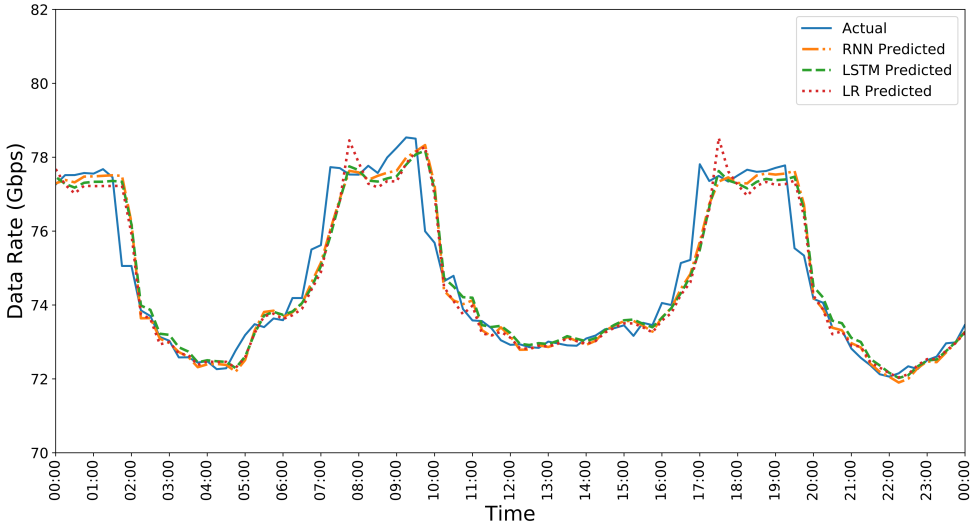
Metric Name	(Cell - SubsCat) ID	LR	LSTM	RNN
MAPE (%)	1 - 1	0.577	0.504	0.512
	1 - 2	0.575	0.512	0.573
	1 - 3	0.579	0.511	0.498
	2 - 1	0.578	0.510	0.499
	2 - 2	0.576	0.510	0.521
	2 - 3	0.585	0.504	0.519
	3 - 1	0.761	0.680	0.735
	3 - 2	0.757	0.688	0.754
	3 - 3	0.750	0.696	0.735
	4 - 1	0.581	0.507	0.487
	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 (Mbs)	1 - 1	189.4	160.7	151.5
	1 - 2	242.6	209.92	238.5
	1 - 3	296.9	257.0	222.2
	2 - 1	188.4	161.7	142.3
	2 - 2	243.7	206.8	196.6
	2 - 3	297.9	247.8	224.2
	3 - 1	228.7	200.7	208.8
	3 - 2	288.7	258.0	275.4
	3 - 3	347.1	319.4	314.3
	4 - 1	189.4	160.7	138.2
	4 - 2	243.7	209.9	184.3
	4 - 3	295.9	248.8	223.2
	5 - 1	189.4	162.8	150.5
	5 - 2	243.7	208.8	201.7
	5 - 3	295.9	248.8	219.9
	Average	**252.9**	**218.1**	**206.8**

Table 5.2. Network Load Prediction MAE Performance - Real Life Data Set [2].

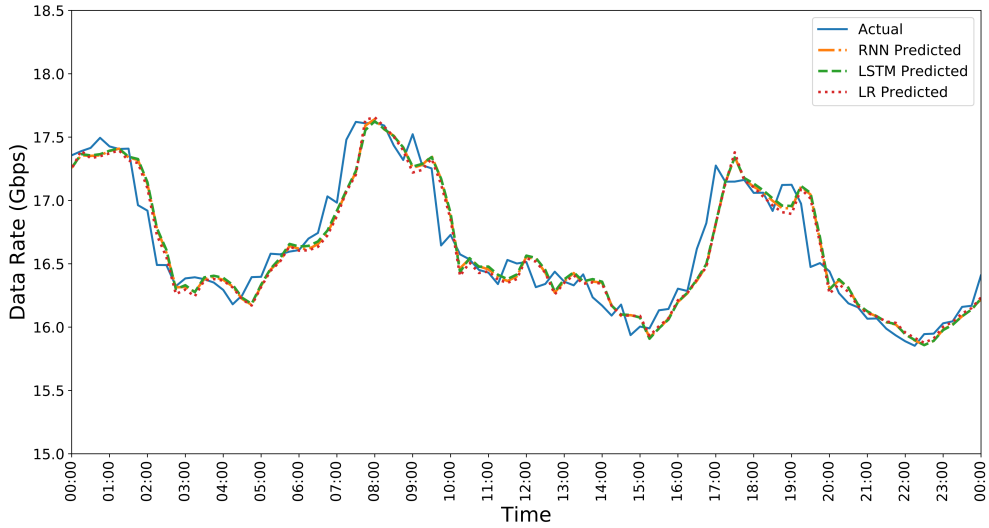
Metric Name	Access Point ID	LR	LSTM	RNN
MAE (Mbs)	1 - ap184074	83.8	78.3	84.4
	2 - ap184149	82.4	78.1	82.4
	3 - ap184202	101.7	73.2	74.7
	4 - ap185135	189.8	177.2	207.6
	Average	**114.4**	**101.7**	**112.2**

The snapshot of the ML prediction results, which was taken from on a random day during the simulation, is shown in Figure 5.2. The comparison of Figure 5.2(a) and 5.2(b), which represents different UE types, subscriber categories, and RRU cells, allows us to come to the following conclusions. First of all, in both figures, LR predictions are less accurate compared to the other models. As for other models, RNN prediction results are closer to LSTM's outcomes for unsteady data rates as it can be observed between 07:00 and 10:00 in Figure 5.2(a) and between 00:00 and 01:30 in Figure 5.2(b). Similarly, LSTM prediction results are more accurate compared to RNN's outcomes for steady data rates as it can be observed between 11:00 and 16:00 in Figure 5.2(a). Since vehicle device type has the highest mean handover ratio as depicted in Table 4.1, Figure 5.2(b) is less steady compared to Figure 5.2(a).

When it comes to the prediction results of the real life data set in Table 5.2, a different conclusion can be made about the performance of the algorithms when it is compared to MAE scores of generated data set results in Table 5.1. Unlike the performance results of the algorithms trained with the generated data set, LR and RNN perform very similarly. However, LSTM outperforms the others with significantly lower error. This result indicates that the real data set carries steady data rates compared to the generated data set. As shown in Figure 5.3, the spikes in the data set occur with a pattern. In fact, approximately every hour there is a sharp increase in the data rate. The pattern of an increase in the data rate during lunch time every day of week, contains general information about the historical data. Thus, LSTM by its nature is able to use it to make predictions with better accuracy.



(a) RRU_3, UE type is cell phone, and SubsCat is gold.



(b) RRU_4, UE type is vehicle, and SubsCat is platinum.

Figure 5.2. Time versus data rate for a sample day with different AI/ML model predictions - Generated Data Set © 2020 IEEE [5].

Overall, the analysis of the given results allows us to conclude that LR performs reasonably well with both of our data sets. However, in busier and wider networks which have more complex patterns in terms of data rate, LR is likely to provide predictions with poor performance due to its simplistic nature.

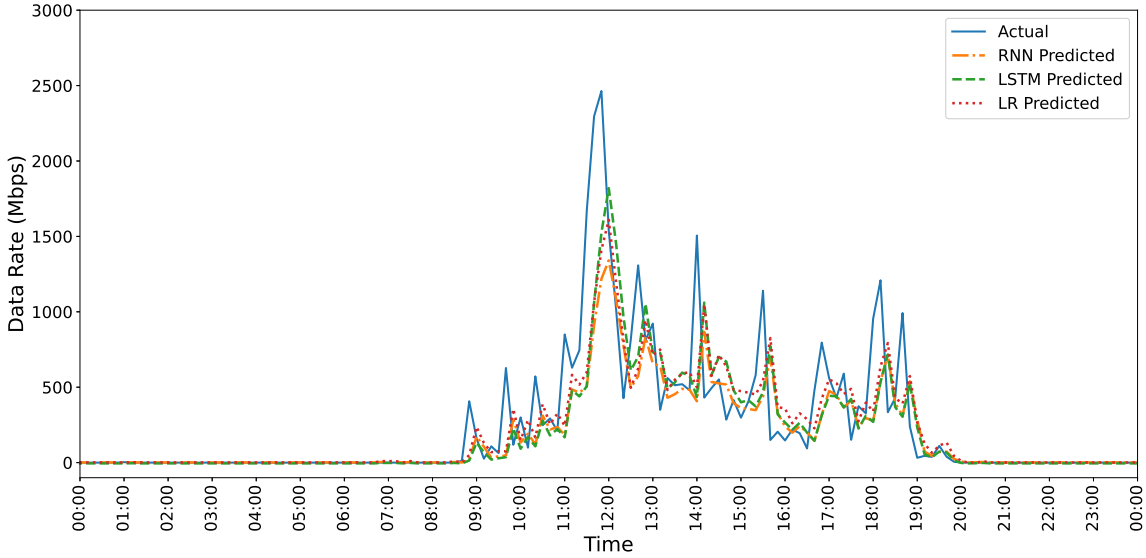


Figure 5.3. Time versus data rate for a sample day with different AI/ML model predictions - Real Life Data Set, AP 2 [2].

Moreover, the performance of more complex ML algorithms such as LSTM and RNN highly depends on the nature of data set. Nevertheless, both of them are efficient predictors compared to baseline model, LR.

5.2.3. Performance Metrics of Anomaly Detection

On the baseline, anomaly detection is a classification problem where ML algorithm tries to distinguish the properties of network with abnormalities from the normal state of the network. In order to measure the performance of classification problem there are several metrics that are used in the literature. Area under the receiver operating characteristics curve (AUROC) is used as a performance metric of anomaly detection in this analysis, and it compares the true positive rate (i.e., sensitivity) P_d and false positive rate P_r. Sensitivity is defined as follows:

$$P_d = \frac{TP}{TP + FN}, \tag{5.3}$$

where TP is the number of true positives, and FN is the number of false negatives acquired from prediction results.

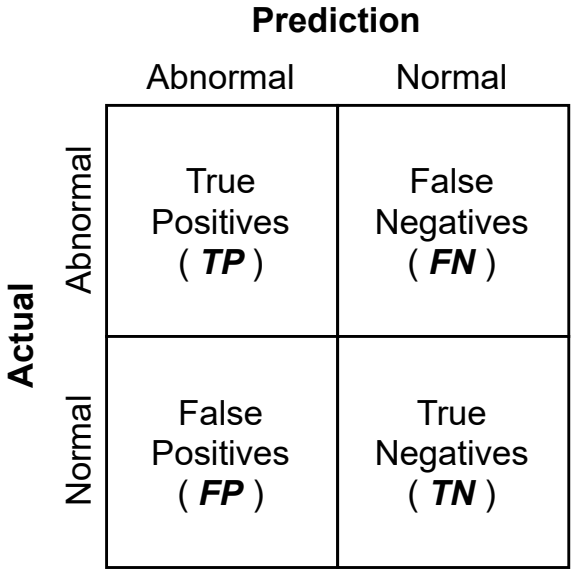


Figure 5.4. Confusion Matrix for Anomaly Detection.

This formula gives the ratio of true positives versus actual true values. Successful predictions of the model for detecting anomalies are represented by P_d. Similarly, the false positive rate can be written as

$$P_r = \frac{FP}{FP + TN}\,, \tag{5.4}$$

where FP is the number of false positives, and TN is the number of true negatives acquired from prediction results. This formula gives the ratio of false positives versus actual false values. Failed predictions of the model are represented by P_r. By using the predictions of ML algorithms the receiving operator characteristics (ROC) curve, which is represented by P_d and P_r, is visualized and the area under curve is calculated. Models that are good at classifications create a ROC curve shape that looks like an elbow going upwards and turning right.

Accuracy and precision are the other metrics used to evaluate performance of classification algorithms. Accuracy is a ratio of correct predictions to the total number of predictions whereas precision is the ratio of correct predictions of positive values to the total positive predictions.

Accuracy may not be an indicator of a good performance when imbalanced number of labels in data are present; however, it is calculated for the sake of comparison. Precision, on the other hand, relates to false positive rate, which is the component of ROC. In other words, high precision indicates a good performance.

The formula of accuracy and precision are

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}\,, \tag{5.5}$$

$$\text{Precision} = \frac{TP}{TP + FP}\,, \tag{5.6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false negatives, and TN is the number of true negatives. The meanings of these definitions are depicted in Figure 5.4.

5.2.4. Anomaly Detection

To detect anomalies in the generated data set, two ML algorithms are used namely logistic regression and XGBoost. These two models are trained and fitted, their performance metrics, namely AUROC , accuracy and precision metrics are calculated as shown in Table 5.3. By analyzing the results given in Table 5.3, it can be clearly stated that XGBoost model outperforms the baseline model which is logistic regression. For each cell and subscriber categories, XGBoost prediction scores are higher, especially it has significantly improved in accuracy score.

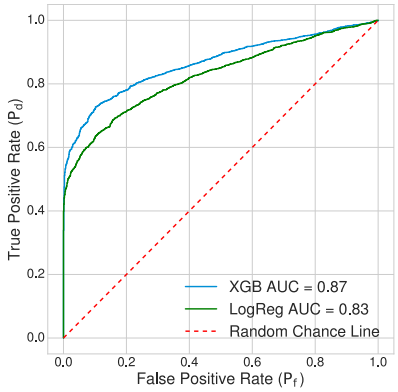
AUROC figures are calculated and are shown in Figure 5.5. The performance of models that are trained for RRU_3 and RRU_4, the cell phone device type and each subscriber category are compared. Comparing Figure 5.5(a) with Figure 5.5(b), Figure 5.5(c) with Figure 5.5(d), and Figure 5.5(e) with Figure 5.5(f), the area under ROC curve for XGBoost model predictions are significantly higher.

Table 5.3. Average Results for Anomaly Predictions (Averaging is Done Over Device Types) © 2020 IEEE [5].

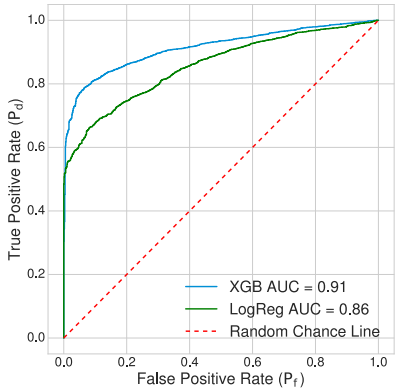
		Logistic Regression			XGBoost		
Cell ID	SubsCat	AUROC	Accuracy	Precision	AUROC	Accuracy	Precision
1	Platinum	88.0%	55.4%	77.8%	91.5%	63.4%	77.5%
1	Gold	87.4%	56.0%	77.4%	91.5%	63.5%	77.9%
1	Silver	87.6%	55.3%	77.4%	91.7%	63.6%	78.0%
2	Platinum	87.3%	55.5%	77.0%	91.4%	63.3%	77.6%
2	Gold	87.6%	55.7%	77.5%	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%**

Considering the network topology for generated data explained in Section 4.3, RRU_3 has a challenging characteristic compared to the other cells due to being the center node in the ecosystem where a high number of handovers occur. Due to this reason, XGBoost performs poorly compared to its performance scores for RRU_4. On the other hand, logistic regression, as a baseline model in this anomaly detection problem, is outperformed by its competitor in all comparisons, which is expected due to its less complex algorithm failing to fit to the generated data set.

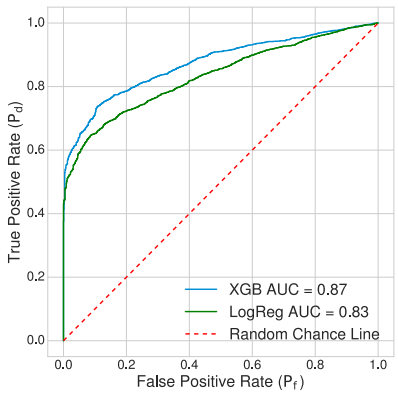
In addition to the conclusions derived from Figure 5.5, the ROC curves in the subfigures belonging to the same cell and different subscriber categories do not vary significantly. This fact explains that subscriber categories do not play a role worthy of attention while detecting anomalies in the network.



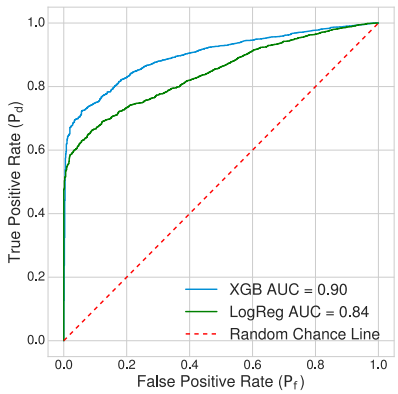
(a) RRU_3, UE type is cell phone, and SubsCat is silver.



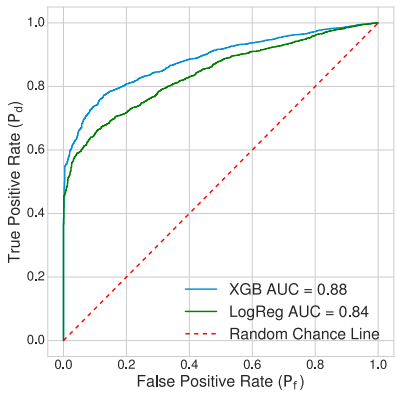
(b) RRU_4, UE type is cell phone, and SubsCat is silver.



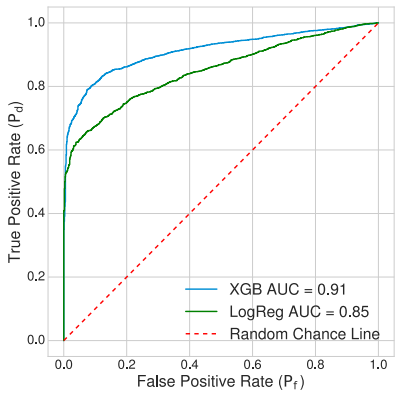
(c) RRU_3, UE type is cell phone, and SubsCat is gold.



(d) RRU_4, UE type is cell phone, and SubsCat is gold.



(e) RRU_3, UE type is cell phone, and SubsCat is platinum.



(f) RRU_4, UE type is cell phone, and SubsCat is platinum.

Figure 5.5. False positive rate versus true positive rate for logistic regression and XGBoost models (AUROC) © 2020 IEEE [5].

6. CONCLUSION

The paradigm shift 5G brings over 4G created an opportunity for new technological advancements to take place in the next generation wireless networking. With increasing scale of cellular networks, an intelligent and adaptive mechanism became vital to manage a network with a set of high standards.

In this thesis, a novel system is presented to obtain an intelligent decision mechanism for network data analytics for 5G networks. NWDAF, which is one of the newly introduced NFs in 5G SBA, is described, and several ML algorithms are trained to fulfill the responsibility of this function partially. Firstly, data sets are analyzed and several features are extracted to assist the performance of ML techniques. Then, network load performance is predicted using LR, LSTM, and RNN algorithms, which are commonly used for time series problems. Afterwards, an anomaly detection algorithm is implemented by using logistic regression and a tree-based classifier, XGBoost. Moreover, a data set generation methodology is described by using the fields defined in 3GPP standards, in order to assess the network data analytics in 5G.

As a conclusion, network load prediction experiments indicate that neural network algorithms outperform linear regression predictor. Specifically, for real life data set [2], LSTM performs better than its competitors significantly whereas for generated data set [1], both neural network algorithms perform well depending on the steadiness of the time period. On the other side of the coin, tree-based XGBoost modal outperforms logistic regression for anomaly detection in the network. To conclude, a very practical usage of NWDAF by using popular and common AI/ML models is shown.

Due to freshness of NWDAF, there are not many researches covering this topic in the literature. Plenty of practical ideas for NWDAF can be put on the table, to widen its application field.

One idea is to make anomaly detection a multi-label classification problem instead of binary, where the network status can be separated into more than two sections to help operators to understand the seriousness level of anomaly in the network. A second idea is enabling NWDAF as an edge NF, namely NWDAF agent, where it can analyze the network information based of the location of the cells, which are connected to that edge server, enabling NWDAF to recognize local patterns in the network. In addition to the second idea, if NWDAF requests and consumes network analytics information from NWDAF agents, it would aid operators and central NWDAF to make more accurate predictions by using the analytics focused on a limited area in the field. When it comes to the generated data set, more fields such as network slice information can be added to study another subsystem of NWDAF. Lastly, NWDAF subsystems can be experimented by using different AI/ML algorithms and training techniques in order to improve the performance while maintaining the time constraints.

REFERENCES

1. Sevgican, S., M. Turan, K. Gökarslan, H. B. Yilmaz and T. Tugcu, *Synthetic 5G Cellular Network Data for NWDAF*, 2019, `https://github.com/sevgicansalih/nwdaf_data`, accessed in December 2019.

2. Sone, S. P., J. Lehtomäki and Z. Khan, "Wireless Network Traffic Time Series of an Enterprise Network", *IEEE Dataport*, 2020.

3. 3GPP, *Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services*, Technical Specification (TS 23.288), 3rd Generation Partnership Project (3GPP), September 2021, version 16.8.0.

4. 3GPP, *5G System; Network Data Analytics Services; Stage 3*, Technical Specification (TS 29.520), 3rd Generation Partnership Project (3GPP), September 2021, version 16.9.0.

5. Sevgican, S., M. Turan, K. Gökarslan, H. B. Yilmaz and T. Tugcu, "Intelligent Network Data Analytics Function in 5G Cellular Networks Using Machine Learning", *Journal of Communications and Networks*, Vol. 22, No. 3, pp. 269–280, 2020.

6. Gupta, A. and R. K. Jha, "A Survey of 5G Network: Architecture and Emerging Technologies", *IEEE Access*, Vol. 3, pp. 1206–1232, 2015.

7. Agiwal, M., A. Roy and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey", *IEEE Communications Surveys & Tutorials*, Vol. 18, No. 3, pp. 1617–1655, 2016.

8. CISCO, *Cisco Annual Internet Report (2018-2023)*, 2020, `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`, accessed in December 2021.

9. Shariatmadari, H., 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 Communications Magazine*, Vol. 53, No. 9, pp. 10–17, 2015.

10. Saad, W., M. Bennis and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems", *IEEE Network*, Vol. 34, No. 3, pp. 134–142, 2020.

11. Letaief, K. B., W. Chen, Y. Shi, J. Zhang and Y.-J. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks", *IEEE Communications Magazine*, Vol. 57, No. 8, pp. 84–90, 2019.

12. Yang, P., Y. Xiao, M. Xiao and S. Li, "6G Wireless Communications: Vision and Potential Techniques", *IEEE Network*, Vol. 33, No. 4, pp. 70–75, 2019.

13. Ali, S., W. Saad, N. Rajatheva, K. Chang, D. Steinbach, B. Sliwa, C. Wietfeld, K. Mei, H. Shiri, H.-J. Zepernick, T. M. C. Chu, I. Ahmad, J. Huusko, J. Suutala, S. Bhadauria, V. Bhatia, R. Mitra, S. Amuru, R. Abbas, B. Shao, M. Capobianco, G. Yu, M. Claes, T. Karvonen, M. Chen, M. Girnyk and H. Malik, "6G White Paper on Machine Learning in Wireless Communication Networks", *arXiv preprint arXiv:2004.13875*, 2020.

14. Andrews, J. G., S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong and J. C. Zhang, "What Will 5G Be?", *IEEE Journal on Selected Areas in Communications*, Vol. 32, No. 6, pp. 1065–1082, 2014.

15. Ekström, H., *Non-standalone and Standalone: Two Standards-Based Paths to 5G*, 2019, `https://www.ericsson.com/en/blog/2019/7/standalone-and-non-standalone-5g-nr-two-5g-tracks`, accessed in December 2021.

16. 3GPP, *System Architecture for the 5G System (5GS)*, Technical Specification (TS 23.501), 3rd Generation Partnership Project (3GPP), September 2021, version

16.10.0.

17. 3GPP, *5G System; Unified Data Management Services; Stage 3*, Technical Specification (TS 29.503), 3rd Generation Partnership Project (3GPP), September 2021, version 16.9.0.

18. Hernandez-Chulde, C. and C. Cervello-Pastor, "Intelligent Optimization and Machine Learning for 5G Network Control and Management", *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pp. 339–342, June 2019.

19. Boccardi, F., R. W. Heath, A. Lozano, T. L. Marzetta and P. Popovski, "Five Disruptive Technology Directions for 5G", *IEEE Communications Magazine*, Vol. 52, No. 2, pp. 74–80, 2014.

20. 3GPP, *Study on Latency Reduction Techniques for LTE*, Technical Report (TR 36.881), 3rd Generation Partnership Project (3GPP), July 2016, version 14.0.0.

21. Chen, T. and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 785–794, ACM, 2016.

22. Barmpounakis, S., 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 Journal on Advances in Telecommunications*, Vol. 11, No. 3, pp. 101–114, 2018.

23. Klautau, A., 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", *2018 Information Theory and Applications Workshop, San Diego*, pp. 1–1, 2018.

24. Koksal, B., R. Schmidt, X. Vasilakos and N. Nikaien, "CRAWDAD Dataset eure-

com/elasticmon5G2019 (v. 2019-08-28)", *Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD)*, August 2019.

25. Sone, S. P., J. J. Lehtomäki and Z. Khan, "Wireless Traffic Usage Forecasting Using Real Enterprise Network Data: Analysis and Methods", *IEEE Open Journal of the Communications Society*, Vol. 1, pp. 777–797, 2020.

26. Shafin, R., 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", *IEEE Wireless Communications*, Vol. 27, No. 2, pp. 212–217, 2020.

27. Jiang, C., H. Zhang, Y. Ren, Z. Han, K.-C. Chen and L. Hanzo, "Machine Learning Paradigms for Next-Generation Wireless Networks", *IEEE Wireless Communications*, Vol. 24, No. 2, pp. 98–105, 2016.

28. Casellas, R., 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", *International Conference on Transparent Optical Networks (ICTON)*, pp. 1–4, July 2018.

29. Moysen, J. and L. Giupponi, "From 4G to 5G: Self-Organized Network Management Meets Machine Learning", *Computer Communications*, Vol. 129, pp. 248–268, 2018.

30. Chen, M., U. Challita, W. Saad, C. Yin and M. Debbah, "Artificial Neural Networks-Based Machine Learning for Wireless Networks: A Tutorial", *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 4, pp. 3039–3071, 2019.

31. Sun, Y., M. Peng, Y. Zhou, Y. Huang and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues", *IEEE Communications Surveys Tutorials*, Vol. 21, No. 4, pp. 3072–3108, 2019.

32. Fang, H., X. Wang and S. Tomasin, "Machine Learning for Intelligent Authenti-

cation in 5G and Beyond Wireless Networks", *IEEE Wireless Communications*, Vol. 26, No. 5, pp. 55–61, 2019.

33. Zhang, C., P. Patras and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey", *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 3, pp. 2224–2287, 2019.

34. Asadi, A., S. Müller, G. H. Sim, A. Klein and M. Hollick, "FML: Fast Machine Learning for 5G MmWave Vehicular Communications", *IEEE Conference on Computer Communications (INFOCOM)*, pp. 1961–1969, 2018.

35. Ye, H., L. Liang, G. Ye Li, J. Kim, L. Lu and M. Wu, "Machine Learning for Vehicular Networks: Recent Advances and Application Examples", *IEEE Vehicular Technology Magazine*, Vol. 13, No. 2, pp. 94–101, 2018.

36. Cheng, N., F. Lyu, J. Chen, W. Xu, H. Zhou, S. Zhang and X. S. Shen, "Big Data Driven Vehicular Networks", *IEEE Network*, Vol. 32, No. 6, pp. 160–167, 2018.

37. Klautau, A., P. Batista, N. González-Prelcic, Y. Wang and R. W. Heath, "5G MIMO Data for Machine Learning: Application to Beam-Selection Using Deep Learning", *IEEE Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018.

38. Gao, X., L. Dai, Y. Sun, S. Han and I. Chih-Lin, "Machine Learning Inspired Energy-Efficient Hybrid Precoding for MmWave Massive MIMO Systems", *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2017.

39. Bai, L., 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", *Wireless Communications and Mobile Computing*, Vol. 2018, 2018.

40. 3GPP, *Technical Specification Group Services and System Aspects; Service Re-*

quirements for the 5G System, Technical Specification (TS 22.261), 3rd Generation Partnership Project (3GPP), September 2021, version 18.40.0.

41. 3GPP, *5G System; Usage of the Unified Data Repository Service for Policy Data, Application Data and Structured Data for Exposure; Stage 3*, Technical Specification (TS 29.519), 3rd Generation Partnership Project (3GPP), September 2021, version 16.8.0.

42. 3GPP, *5G System; Network Exposure Function Northbound APIs; Stage 3*, Technical Specification (TS 29.522), 3rd Generation Partnership Project (3GPP), September 2021, version 16.9.0.

43. 3GPP, *5G System; Background Data Transfer Policy Control Service; Stage 3*, Technical Specification (TS 29.554), 3rd Generation Partnership Project (3GPP), June 2021, version 16.7.0.

44. 3GPP, *T8 Reference Point for Northbound APIs*, Technical Specification (TS 29.122), 3rd Generation Partnership Project (3GPP), September 2021, version 16.11.0.

APPENDIX A: COPYRIGHT PERMISSION GRANTS

For [5], in reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Boğaziçi University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

© 2020 IEEE. Reprinted, with permission, from S. Sevgican, M. Turan, K. Gökarslan, H. B. Yilmaz and T. Tugcu, "Intelligent Network Data Analytics Function in 5G Cellular Networks Using Machine Learning", Journal of Communications and Networks, June 2020.

For [5], requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis: In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.