Combining Network Data Analytics Function and Machine Learning for Abnormal Traffic Detection in Beyond 5G

Abdelkader Mekrache  
EURECOM  
Sophia Antipolis, France  
abdelkader.mekrache@eurecom.fr

Karim Boutiba  
EURECOM  
Sophia Antipolis, France  
karim.boutiba@eurecom.fr

Adlen Ksentini  
EURECOM  
Sophia Antipolis, France  
adlen.ksentini@eurecom.fr

Abstract—The Network Data Analytics Function (NWDAF) is a key component of the 5G Core Network (CN) architecture whose role is to generate analytics and insights from the network data to accommodate end users and improve the network performance. NWDAF allows the collection, processing, and analysis of network data to enable a variety of applications, such as User Equipment (UE) mobility analytics and UE abnormal behaviour. Although defined by 3GPP, realizing these applications is still an open problem. To fill this gap: (i) we propose a microservices architecture of NWDAF to plug the 3GPP applications as microservices enabling greater flexibility and scalability of NWDAF; (ii) devise a Machine Learning (ML) algorithm, specifically an LSTM Auto-encoder whose role is to detect abnormal traffic events using real network data extracted from the Milano dataset [1]; (iii) we integrate and test the abnormal traffic detection algorithm in the NWDAF based on OpenAirInterface (OAI) 5G CN and RAN [2]. The experimental results show the ability of NWDAF to collect data from a real 5G CN using 3GPP-compliant interfaces and detect abnormal traffic generated by a real UE using ML.

Index Terms—NWDAF, 5G, ML, LSTM Auto-encoder, Abnormal traffic detection, OAI

I. INTRODUCTION

The current advances in 5G networks have led to thriving mobile communication by targeting broad services and applications. In contrast to previous generations, which are consumer-based technologies, 5G and beyond systems, besides the regular customers, target industry verticals, such as Industry 4.0, Virtual Reality (VR), and Internet of Things (IoT). These new vertical applications require high reliability, particularly at the network level. Further, due to the huge size of data and its complexity, model-based approaches became less efficient, while data-driven approaches have become promising solutions to improve and optimize network performances. In this context, the Network Management System (NMS) should: (i) monitor data from different sources (Network Functions (NFs), Radio Access Network (RAN), Infrastructure, etc.); (ii) derive analytics and insights from the collected data in order to detect events and react by applying relevant actions. For example, the NMS monitors the CPU and RAM usage of the NFs as well as the users’ generated load to predict anomalies that may happen due to insufficient RAM resources. Hence, the NMS should increase the resources dedicated to the concerned NFs. However, a large proportion of NFs/RAN vendors use closed and proprietary systems, making collecting data difficult or even impossible. To overcome this limitation, the 3rd Generation Partnership Project (3GPP) introduced the Network Data Analytics Function (NWDAF) starting from Release 15 and maintaining till the current standard (i.e., Release 17) [3].

The NWDAF is a data-aware NF that aims to provide data analytics services to support network optimization, service assurance, and business intelligence. The NWDAF can provide insights about network performances, traffic patterns, and users’ behaviour, which can help network operators optimize their networks for better efficiency, reliability, and Quality of Service (QoS) support. Besides, NWDAF contributes to the entire life cycle of network planning, construction, operations, and maintenance.

Table I summarizes the NWDAF services according to Release 17 of the 3GPP standard and briefly describes each service. To provide analytics services, the NWDAF collects raw data from different sources, including NFs, Application Functions (AFs), and Operation Administration and Maintenance (OAM) using well-established Application Programming Interfaces (APIs) standardized by 3GPP. Then, the NWDAF performs intelligent analytics on the raw data and outputs the analytics results to consumers, including other NFs and the OAM. Thus, using the provided analytics, NWDAF consumers (such as NMS) can make smart decisions about network operations, service delivery, and customer experience.

Among the key use cases of NWDAF, detecting User Equipment (UE)s Abnormal behaviour attracted the attention of mobile operators. Indeed, the number of connected devices and the data volume are growing exponentially, making the abnormal detection task very challenging for mobile operators. Moreover, only limited research has been conducted in the area, and only a few functional 5G Core Network (CN) and NWDAF experimental prototypes are available, such as [5]. Until this point, most of the previous works have relied on using 4G cellular and internet traffic to extrapolate relationships for 5G core networks. Authors of [6] explain
<table>
<thead>
<tr>
<th>Categorie</th>
<th>Event ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Conditions</td>
<td>slice_load_level</td>
<td>Network slice load level computation and prediction.</td>
</tr>
<tr>
<td></td>
<td>nsi_load_level</td>
<td>Network slice instance load level computation and prediction.</td>
</tr>
<tr>
<td></td>
<td>if_load</td>
<td>Load analytics information and prediction for a specific network function.</td>
</tr>
<tr>
<td></td>
<td>user_data_congestion</td>
<td>Congestion information—current and predicted for a specific location.</td>
</tr>
<tr>
<td></td>
<td>network_performance</td>
<td>Network performance computation and prediction on the gNB status information.</td>
</tr>
<tr>
<td></td>
<td>qos_sustainability</td>
<td>QoS sustainability—reporting and predicting QoS change.</td>
</tr>
<tr>
<td>Device Behaviour</td>
<td>ue Mobility</td>
<td>UE mobility analytics and expected behaviour prediction.</td>
</tr>
<tr>
<td></td>
<td>ue_comm</td>
<td>UE communication analytics and pattern prediction.</td>
</tr>
<tr>
<td></td>
<td>abnormal Behaviour</td>
<td>UE abnormal behaviour and anomaly detection, e.g., being misused or hijacked.</td>
</tr>
<tr>
<td>Service Experience</td>
<td>service experience</td>
<td>Service experience computation and prediction for an application or UE group.</td>
</tr>
<tr>
<td></td>
<td>red_trans_exp</td>
<td>Redundant Transmission Experience related analytics. These analytics may be used by the SMF to determine whether redundant transmission shall be performed, or shall be stopped.</td>
</tr>
<tr>
<td></td>
<td>wlan_performance</td>
<td>Quality and performance of WLAN connection of UE computation and prediction.</td>
</tr>
<tr>
<td></td>
<td>dn_performance</td>
<td>User plane performance computation and prediction.</td>
</tr>
<tr>
<td>Network Planning</td>
<td>dispersion</td>
<td>Location or network slices where UEs disperse most of their data volume and sessions transactions.</td>
</tr>
</tbody>
</table>

abnormal behavior detection in NWDAF according to 3GPP. Furthermore, they extensively review the related work, summarize open problems, and provide possible future research directions. However, they did not provide a solution to the problem. The authors of [5] show the integration of Open5GS 5G CN with a NWDAF prototype. However, they only focused on NFs interaction data, ignoring UE data. Authors of [7] proposed a functional prototype of NWDAF integrated with Open5GS 5G CN. However, they did not leverage ML to provide advanced use cases.

In this paper, we design a fully 3GPP-standardized NWDAF based on the microservices architecture that allows new use cases to be added in a plug-and-play fashion. The proposed architecture includes three layers: (i) the Exposure layer that implements the 3GPP NWDAF open API [8] to provide a fully 3GPP-standardized northbound interface; (ii) the Analytics layer that implements the NWDAF intelligence using Machine Learning (ML) techniques; and (iii) the Monitoring layer that uses the NFs service based interface to collect data from NFs, the O-RAN xApps to collect data from the RAN and the Virtual Infrastructure Manager (VIM) to collect data from the infrastructure. Then, we integrated our NWDAF with OpenAirInterface (OAI) 5G CN1, mainly with AMF and SMF. Moreover, we designed and implemented a traffic anomaly detection algorithm based on Long-Short-Term-Memory (LSTM) Auto-encoder and plugged the algorithm into the NWDAF architecture as a microservice. We trained the LSTM Auto-encoder using real data from the Milano dataset [1]. Finally, we tested the algorithm at EURECOM 5G facility [9] using Commercial Off-The-Shelf (COTS) UE and real network conditions. The results show the efficiency of the proposed algorithm in detecting traffic anomalies.

The remaining sections of this paper are structured as follows: Section II describes the NWDAF realization. In Section III, we present use cases that highlight the practical applications of NWDAF. Section IV provides an analysis of NWDAF performance. Finally, Section V concludes the paper.

II. NWDAF REALIZATION

The section is structured into three subsections, covering different aspects of NWDAF. Subsection II-A provides an overview of NWDAF's architecture, while subsection II-B delves into technical details of the NWDAF's implementation. Finally, subsection II-C explores NWDAF interoperability aspects with OAI 5G CN.

A. NWDAF Architecture Aspects

The overall architecture of NWDAF is depicted in Figure 1, which comprises three layers: (i) Exposure, (ii) Monitoring, and (iii) Analytics. The Exposure layer is the system’s entry point, where clients can submit requests following 3GPP specifications. The Monitoring layer collects data from the infrastructure and stores it in a database. The Analytics layer performs computing tasks to derive the requested information.

1) The Exposure layer: The Exposure Layer is responsible for communication with external entities, such as service providers and NFs. One of its key components is the North-Bound Interface (NBI) Gateway module, which acts as a 3GPP-compliant interface between NWDAF and its clients. The latter can either request information or subscribe for an event (i.e., the client waits till an event occurs to receive information from the NWDAF). The client request is routed to either the NBI Analytics Info module or the NBI Events Subscription module, depending on the nature of the request. On the one hand, the NBI Analytics Info module enables external entities to obtain real-time analytical data from the NWDAF, offering valuable insights into network performance. Its main responsibilities include: (i) handling requests from the NBI Gateway module; (ii) requesting the Analytics layer to compute the desired information; (iii) and generating and

1https://gitlab.eurecom.fr/oai/cn5g/oai-cn5g-fed
transmitting responses to clients based on the Analytics layer’s output. On the other hand, the NBI Events Subscription module allows external entities to subscribe to specific network events and receive notifications when those events occur. This component performs several tasks, including: (i) handling requests from the NBI Gateway module; (ii) keeping track of existing subscriptions; (iii) communicating with the Analytics layer to compute the desired information; (iv) and delivering notification messages to the corresponding clients based on the responses received from the Analytics layer.

2) The Analytics layer: Within the microservices architecture of NWDAF, each service (summarized in Table I) is mapped to an engine located in the Analytics layer. Each engine performs the required computations to derive the target service information. The Exposure layer is responsible for selecting the appropriate engine based on the type of service requested. For instance, The Anomaly Detection Engine depicted in Figure 1 employs an ML-based model to calculate the probability of traffic anomalies based on the history of the UE traffic patterns. The latter are retrieved from the database of the Monitoring layer by the engines. By comparing the current data pattern with the past, the engines can detect potential issues, and network administrators can take preventive measures. The ML models are stored in a ML repository which is populated by the network administrator as shown in Figure 1. The engines pull the ML model that corresponds to their proposed service.

3) The Monitoring layer: The SouthBound Interface (SBI) component is a vital element responsible for collecting and storing network data. It communicates with several entities, such as NFs for CN-related data, VIM for Edge Network-related data, and xApps for RAN-related data. Upon startup of the NWDAF, the SBI module subscribes to the CN NFs to receive notifications and stores the received notification data in the database. Furthermore, it requests VIM to collect the RAM and CPU utilization of various NFs in the Edge network. For more details on the KPIs collection mechanism, readers can refer to [10]. The NWDAF can also request RAN information leveraging O-RAN xApps and KPM Service Model (SM) [11].

B. NWDAF Implementation Aspects

NWDAF was built using multiple technologies and programming languages. The RESTful API architecture is used for all components due to its scalability, ability to manage multiple data formats, and utilization of HTTP, a widely used network communication protocol. Python is used for the development of engines that use ML models due to its extensive libraries and frameworks for data analysis and machine learning. GoLang is used for the remaining system components due to its efficient code generation, essential for high-performance system components. This language combination allows greater system flexibility and efficiency, resulting in a more robust architecture. MongoDB is selected as the database due to its scalability and ability to manage unstructured data.

C. NWDAF Interoperability Aspects

The few experimental NWDAF prototypes are coupled with Open5GS CN. The latter does not expose a standardized Event Exposure API. Thus, we selected OAI 5G CN, which implements the 3GPP-compliant Event Exposure API, to test our NWDAF. OAI offers a 3GPP-compliant Release 16 5G CN and RAN that support COTS UEs in real radio conditions. The integration with the CN is done by subscribing to AMF and SMF Event Exposure APIs while the integration with the RAN is done using FlexRIC [12] as O-RAN RIC. The latter hosts monitoring xApps used by the NWDAF to collect radio information.
III. NWDAF Use cases

As mentioned earlier, the NWDAF provides two main services: event subscriptions and analytics information. The NBI Events Subscription module enables clients to subscribe or unsubscribe to/from various analytics events, while the NBI Analytics Info module allows clients to request and receive specific types of analytics information. Our NWDAF provides both Core and ML-based services, which we will explore further.

A. Core services

Our proposed NWDAF supports three Core services:

- **Network Performance**: The NWDAF provides support for two types of network performance events: “num_of_ue,” which measures the number of attach requests during a time window, and “sess_succ_rate,” which measures the session success rate during a time window specified in the request. The NWDAF computes “num_of_ue” using SMF notifications, specifically, the registration event. The NWDAF also supports filtering the number of attach requests according to a specific network area or a specific operator. “sess_succ_rate” is computed using the PDU session establishment event. The NWDAF also supports filtering according to a specific data network name or a network slice.

- **UE communications**: “ue_comm” refers to the number of packets and bytes exchanged in the uplink and downlink directions for each PDU session. To incorporate these statistics into SMF notifications, OAI-UPF-VPP was used during core network deployment. The SMF collects measurement reports from the UPF using the N4 interface for the usage report procedure of the N4 interface.

- **NF Load**: As mentioned earlier, the NWDAF computes the “nf_load” event data using data received from the MongoDB database. This data includes information on the CPU and RAM consumption of each NF.

Whenever a client subscribes for an event, the associated engine retrieves the required data from the MongoDB database and performs computation on the requested time window, i.e., the time interval of the client’s interest.

B. ML-based services

One of the key services of our NWDAF is ML-based abnormal traffic detection, which is identified in the 3GPP standard by the event ID “abnormal_behaviour” and exception ID “unexpected_large_rate_flow”.

- **Abnormal Traffic**: The NWDAF clients can subscribe to the “abnormal_behaviour” event to receive periodic updates on the probability of abnormal traffic. The latter is computed by the Anomaly Detection Engine leveraging an LSTM-based Auto-encoder. The LSTM-based Auto-encoder learns the traffic pattern using the history of the UE data measured from the UPF in both uplink and downlink directions. The traffic pattern contains information such as the data size in bytes at a given timestamp during a week. Figure 2 depicts a basic illustration of the LSTM Auto-encoder model architecture.

An Auto-encoder is a type of artificial neural network that consists of an encoder and a decoder. During the encoding phase, the input data, represented as \( X = \{x_1, x_2, \ldots, x_n\} \), is compressed into a lower-dimensional space according to Equation 1.

\[
Z = \sigma(WX + b)
\]  
(1)

Where \( Z \) is the output of the LSTM encoder, \( \sigma \) is the activation function of the LSTM encoder, \( W \) is the weights of the encoder layers, and \( b \) is the bias vector of the encoder layers. Similarly, the decoding phase is trained according to Equation 2 to obtain the output data \( X' = \{x'_1, x'_2, \ldots, x'_n\} \) that is similar to the input dimension.

\[
X' = \sigma'(W'Z + b')
\]  
(2)

Where \( Z \) is the input of the LSTM decoder, \( \sigma' \) is the activation function of the LSTM decoder, \( W' \) is the weights of the decoder layers, and \( b \) is the bias vector of the decoder layers. The Auto-encoder architecture’s motivation is that it reduces the variance between input and output by training on only normal traffic without considering abnormal traffic. Whereas, the motivation behind combining LSTM and the Auto-encoder architecture is due to the nature of the input, which is correlated with time, e.g., traffic during the night is different from traffic during the day. Besides, the LSTM is well known for dealing with the vanishing gradient problem [13], which makes the training more stable. In Equation 3, \( \sigma \) represents the Mean Absolute Error (MAE) between the input and output of the Auto-encoder.

\[
\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x'_i - x_i| = \frac{1}{n} \sum_{i=1}^{n} |e_i|
\]  
(3)

Let \( \beta \) denote the average MAE over the training data and \( \beta' \) the MAE of the inference data. Equation 4 calculates the traffic anomaly probability \( p \) using the difference between \( \beta \) and \( \beta' \).

\[
p = \min(\alpha \times |\beta' - \beta|, 1)
\]  
(4)

Where \( \alpha \) is a weight to control the impact of the distances scale, i.e., when \( \alpha \) is small enough, the distances will have less impact on the probability.
- **ML-based services perspectives:** The suspicion of Distributed Denial Of Service (DDOS) attack service, which is identified by the event ID “abnormal Behaviour” and exception ID “suspicion of ddos attack,” can be performed by analyzing the number of users in a similar way to abnormal traffic detection. The Auto-encoder will be trained with normal historical data, and if the current data pattern deviates significantly from past patterns, the likelihood of a suspicion of DDOS attack increases. This probability can be used in a closed-control loop to help the 5G CN mitigate DDOS attacks [14]. Additionally, radio link failures, identified by the “abnormal Behaviour” event ID and “unexpected radio link failures” exception ID, can be predicted using a combination of LSTM and Support Vector Machine (SVM), as proposed in [15].

IV. PERFORMANCE EVALUATION

The section is structured into three subsections: Experimentation setup, which details the experimental setup; Experimentation results, which presents and analyzes the performance of the NWDAF; and Experimentation conclusion, which offers additional insights related to the experiment.

A. Experimentation Setup

Our experimental setup includes two machines, each equipped with 36 Intel(R) Xeon(R) Gold 6154 CPUs running at 3.00GHz. One of these machines is used to run the gNB based on OAI and is connected to a USRP B210, while the second machine hosts the 5G Core Network based on OAI and the NWDAF. In addition, we have a laptop with Ubuntu Operating System (OS) connected to a Quectel RM500Q-GL module, which is considered as a 5G UE. To train our LSTM Auto-encoder, we leveraged the Milano dataset [1]. This dataset was collected by Telecom Italia for a year. It contains various information regarding user connectivity events. Our experiment focused on the volume of data exchanged with users. We filtered the dataset and organized the records into the following structure: (weekday, hour, minute, and internet data). We included the weekday as traffic varies from day to day, and the hour as traffic during night time differs from that during the day. The internet data metric, introduced by Telecom Italia, is proportional to the traffic volume and is used to conceal the true values. To test the LSTM Auto-encoder, we generate the anomalies by introducing long traffic flows that do not follow the Milano dataset pattern. The LSTM Auto-encoder includes two hidden layers for both the encoder and the decoder. We employ a learning rate of 0.001, batch size of 128, tanh activation function, and train the model using Adam optimizer for 20 epochs. The input sequence size is set to 12. An expert opinion was considered to set the weight \( \alpha \) of Equation 4 to 0.6, as it produces reasonable anomaly probabilities and a realistic number of anomalies.

B. Experimentation Results

Figure 3 shows the training and validation loss over the number of epochs. We observe that the loss trend is similar for
training and validation sets and converges after approximately ten epochs. On the other hand, Figure 4 illustrates the Milano internet data for one week and the generated data using LSTM Auto-encoder. The figure shows that our LSTM Auto-encoder is able to learn the Milano dataset data, which is characterized by high data volume during the day hours and low data volume during the night hours. In addition, the Milano training data with injected anomalies and anomaly probabilities are plotted in Figure 5. This figure demonstrates that the LSTM Auto-encoder reconstructs input data but deviates when anomalies occur. Further, the distance between input and generated data correlates with the anomaly probability. The probabilities increase as the generated data diverges from the input data. The anomaly probability threshold is set to 0.5 to determine whether or not the traffic is an anomaly. Consequently, sequences that deviate from normal behaviour are highlighted in Fig 6. For instance, Internet data values that are unusually high or unusually low are flagged as potential problems. As previously stated, the sequence size is 12 time steps. As a result, points preceding and following the anomalies are highlighted, as they are in the sequence containing the anomaly.

We have created a video demonstration to showcase the utilization of the LSTM Auto-encoder in NWDAF. The video link can be found here: https://lnkd.in/e-b2jaHK. In the video, we have shown the subscription of an anomaly detector client to NWDAF to obtain abnormal traffic probability. We first generated normal uplink data at the UE using the Milano data pattern, which resulted in receiving low probabilities. We then generated abnormal traffic, and the client received a high abnormal traffic probability. Finally, we reverted the traffic to normal, and the probability also decreased.

C. Experimentation Conclusion
According to the outcomes of our experimentation, NWDAF can contribute significantly to improving 5G network orchestration. For instance, it can inform clients about the probability of abnormal traffic, which can be used to conduct corrective actions. Then, these clients can use Root Cause Analysis (RCA) tools to identify the underlying causes of network failures. For example, eXplainable AI (XAI) and Machine Reasoning-based techniques can be used in conjunction with expert knowledge to identify the reasons for anomalies using other NWDAF services. In the context of 6G Network’s Zero-touch Service Management (ZSM), this can lead to an AI-based closed-loop fault management that will enable a self-managed network and minimize human intervention. This can also be integrated into NWDAF services by plugging a new microservice Engine for RCA that alerts NWDAF clients of the root cause.

V. CONCLUSION
In this paper, we have proposed a microservices architecture for NWDAF to address the challenge of realizing the various use cases defined by 3GPP. By plugging the 3GPP use cases as microservices, we have increased the flexibility and scalability of the NWDAF. We have also designed and implemented an LSTM Auto-encoder algorithm to detect abnormal traffic events. The proposed model is integrated in the NWDAF as a microservice. The experimental results showed the ability of NWDAF to collect data from a real 5G CN and detect abnormal traffic generated by a real UE. Our solution can help enable a self-managed network in the context of 6G’s Zero-touch Service Management and reduce human intervention, ultimately improving the network performance and end-user experience. Overall, this research contributes to the development of NWDAF and its applications in Beyond 5G networks, particularly abnormal traffic detection.

ACKNOWLEDGMENT
This work is partially supported by the European Union’s Horizon Program under the 6GBricks project (Grant No. 101096954) and the Imagine-B5G project (Grant No. 101096452).

REFERENCES