QUADRI-DIMENSIONAL APPROACH FOR DATA ANALYTICS IN MOBILE NETWORKS

by

MAMPAKA MALUAMBANZILA MINERVE

submitted in accordance with the requirements for
the degree of

MASTER OF TECHNOLOGY

in the subject

ENGINEERING: ELECTRICAL

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: Dr M. Sumbwanyambe

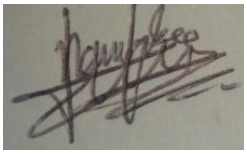
October 2018

DECLARATION AND COPYRIGHT

I declare that "*Quadri-Dimensional Approach for Data Analytics in Mobile Networks"* is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I have not previously submitted this work, or part of it, for examination at UNISA for another qualification or at any other higher education institution.

MM MAMPAKA DATE



_____ 08.10.2018

ACKNOWLEDGMENTS

I would like to thank my amazing wife Sonia Kiangala for her support and advices throughout this study. Sonia, you have been a great motivation for me.

I am equally grateful to my supervisor Dr. Mbuyu Sumbwanyambe for all his inputs and guidance.

Above all, I want to thank the GOD almighty for granting me life, intellect, perseverance and the ability to conduct this study.

ABSTRACT

The telecommunication market is growing at a very fast pace with the evolution of new technologies to support high speed throughput and the availability of a wide range of services and applications in the mobile networks. This has led to a need for communication service providers (CSPs) to shift their focus from network elements monitoring towards services monitoring and subscribers' satisfaction by introducing the service quality management (SQM) and the customer experience management (CEM) that require fast responses to reduce the time to find and solve network problems, to ensure efficiency and proactive maintenance, to improve the quality of service (QoS) and the quality of experience (QoE) of the subscribers. While both the SQM and the CEM demand multiple information from different interfaces, managing multiple data sources adds an extra layer of complexity with the collection of data.

While several studies and researches have been conducted for data analytics in mobile networks, most of them did not consider analytics based on the four dimensions involved in the mobile networks environment which are the subscriber, the handset, the service and the network element with multiple interface correlation.

The main objective of this research was to develop mobile network analytics models applied to the 3G packet-switched domain by analysing data from the radio network with the Iub interface and the core network with the Gn interface to provide a fast root cause analysis (RCA) approach considering the four dimensions involved in the mobile networks. This was achieved by using the latest computer engineering advancements which are Big Data platforms and data mining techniques through machine learning algorithms.

Keywords: Telecommunication, Mobile Networks, Packet-Switched, QoS, QoE, SQM, CEM, Root Cause Analysis, Data Analytics, Big Data, Machine Learning, Artificial Intelligence, ANN, Deep learning.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	Definitions
1G	1st Generation mobile network
2G	2nd Generation mobile network
3G	3rd Generation mobile network
4G	4th Generation mobile network
5G	5th Generation mobile network
AMPS	Advanced Mobile Phone System
ANN	Artificial Neural Network
API	Application Programming Interface
APN	Access Point Name
AR	Augmented Reality
ATM	Asynchronous Transfer Mode
BSS	Base Station Subsystem
CART	Classification And Regression Trees
CDN	Content Delivery Network
CDR	Call Data Record
CEM	Customer Experience Management
CNN	Convolutional Neural Networks
CSP	Communication Service Provider
CSSR	Call Setup Success Rate
D-AMPS	Digital-Advanced Mobile Phone System
DBN	Deep Belief Networks
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCH	Dedicated Channel
DNN	Deep Neural Network
DNS	Domaine Name Server
DPI	Deep Packet Inspection
DT	Decision Tree
E2E	End-to-End
FACH	Fast-Access Channel

FM	Fault Management
FP	Frame Protocol
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
GMDH	Group Method of Data Handling
GMM	GPRS Mobility Management
GPRS	General Packet Radio Services
GPU	Graphical Processing Unit
GSM	Global System for Mobile communications
GTP	GPRS Tunnelling Protocol
HDFS	Hadoop Distributed File System
HMM	Hidden Markov Model
HSPA	High Speed Packet Access
HTTP	Hypertext Transfer Protocol
ICT	Information and Communications Technology
IMEI	International Mobile Equipment Identity
IMSI	International mobile subscriber identity
IoT	Internet of Things
IP	Internet Protocol
ISP	Internet Service Provider
ITU	International Telecommunication Union
KNN	K-Nearest Neighbors
KPI	Key Performance Indicator
KQI	Key Quality Indicator
K-SVD	K-Singular Value Decomposition
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-agnostic Explanations
LSTM	long short-term memory network
LTE	Long Term Evolution
M2M	Machine to Machine
MAE	Mean Absolute Error
MBB	Mobile Broadband
MIMO	Multiple-in Multiple-out

MLP	Multi-Layer Perceptron
MLPWD	Multi-Layer Perceptron with Weight Decay
MM	Mobility Management
MMH	Maximum Margin Hyperplane
MOS	mean opinion score
MSC	Mobile Switching Centre
MSE	Mean Squared Error
MTTR	Mean-Time To Repair
NAS	Non-Access Stratum
NBAP	Node-B Application Part
NFV	Network Function Virtualization
NLP	Natural Language Processing
NMT	Nordic Mobile Telephone
NOC	Network Operation Centre
NPS	Network Promotor Score
NRA	Network Authority Regulators
OMC	Operations and Maintenance Centre
OPEX	Operational Expenditure
OSS	Operations Support Systems
OTT	Over-The-Top
P2P	Peer-to-Peer
PDP	Packet Data Protocol
PLMN	Public Land Mobile Network
PM	Performance Management
QoE	Quality of Experience
QoS	Quality of Service
RAB	Radio Access Bearer
RACH	Random-Access Channel
RANAP	Radio Access Network Application Part
RAT	Radio Access Technology
RBM	Restricted Boltzmann Machines
RCA	Root Cause Analysis
RF	Random Forests

RLC	Radio Link Control
RMSE	Root Mean Square Error
RNC	Radio Network Controller
RNN	Recurrent Neural Networks
RRC	Radio Resource Control
SDN	Software-Defined Networking
SGSN	Serving GPRS Support Node
SM	Session Management
SMEs	Small and Medium-sized Enterprises
SMS	Short Message Service
SOC	Service Operating Centre
SQI	Service Quality Index
SQL	Structured Query Language
SQM	Service Quality Management
SVM	Support Vector Machines
SVR	Support Vector Regression
TACS	Total Access Communication System
TCP	Transmission Control Protocol
TD-SCDMA	Time Division Synchronous Code Division Multiple Access
UDP	User Datagram Protocol
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
URL	Uniform Resource Locator
UTRAN	UMTS terrestrial radio access network
VM	Virtual Machine
VOD	Video On Demand
VR	Virtual Reality

List of Publications

The following papers were published for this research:

1. **Mampaka Maluambanzila Minerve** and Mbuyu Sumbwanyambe, “A Quadri-Dimensional Approach for Poor Performance Prioritization in Mobile Networks Using Big Data”, Journal of Big Data (2019)6:10, February 2019. https://doi.org/10.1186/s40537-019-0173-8

2. **Mampaka Maluambanzila Minerve** and Mbuyu Sumbwanyambe, “Poor Data Throughput Root Cause Analysis Using Deep Neural Network”, *IEEE Wireless Africa Conference (WAC) 2019*, Pretoria, South Africa, August 2019.

3. **Mampaka Maluambanzila Minerve** and Mbuyu Sumbwanyambe, “Facebook Video Traffic Classification Using Artificial Neural Network”, *in Proc. SATNAC conference*, Western Cape, South Africa, pp. 236-241, September 2018.

4. **Mampaka Maluambanzila Minerve** and Mbuyu Sumbwanyambe, “User-Based QoS Categorization from Radio Conditions Using Machine Learning Techniques”, *in Proc. SATNAC conference*, Western Cape, South Africa, pp. 44-49, September 2018.

Publications not related to this research:

5. Lisungu Oteko Tresor, **Mampaka Maluambanzila Minerve** and Sumbwanyambe Mbuyu, "An Objective MOS Prediction Approach Based on the Nb Interface Parameters”, *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, Bloemfontein, South Africa, 2019, pp. 171-176.

Chapter 1: Introduction

1.1. Background of the study

The mobile networks services with the explosion of new generations of devices, smartphones and tablets have been evolving from simple calls and internet browsing to data intensive applications such as video streaming, social media, online gaming, internet protocol television (IPTV) and even security demanding applications for banking or mobile money [1]. This evolution leads to a complexity in the monitoring and management of the networks with a high number of services to consider in a typical mobile network environment. The complexity is further increased with the introduction of the internet of things (IoT), machine to machine (M2M) communications [2] and the network function virtualization (NFV) with the software-defined networking (SDN) [3].

To manage the complexity of the current mobile networks ecosystems, there has been a paradigm shift from network- to service-oriented management with the subscribers' satisfaction in the centre of the network management. The network operation centre (NOC) has become a lower layer of the so-called service operation centre (SOC) with the implementation of the service quality management (SQM) and the customer experience management (CEM) [4].

The implementation of the SQM has become a huge challenge for communication service providers (CSPs), since it involves several areas of expertise needed to handle both the quality of service (QoS) and the quality of experience (QoE) which are both linked to the profitability of the organization.

A customer experience lifecycle includes the experience before the subscription, while using the network and when leaving the network in case of churn [5]. The SQM requires a deep knowledge of the technical and business aspects of the mobile network environment, a collection and correlation of customers' information from various data sources such as the network probing, the performance management (PM), the fault management (FM) and the billing information for different services [6]. All these variations of data sources lead to a need to perform mobile network analytics based on multiple dimensions such as the subscriber, the handset, the service and the network element contributing to the management diversity by using an intelligent approach for data collection, data mining and analysis [4].

1.2. Problem statement and purpose of study

1.2.1. Problem statement

Previous researches in mobile network analytics tend to only focus on one dimension at a time, without considering the links and the impacts between different dimensions in mobile networks. An efficient SQM should assume that the subscriber's experience could deteriorate due to the usage of underperforming handsets. In the same vein some handsets might be underperforming only because they are mostly used in a cell with poor coverage. For this, there is a need to bear in mind the correlations and links between the four dimensions (the subscriber, the handset, the service and the network element) during the collection of data, the processing, the root cause analysis (RCA) and the analytics.

1.2.1.1. Sub-problem 1: Data collection and processing

To ensure the implementation of an SQM, there is a need to process and store a large amount of historical aggregated data for the feeding of the SOC [7]. The traditional database management systems based on the structured query language (SQL) have shown their limitation to store and retrieve huge amount of data from multiple interfaces. There is therefore a need to use Big Data platforms to ensure a faster and efficient management of data.

1.2.1.2. Sub-problem 2: Reporting and root cause analysis (RCA)

To unleash the business benefits from the collected information in the mobile networks and increase the revenue, two of the most critical barriers are the organization processes and the data complexity [4]. Thus, there is a need for an intelligent reporting approach to manage the data complexity in the mobile network by understanding different network processes. This will help in building useful reports and most importantly derive the RCA to reduce the troubleshooting time and hence improve the efficiency while optimizing the operation expenditure (OPEX).

1.2.1.3. Sub-problem 3: Analytics

One of the leading trends in SQM implementation is the predictive analytics [8]. Depending on the types of data collected, there is a need to apply analytics for any layer of the mobile networks starting from the NOC in the prediction of network elements, the SOC to predict the behaviour of services, up to the marketing department so as to understand the customers' segmentation and preferences for marketing campaigns or product promotions.

1.2.2. Purpose of study

The purpose of this research was to use some of the advancements in the area of computer engineering and data science to provide a Big Data model for data mining and analytics on the mobile networks, particularly, on the packet-switched domain as its usage is tremendously increasing. The study's contribution is to investigate and provide new directions in mobile networks investigation into managing the packet-switch domain and providing customer- and service-oriented RCA with an upper layer of artificial intelligence. The results of this research were submitted or published to local and international journals and conferences on telecommunication and Big Data.

1.3. Research questions

In light of the problem statement above, this study attempted to address the following questions:

1. Which methods have been used for data analytics in mobile networks previously?
2. How to develop a quadri-dimensional (including the subscriber, the handset, the service and the cell) reporting structure for performance monitoring in mobile networks.
3. How to use machine learning techniques to classify and predict the type of data traffic in the network?
4. How to use machine learning techniques to determine the relationship between the radio conditions and the perceived subscribers' QoS in mobile networks?
5. How to develop a model using machine learning techniques, the radio conditions and the core network performance metrics for the RCA of the poor QoS in the mobile networks?

1.4. Research methodology

The research methodology approach used in this study was the design science paradigm. The design science research (DSR) methodology not only provides investigation flow to determine the problems and the objectives, but also focuses on the development and the design of a valuable artefact [9].

The design science methodology has been widely accepted and approved in research circles for the design of engineering products and applications. There are case studies in the fields of data warehouse, software measurement and telecommunication software that demonstrate the successful implementation of the DSR methodology [10].

The phases considered during the execution of this research/study are described as follows:

1. Literature review: To deeply analyse the problems identified in this study, the motivations leading to the objectives and the requirements of the implementation. This first phase consisted of intensive literature survey in mobile network analytics.
2. Design and implementation: To produce a valuable artefact, the research orientation was a solution-based one to design and implement:
    a. Models based on artificial neural network (ANN) to classify video traffic from a sample of Facebook data containing video, chat and browsing traffic. The best model should be selected based on the accuracy as the evaluation metric.
    b. A structured reporting model to enable hierarchical aggregation following a tree approach to speed up troubleshooting considering the main four dimensions in the mobile network.
    c. Models based on machine learning techniques to predict user-based QoS categories using radio conditions parameters. Different models based on different machine learning techniques which are the decision tree (DT), the random forests (RF), the support vector machines (SVM) were trained, tuned and compared to select the best model based on the accuracy.
    d. A model based on deep neural network (DNN) to analyse the performance metrics from both the radio and the core networks for the RCA of a poor throughput performance in the mobile network.
3. Demonstration and evaluation: The results of the research were evaluated based on a quantitative approach of measuring the performance metrics of the system and a qualitative approach of assessing the accuracy of the results.
4. Communication of the research results: The results of this research will be submitted and published in journals and conferences.

1.5. Research objectives

The main objectives of this research were:

1. To provide a data analytics model applied to the packet-switched domain of the mobile network and build an intelligent reporting model.
2. To analyse the data from the radio network on the Iub interface, the core network on the Gn interface.

3. To establish the correlation between the performance metrics of the network and the perceived experience by the subscribers.
4. To use machine learning techniques for classification problems in mobile networks and to develop an RCA model for poor network performance.

1.6. Scope and delimitation

The scope of this research was limited to the following assumptions and delimitations:

1. The study only focused on the universal mobile telecommunications system (UMTS) 3G packet-switched domain since most of the services, applications and handsets contributing to the complexity of the mobile network environment are operating under that domain.
2. Although there are several interfaces in the packet-switched domain, this research considered only the main interfaces of the packet-switched domain of a UMTS 3G mobile network. The interfaces considered for this study are as follows:
    a) Iub interface: The Iub interface is the radio 3G interface between the Node-B and the radio network controller (RNC) carrying the radio resource control (RRC) measurement reports that provide information about the propagation delay (Distance at which the customers are generating their radio activities), the radio signalling strength and radio signalling quality.
    b) Gn interface: The Gn interface is the core network interface between the general packet radio services (GPRS) support nodes such as the serving GPRS support node (SGSN) and the gateway GPRS support node (GGSN) by carrying the packet data protocol (PDP) information and the user-plane information such as the bytes usage, the latency or the packets retransmission.
3. This research used a physical computer for hosting the application programming interfaces (APIs) used to write the relevant programming codes and a virtual environment to host the Big Data platform for data storage and analytics.

Chapter 2: Literature Review

2.1. Introduction to the mobile networks

The mobile networks have a direct impact in our daily lives, starting with the devices we use to communicate through calls, short message service (SMS), social media chats or just for simple researches on the internet. For the past years, the use cases in the mobile network environment have evolved to address new requirements such as online gaming with the virtual reality (VR) or the augmented reality (AR). These are both very demanding in terms of higher throughputs and lower latencies. Others are applications such as the mobile money services which require high security features [1]. The mobile networks are intensively using the latest evolutions of computer engineering such as the network function virtualization (NFV) with software-defined networking (SDN), artificial intelligence with machine learning algorithms and Big Data concepts to enhance the capability and flexibility of CSPs. These are all implemented so as to address the new emerging use cases while improving the network monitoring and customer's satisfaction [4].

2.2. The mobile networks evolution

The mobile networks started with the introduction of the 1st generation (1G) mobile networks that consisted of purely analogue systems supporting only the voice services. Some of the famous 1G mobile networks technologies are the advanced mobile phone system (AMPS), the Nordic mobile telephone (NMT) and the total access communication system (TACS). The 2nd generation (2G) mobile networks such as the digital-advanced mobile phone system (D-AMPS), the global system for mobile communications (GSM) and the general packet radio services (GPRS) introduced the concept of digitization of information with the combination of both the voice and the data services. In the early 2000s, the 3rd generation (3G) mobile network was introduced with new modulation techniques that enabled the mobile broadband (MBB) with the high-speed packet access (HSPA) plus the downlink with the carrier aggregation that used multiple-in and multiple-out (MIMO) techniques. Some of the 3G technologies available today are the UMTS and the time division synchronous code division multiple access (TD-SCDMA). The 4th generation (4G) mobile networks such as the long-term evolution (LTE) were introduced after 2008, to enable higher data rates and for faster and better data connectivity. This was mainly done for services requiring richer contents and more connections [11]. Lately, different telecommunication standardization bodies are writing for the specifications of the 5th generation

of telecommunication (5G) technologies that will probably have first commercial deployments from 2020 [3].

2.3. Mobile networks trends

The trend of the mobile networks is strongly influenced by the types of use cases that need to be addressed. The amount of applications and services increases with the number of smartphones [4] and software capabilities facilitated by APIs and opensource platforms. Most of the developed countries are taking huge strides in the deployment of the 4G networks and the testing of the 5G prototypes. Interestingly, while the European mobile operators have more than 90% coverage on the 4G networks and expecting first deployments of 5G networks around 2020 [7], some of the African countries are still lagging with the deployment of the 4G networks as shown in Figure 1 [12] where the percentage of 4G connections will still be below 50% by 2021. Therefore, for most of the African countries the packet-switched services will still be delivered by the legacy 2G/3G technologies.

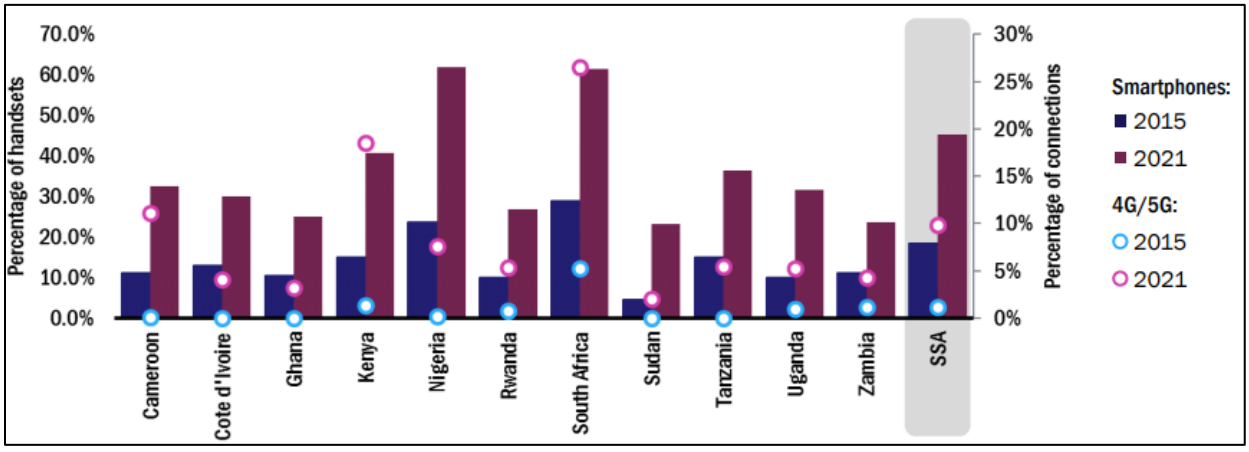


Figure 1: Smartphones as a percentage of handsets, Sub-Saharan Africa 2015 and 2021 [12]

Another factor influencing the trend of the mobile networks is the types of devices. The introduction of smart electronic devices and smart appliances that are technologically wrapped around machine to machine (M2M) communication and the internet of things will contribute to the evolution of smart cities and the evolvement of the industry 4.0 [13]. Even though these new

devices represent opportunities in terms of revenue, there are also new challenges for CSPs in terms of services and network management complexity.

2.4. Service quality management (SQM)

The evolution of different services in the network and the complexity in the management of the mobile networks have a major impact on the QoE [14] which is the measurement of the user's experience based on the service value perception. The QoE can be subjective or objective. A subjective QoE is based on the user's opinion of a certain aspect of a service such as customer service, easy-to-use, cost, billing or performance. This type of QoE does not provide, in general, the difference between the expected and the delivered quality. The objective QoE instead, establishes the relationship between the factors influencing the QoE based on the network information [15] to demonstrate that, although the QoE is linked to the end-user perception, it can have a relationship with the traditional network-centric QoS contributing to the end-user dissatisfaction or satisfaction [16]. This relationship is shown in Figure 2 [17] which describes the four viewpoints of the QoS where the customer's perception of the QoS plays a major role in the overall QoE.

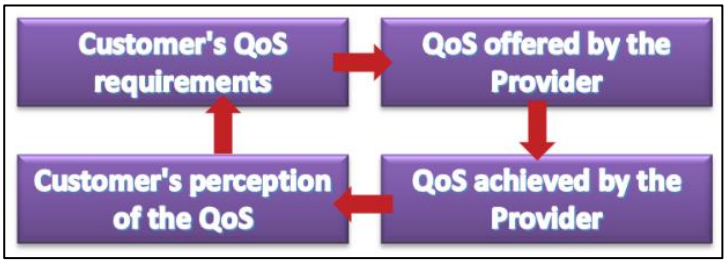


Figure 2: ITU Four Viewpoints of QoS [17]

One of the important metrics used to determine the customers' perception of the network is the network promotor score (NPS). The NPS is survey used to evaluate subscriber's satisfaction through a score ranging from zero to ten (0 to 10) and is based on the subscriber's recommendation of the consumed product [18]. Therefore, the CSPs use the NPS to determine the organization efficiency and effectiveness for the CEM. Figure 3 [18] shows the correlation between the NPS and the churning rate of customers based on the analysis of the throughput for 6 months for small and medium-sized enterprises (SMEs), indicating the relationship between the NPS and the QoE. Generally, poor QoE can be understood through NPS detractors.

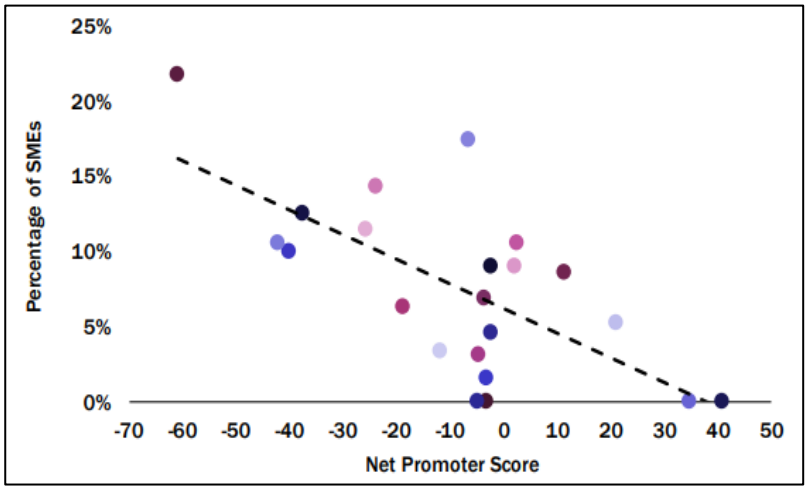


Figure 3: Percentage of SMEs intending to churn within six months and corresponding NPS [18]

The NPS of some of the operators in the African countries are shown in Figure 4 [19]. Although some of the operators in South Africa and in Kenya seem to have high NPS as compared to others, the overall NPS is still poor, leading to high churning rate in most African countries [19].

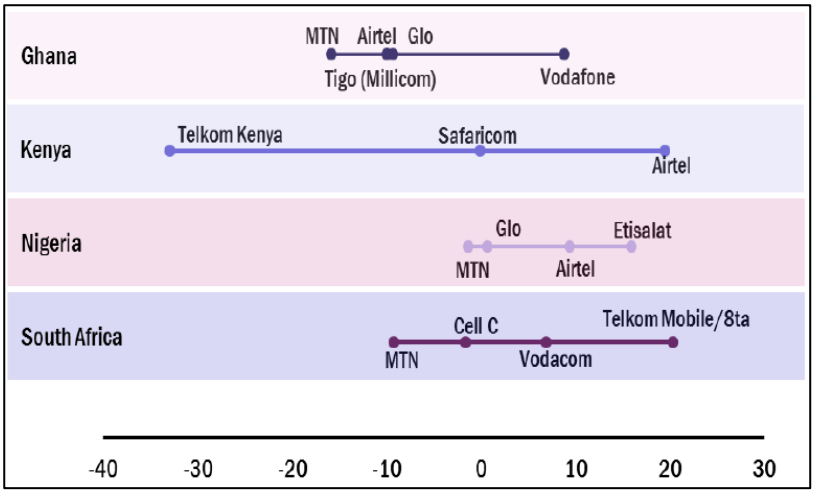


Figure 4: NPS for samples of African operators [19]

The NPS detractors usually churn from the network based on distinctive top three network-related reasons on churn. For most mobile networks, the reasons why customers churn, are based on the service quality, network coverage and data speed as shown in Table 1 [18]. The

data speed which is perceived by the customers as the user throughput plays a very important role in the QoE for the packet-switched domain of the mobile network. As most of the mobile networks in Africa have 3G networks deployment, the optimization of the 3G packet-switched mobile network is still crucial in ensuring better QoS.

Table 1: Top three reasons for churn by enterprise size and service type [18]

Service Type	SMEs	Large enterprises
Mobile	1. Price	1. Customer service
	2. Customer service	2. Price
	3. Network coverage	3. Network quality and data speed
Fixed	1. Data rate or bandwidth	1. Customer service
	2. Price	2. Price
	3. Customer service	3. Network coverage

The traditional NOC has been very inefficient in terms of problem finding, handling and resolution. As more and more people use the mobile networks with multiple connectivity, acquiring a new customer is more difficult than it is for the existing customers to churn. While the traditional mobile networks monitoring strategies followed a bottom-up approach, that is starting with the network elements management, the network alarming and troubleshooting through historical key performance indicators (KPIs) monitoring; the SQM follows the top-down approach starting with the very aggregated service quality index (SQI), down to the KPIs. This assists the CSPs in reacting, in near real-time, not only to issues but also, based on the historical and statistical values, to applying predictive and proactive maintenance [4].

Some of the benefits of the SQM are the reduction of the OPEX, the reduction of the time-to-market, the reduction of the mean-time to repair (MTTR) and the increase of the revenues. Even if these seem to be clear benefits of the SQM approach, however, there are still barriers regarding its full implementation due to delays in digital transformation for certain CSPs, the complexity of the services and the processing of huge amount of the mobile network data before extracting any values. To extract values from the data, the correlation between different parts of the network is a necessity for any SQM to provide an end-to-end (E2E) QoS and QoE. The collection and the correlation of information from different parts of the network allows the computation of

customer-centric metrics which are derived from weighted functions of the aggregated SQI attributes and KPIs from different dimensions in the mobile network [7]. An SQM takes information from multiple data sources including the call data records (CDRs), the measurement reports and the operations and maintenance centre (OMC) data to provide and E2E visibility [20]. The value from the data is then extracted using Big Data and machine learning for predictive analysis [21]. Since the CSPs spend a lot of time when there is an issue or outage in the network because of the complexity of services and the number of elements involved in the mobile networks, the evaluation of the efficiency in detecting and solving issues is measured by the MTTR [22].

2.5. 3G packet-switched mobile networks interfaces

Most of the services and applications introduced with the proliferation of devices are in the packet-switched domain delivered through the mobile broadband (MBB) [14]. In MBB, the network related reasons to churn are due to poor network quality and data speeds [18]. The replacement of the traditional circuit-switched services and the continuous increase in customers' data quality expectation resulted in different SQM model for packet-switched services compared to the traditional circuit-switched [23].

While the 3G packet-switched domain has several interfaces as shown in Figure 5 [99], most of the information required to perform an E2E analysis can be captured from the Iub interface of the radio network and the Gn interface of the core network.

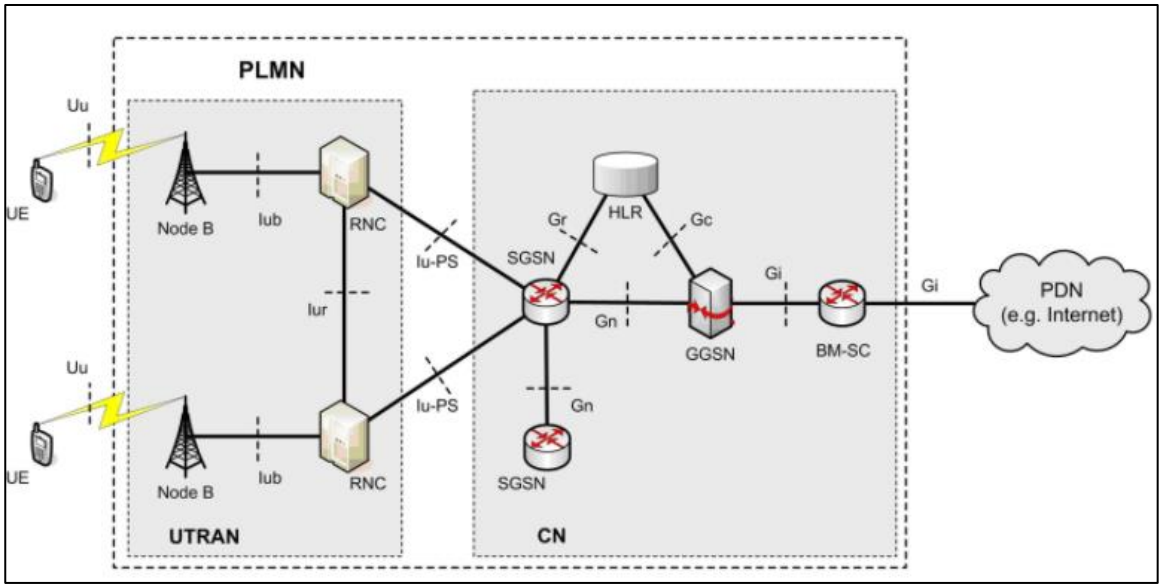


Figure 5: 3G packet-switched network topology [99]

2.5.1. Iub interface

The Iub is the logical interface between the Node-B and the radio network controller (RNC) in the UMTS terrestrial radio access network (UTRAN). It is used to ensure the interconnection between the two nodes, to manage the radio resources as well as to ensure the transport of the non-access stratum (NAS) information transferred from the user equipment (UE) directly to the RNC. The Iub interface supports different data streams depending on the types of the channel used. Some of the Iub interface channels are the dedicated channel (DCH), the fast-access channel (FACH) and the random-access channel (RACH). These channels provide the transport of information during the connection and the idle modes of the network system [24]. The main functions of Iub interface are the management of the radio link, the cell configuration, the radio network performance, the radio resource and the measurement reports.

a) Iub protocol stack

The Iub protocol stack consists of the transport and the network layers. The transport layer is the underlaying layer based on either the asynchronous transfer mode (ATM) or internet protocol (IP). The network layer is composed of the Node-B application part (NBAP), the radio link control (RLC), the radio resource control (RRC) and the upper NAS as shown in Figure 6 [24].

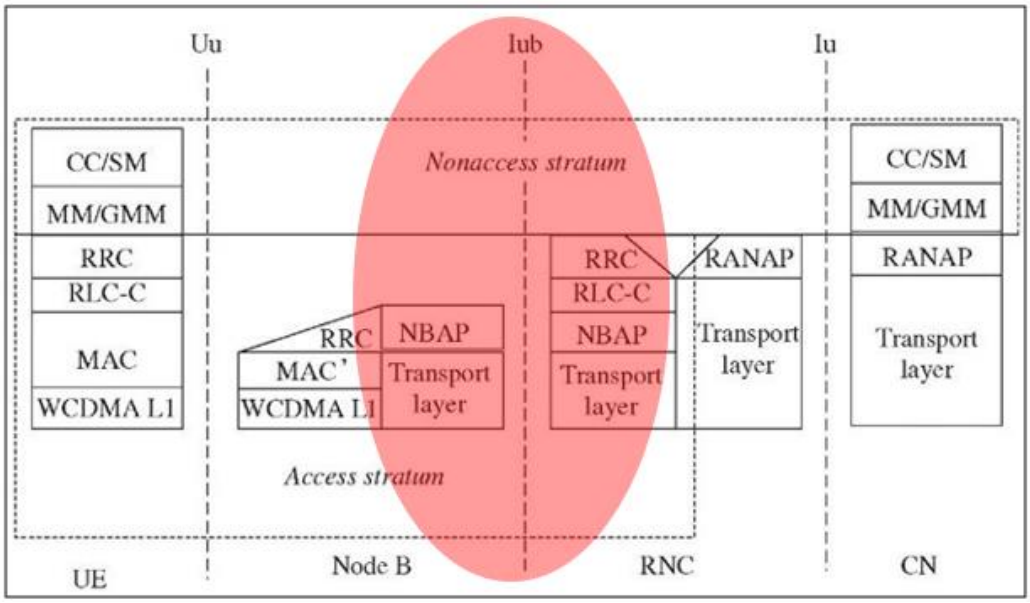


Figure 6: Iub interface protocol stack [24]

b) Iub interfaces messages and procedures

Most of the 3G radio procedures are transported through the RRC protocol. Some of them are:

1. Upper layer messages routing for mobility management (MM) and session management (SM) to ensure the communication between the UE and the core network.
2. Radio bearer management
3. UEs paging
4. System information broadcasting
5. Handovers management
6. Power control
7. Lower layer configuration
8. Measurement reports management

Any upper layer procedure requires a prior RRC establishment to ensure the communication between the UE and the RNC. After the RRC connection request from the UE to the RNC via the Node-B, the NBAP protocol sets the radio links between the Iub bearer and the frame protocol (FP) before completing the RRC setup as shown in Figure 7 [25].

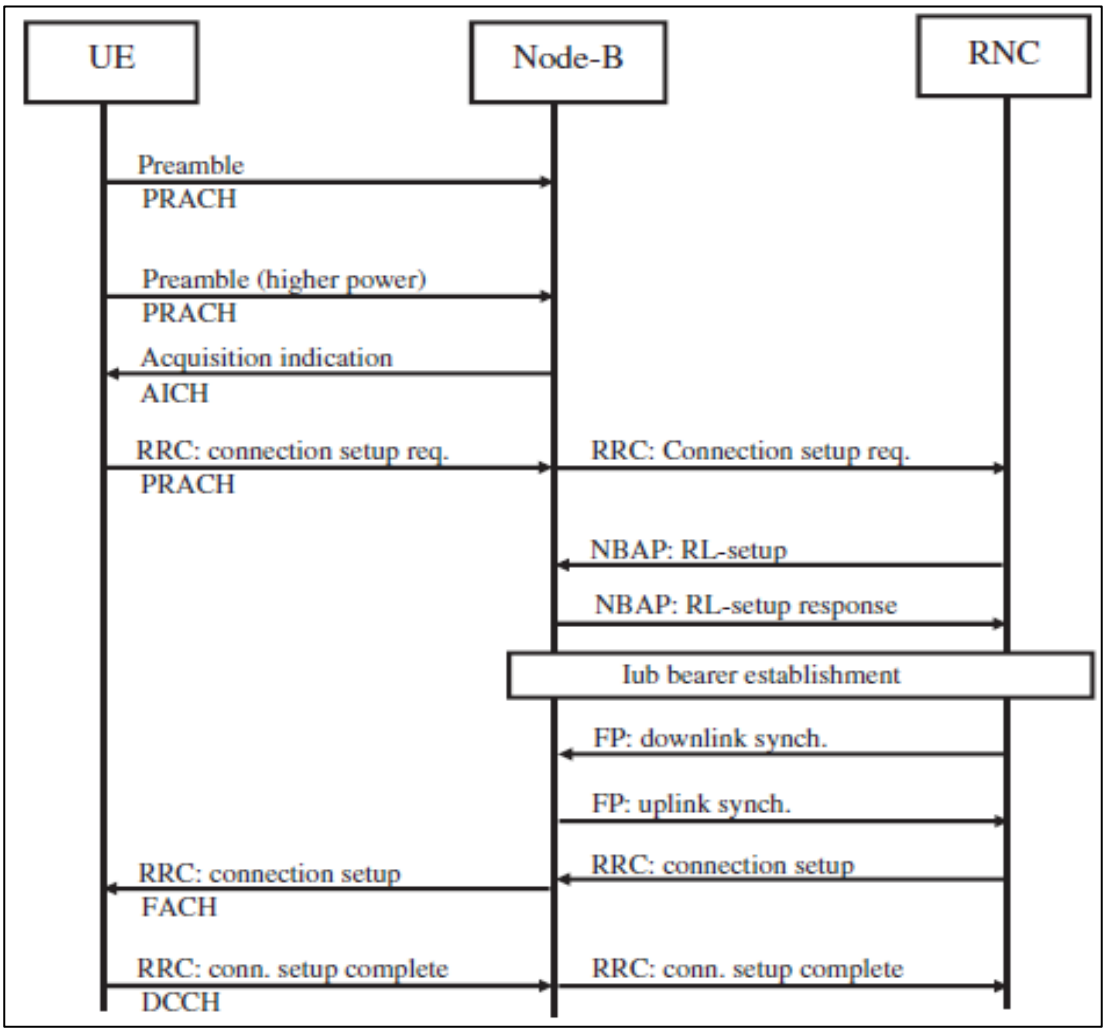


Figure 7: RRC establishment steps [25]

Among the important messages for the radio optimization, there are the measurement reports. These are messages exchanged over the RRC layer to enable communication between the UEs and the RNC. The RNC sends sets of conditions to the UE via the measurement commands and expects periodic and event-triggered measurement reports from the UE to request for the appropriate actions from the RNC after any of those conditions are modified [26]. Table 2 summarises the measurement reports groups and their trigger conditions.

Table 2: 3G Measurement reports summary

Measurement Type	Event-ID Group	Typical Tasks
Intra-frequency measurement	e1	Triggers the softer or soft handover if necessary
Inter-frequency measurement	e2	Triggers the hard handover if necessary
Inter-RAT measurement	e3	Triggers handover from the UTRAN another technology if necessary
Traffic Volume measurement	e4	Triggers the change of RRC state while the packet data protocol (PDP) context stays active (channel type switching)
Quality Reporting	e5	Informs the source RNC that a predefined number of cyclic redundancy check (CRC) errors is exceeded on the UE side
UE Internal measurement	e6	Delivers the information about UE Tx power (i.e.: If maximum Tx power is reached)
UE Positioning reporting	e7	Informs the network about problems with positioning

2.5.2. Gn interface

The Gn is the interface between two GPRS supporting nodes that could be an SGSN or a GGSN within the same public land mobile network (PLMN) and is equivalent to the Gp interface in case of different PLMNs interconnection while keeping the same protocol stack.

a) Gn protocol stack

The Gn interface protocol stack consists of the transport and the application layers. The transport layer consists of the physical, the data link, the IP and user datagram protocol (UDP) layer while the application layer consists of the GPRS tunnelling protocol (GTP) as shown in Figure 8 [27].

The GTP is divided into two sub-protocols which are the GTP-C for the control plane and the GTP-U dedicated for the user plane [27].

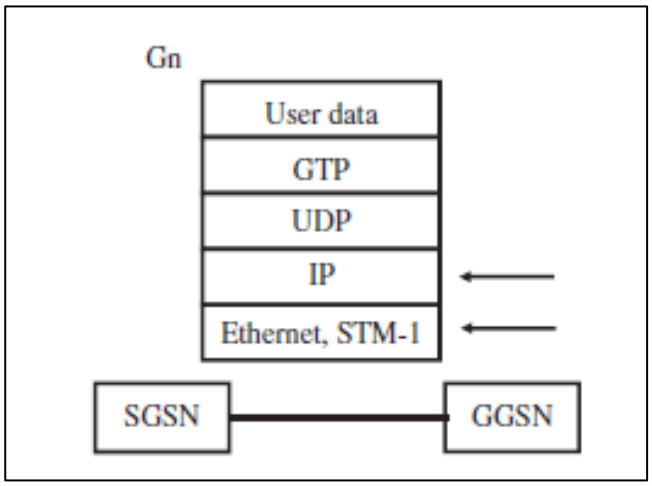


Figure 8: Gn interface protocol stack [27]

b) Gn interface messages and procedures

The main procedures of the GTP-C are the GPRS mobility management (GMM) and the session management (SM). The GMM takes care of the mobility procedures such as the location update to track the serving cell of the UE [28]. The SM controls the sessions by means of PDP context establishment, update and deletion. The PDP context request message contains information such as the QoS, the international mobile subscriber identity (IMSI), the international mobile equipment identity (IMEI), the network elements information (The SGSN and the GGSN IPs) and the access point name (APN) [28]. The PDP Context establishment steps are provided in Figure 9 [27].

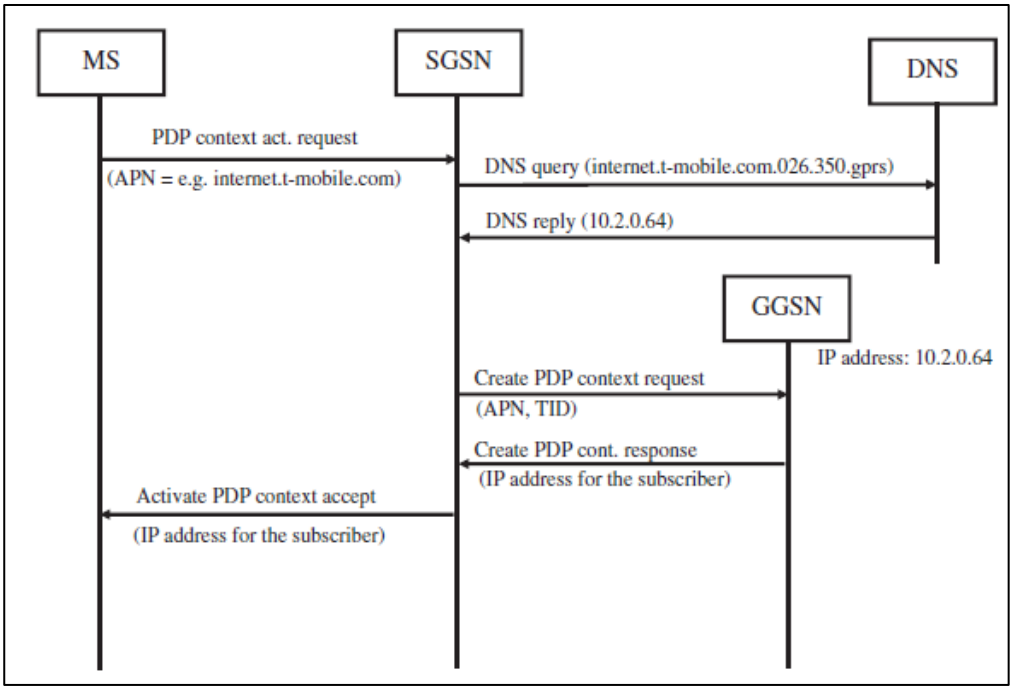


Figure 9: PDP context establishment steps [27]

The user plane of the GPRS tunnelling protocol (GTP-U) carries the data of the subscribers related to the PDP context created. The packets transmission is regulated according to the QoS negotiated during the establishment of the PDP assignment. The QoS configuration depends on the types of subscribers, the types of services used: Browsing, video streaming or peer-to-peer (P2P) [29]. Some of the important KPIs used to measure the performance of data services are the throughput, the latency and the retransmission. To get to a deeper level of applications classification, the GGSN is often extended with the deep packet inspection (DPI) engine which uses either the domain name system (DNS) requests for the uniform resource locator (URL) resolution from IP addresses or the signature and the pattern of the traffic to detect the types of application or protocols used.

2.6. QoS and QoE in packet-switched mobile networks

Several researches have been done to evaluate and optimize both the QoS and the QoE in the packet-switched mobile network environment. Ouyang and Fallah [23] presented a study based on statistical models for a packet-switched network operation. The simulation used a service call generator tool to trigger different traffic types and collect performance from the GGSN. The main QoS performance assessed were the average utilization, the latency, the packet loss, the throughput and the DNS failures. Rawal and Gupta [30] used OPNET modeler to simulate a

packet-switched mobile network with QoS parameters such delay and throughput. The paper demonstrated that the configuration parameters have an impact on the overall QoS of the network. Schwind et al. [16] built a testbed that consisted of 250 access nodes hosted in volunteers' home and for mobility reason in public transport cars, buses and trains. The traffic was sent to the core network of three CSPs in three different European countries and the user measurement results were periodically transferred to a centralized repository. The correlation was then established between the experimental measurements and the QoS of the network to derive the link between the QoS and the QoE.

Unlike previous researches focusing only on simulations and lab tests, the research done by Casas et al. [31] considered another source of information for QoE based on crowdsourcing to receive real users experience feedback from the users' devices. Using crowdsourcing applications on end-user's devices is getting popular. Not only for network authority regulators (NRA) but also for the CSPs. Some of the common crowdsourcing applications are Netalyzer, YoMoApp, Mobiperf. The results from the passive monitoring could therefore be correlated with the crowdsourcing application to demonstrate the link between the QoE and the QoS for a packet-switched mobile network. The metrics used to measure the QoE was the mean opinion score (MOS) which usually ranges from one to five (1 to 5). In this case, 1 is considered as being a poor QoE and 5 being an excellent one.

Recent academic developments such as the one done by Monserrat et al. [7] used questionnaires to determine the NPS which is a metric model used to evaluate customers satisfaction and loyalty through rating of 0 to 10 based on subscriber's perceived recommendation of the consumed product or service. Upadhyaya et al. [32] proposed a different model of collected QoE for the web services based on online reviews of the perception of the services delivered. Thousands of online reviews were analysed to determine the QoE attributes for the web services and create the relationship with the QoS parameters. To enable simple feedback, the online review was captured in native language, and was collected using crawling on the web, then saved in a database for text analysis. The text analysis consisted of categorization of words for positive feedback such as beautiful, nice, happy; and a negative feedback, based on words such as bad, terrible, disappointed. The reviews were clustered to determine the factors that influenced the user's perception and the QoE in correlation with the QoS parameters. The study done by Fiedler et al. [33] presented, through longitudinal user study, the relationship between the QoE, the volume of data usage and the churn risk. The study

demonstrated that most of the churners tend to have negative QoE followed with a less usage of the network before a potential churn.

2.7. Introduction to data mining techniques

Machine learning involves several domains of interest in the telecommunication environment for data mining and predictive analytics. While data mining involves procedures to discover trends and hidden patterns within large datasets, predictive analytics focuses on extracting useful information from those datasets to perform predictions or estimation about potential outcomes in the future [34]. Figure 10 [100] shows different intersections between data mining and other fields.

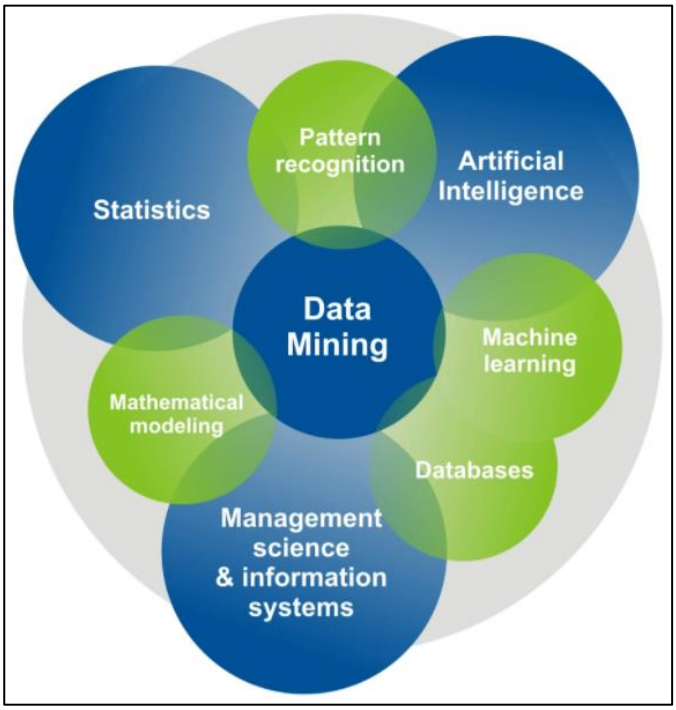


Figure 10: Data mining ecosystem [100]

The telecommunication environment uses both supervised and unsupervised machine learning algorithms. Supervised machine learning algorithms such as regression, statistical analysis and classification imply providing sample of results to the learning model during the training phase to derive the link between the predictors and the prediction. Unsupervised machine learning algorithms such as clustering involve finding patterns from data without prior knowledge of the results [35], [36]. Some of the machine learning algorithms used for classification are the linear discriminant analysis (LDA), the K-nearest neighbors (KNN), the classification and regression

trees (CART), the random forests (RF) and the support vector machines (SVM). Depending on the types of data to be used, the distribution of the data, the number of examples in the dataset and the hardware resources, machine learning algorithms can behave differently.

2.7.1. Linear discriminant analysis (LDA)

The LDA is a statistical approach based on a generalization of Fisher's linear discriminant used for both classification and dimension reduction. LDA is based on variance analysis where projection is used to characterize or separate classes of data after finding the linear combination of the features [37].

2.7.2. Classification and regression trees (CART)

The CART method is one the decision tree algorithms used to produce strictly binary decision trees where each decision node has 2 branches. Through recursive partitions, training data are grouped into subsets based on similarities of values for the target attribute. The tree grows following an exhaustive search of available variables and a possibility of splitting values to select the optimal split [34].

2.7.3. K-nearest neighbors (KNN)

The KNN is an instance-based learning algorithm that can be used for classification and numeric prediction. The algorithm classifies labelled data into categories and each unlabelled record is categorized based on the similarity with the nearest k records in the training dataset and assigned to a specific class that includes most of the nearest neighbors [34].

2.7.4. Support vector machines (SVM)

The SVM method uses the concept of hyperplane to set the boundaries between the data points representing learning examples and their feature values to divide a high-dimension space into homogeneous partitions. The SVM can be used both in classification and numeric prediction learning tasks. Since multiple hyperplane could separate the data, a search for the maximum margin hyperplane (MMH) is required to achieve the greatest separation. The goal is to maximise the margin by creating the largest possible distance between the separating hyperplane and the instances on either side. Both the linear and the non-linear algorithms exist to define the hyperplane [35].

2.7.5. Random forests (RF)

The RF is an ensemble-based technique that is based only on ensembles that are decision trees and a voting model to combine the trees' predictions. The RF method uses both the bagging principle and the random feature selection for additional diversity on the decision tree models [35].

2.7.6. Artificial neural network (ANN)

The ANN is a machine learning technique that attempts to imitate the structure of the natural neurons as shown in Figure 11 [97], to fire non-linear learning tasks. Just like the brain is made of several cells called neurons that are interconnected to each other's to handle stimuli from sensory sources massively processed in parallel, the ANN is an extremely powerful machine learning method that uses interconnected neurons or nodes to model the relationship between a set of input signals and the output signal [38]. ANN machine Learning methods are nowadays intensively used both for supervised learning (classification or numeric prediction) and for unsupervised learning (pattern recognition) and the domains of application range from speech recognition, image processing, self-driving cars, etc. [34].

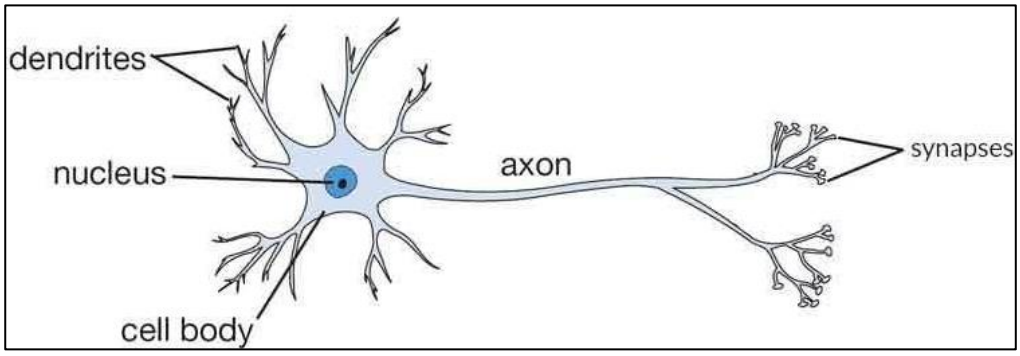


Figure 11: Biological Neuron [97]

The ANN artificial neurons as shown in Figure 12 [97] pass the linear combination of inputs and weights to an activation function to simulate this nonlinear behaviour of firing when a specific threshold is reached [35].

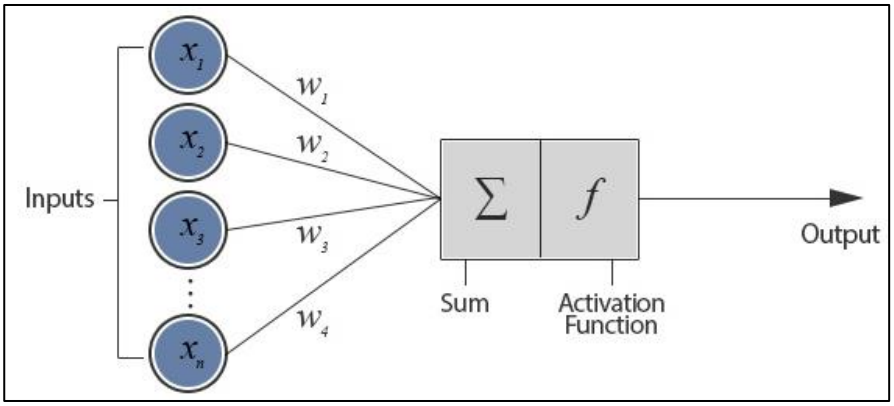


Figure 12: Artificial Neurons [97]

A typical ANN with n input signal can be represented by the following formula:

$$y(x) = f\left(\sum_{i=1}^{n} w_i\, x_i + b_i\right) \tag{1}$$

With x_i as set of inputs, w_i as set of weights, n as the number of observations during training, b_i as set of bias values, f as the activation function, $y(x)$ as the output.

a) Determination of the number of hidden neurones

Several researches have been done to determine the number of hidden neurons for an ANN knowing the number of the input and output nodes.

Heaton [48] proposed a method called the thumb's rule as follows:

1. The number of hidden neurons is between the number of input neurons and output neurons.
2. The number of hidden neurons should be equal to 2/3 of the sum of the number of input neurons and output neurons.
3. The number of hidden neurons should be less than double the size of input neurons.

The researches done by Li et al. [49] as well as Sheela and Deepa [50] proposed an arbitrary functions to obtain the number of hidden neurons based on the number of input nodes while

Shibata and Ikeda [51] as well as Hunter et al. [52], considered both the number of inputs nodes and the output nodes in the arbitrary function to provide the number of hidden neurons.

Interestingly, Vujicic et al. [53] provided a comparative analysis of ANN topologies with different number of hidden neurons. The metric to benchmark the topologies was the mean squared error (MSE) and the number of hidden neurons was calculated by using arbitrary functions such as the ones proposed in [49], [50] and [51]. Two different datasets were used to train the models. The first one was with smaller number of input nodes and second one was with larger number of input nodes. The study concluded that, the methods that performed well on a small dataset did not necessary perform well with the larger ones.

b) Deep Learning

Deep learning is one of the machine learning techniques that put focus on the data representations and features learning instead of individual tasks for both supervised and unsupervised learning [39]. The deep learning training methods require large amount of data and several iterations to reach the convergence. These methods perform better than traditional algorithms and take advantage of the hardware infrastructure for parallel computing or graphical processing unit (GPU) computing power. Often selected an excellent choice to solve difficult tasks with large sets of data, deep learning has gained a lot of popularity ranging from speech recognition, natural language processing (NLP), image recognition, machine translation. The baseline of deep learning algorithms is the usage of multilayer perceptron to represent high level representations into simpler ones. Deep learning architectures include the deep ANN that will be referred throughout this research as deep neural network (DNN), the deep belief networks (DBN), recurrent neural networks (RNN) and convolutional neural networks (CNN). Those architectures can be used for different types of classification problems [40].

c) Deep neural network (DNN)

The DNN which is a deeper structure of multi-layered neural network as shown in Figure 14 [43], can be considered as an extension of the multi-layer perceptron (MLP) shown in Figure 13 [43], but with more than one hidden layer to ensure the non-linearity within the distribution of the data is captured. The DNN can achieve better performance compared to the shallow networks since it can extract high-level abstraction from low-level features or the raw data [40]. Multiple hidden

layers involve additional complexity to the model and the training process, therefore deep learning techniques are required to ensure hyper parameters tuning [41].

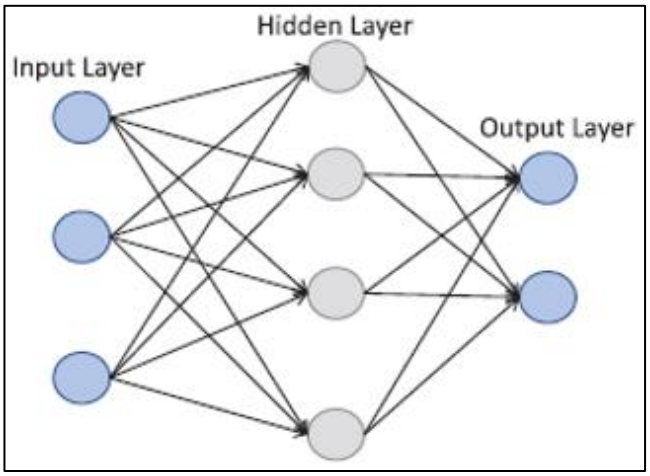


Figure 13: Shallow ANN architecture [43]

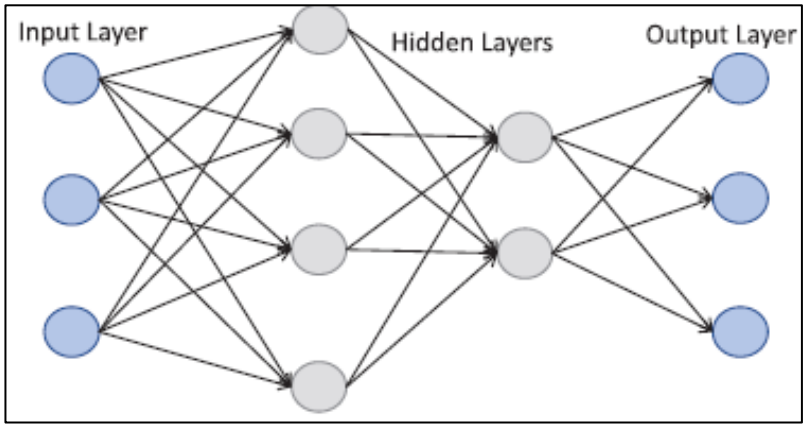


Figure 14: DNN architecture [43]

d) Deep belief network (DBN)

The DBN is a deep generative classifier commonly used to model deep learning algorithms with effective results as unsupervised learning method involving learning features from high-dimensional and complex datasets [42]. The core of the its structure is composed of restricted Boltzmann machines (RBM) where a greedy training is separately done for each layer considering information learned from the previous layer as input. An RBM is a probabilistic model

based on graph with one hidden layer and without direct connection between visible units or between hidden units [40] as shown in Figure 15 [43]. To tune the DBN, the backpropagation or the SVM can be used in a supervised or semi-supervised stage as intermediary layers [41]. But one of the drawbacks of the RBM is the tractability of the joint distribution [47].

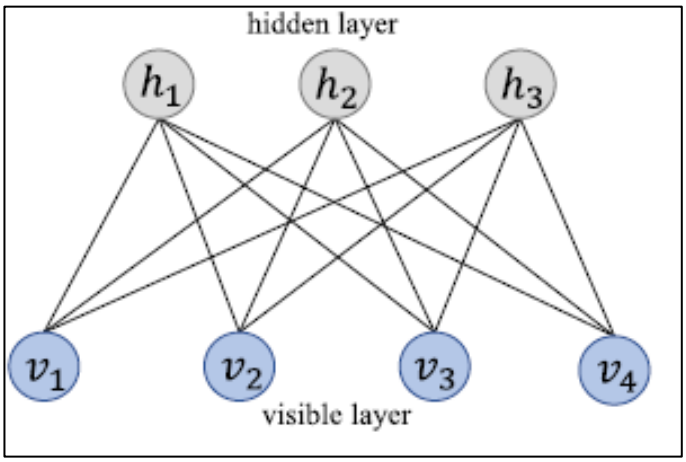


Figure 15: RBM architecture [43]

e) Recurrent neural network (RNN)

The RNN is a deep architecture with a feedback looping between layers as shown in Figure 16 [43]. There are two types of RNN models, the Elman and the Jordan models. The Elman model is based on a simple feedback, looping layer by layer. While the Jordan model has a feedback looping throughout all the nodes of a layer up until the next layer [42]. The RNN architecture has direct cyclic connections between internal neurons. Generally, RNN uses backpropagation learning, to processes input sequences by using the internal memory to keep a link between the output and its previous computation. As such, RNN is suitable for dynamic temporal behaviour analysis such as stock market prediction, time series prediction [45] and modelling sequence of data such as text and speech analysis [43].

One of the variants of RNN is the long short-term memory network (LSTM) which is easy to train since it solves the problem of the vanishing gradient [46]. The LSTM is a deep learning architecture suited for complex patterns prediction and can be used for both supervised and

unsupervised learning. training LTSM models requires a large amount of data and several iterations before converging [41].

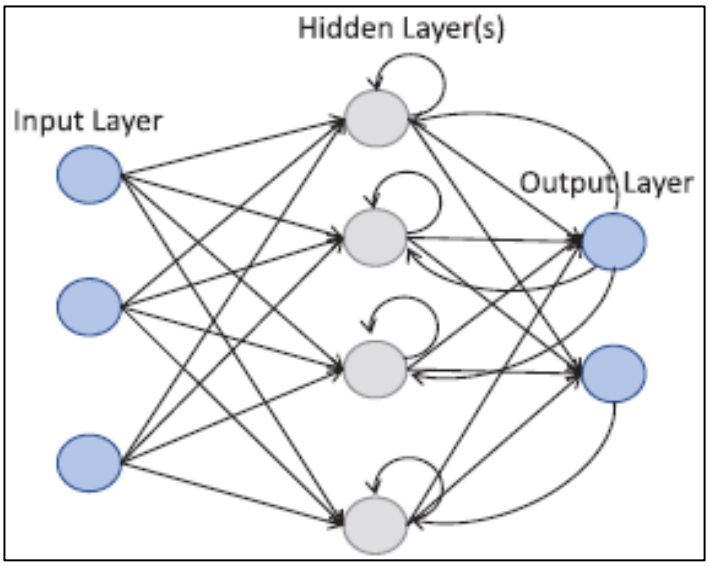


Figure 16: RNN architecture [43]

f) Convolutional neural network (CNN)

The CNN is a descriptive deep architecture containing an array of one or more convolutional and pooling layers to create a multilayer neural network [42] as shown in Figure 17 [101]. The convolutional layer is made of learnable filters also referred to as kernel [46], that extracts location invariant patterns from input objects such as images to detect specific types of features [45]. The pooling layer on the other hand performs translational invariant properties by sampling the output from the convolutional layer. The CNN has an advantage of easy training, since it has less parameters as compared to other connected networks with the same number of hidden nodes [42]. Although there is a similarity between the CNN and the ordinary artificial neural networks as they both have neurons with tuneable biases and weights, the CNN has better performance to overcome the burden of dimensionality that causes other deep learning structures to underperform when the input becomes very large and complex like in high resolution images [43].

The CNN is used in several applications that include video analysis and natural language processing (NLP) [45]. Others are robotics, speech and image recognition where some popular CNN approaches such as LeNet, AlexNet, VGG, RestNet and GoogLeNet are used [44].

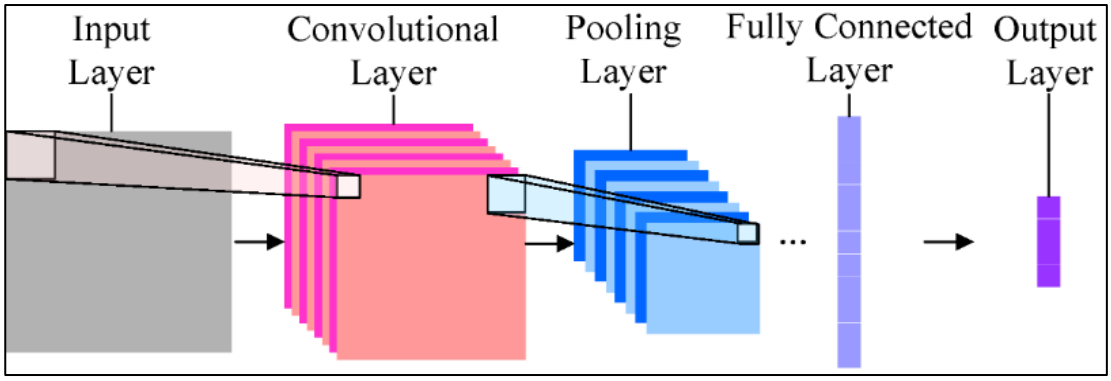


Figure 17: CNN architecture [101]

2.7.7. Data mining techniques performance evaluation

There are different methods to evaluate the performance of machine learning models depending on whether the problem is a regression one or a classification one. Since the focus of this research is on classification methods, we will discuss only the performance evaluation metrics for classification methods. Most of the classification evaluation metrics used for prediction were based on the confusion matrix (Which is a summary table benchmarking the actual values or reference value with the predicted value). Table 3 shows a generic confusion matrix [35]:

Table 3: Generic confusion matrix

	Reference	
Prediction	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

1. TP: True positive, amount of correctly predicted positive class
2. FP: False positive, amount of wrongly predicted positive class
3. TN: True negative, amount of correctly predicted negative class
4. FN: False negative, amount of wrongly predicted negative class.

Some common evaluation metrics that were defined based on information from the confusion matrix are:

1. The accuracy as shown in equation (2), which is the success rate of the prediction model.
2. The precision as shown in equation (3), which is the positive predictive value demonstrating the correctness of the predicted positive class.
3. The sensitivity as shown in equation (4), which is the true positive rate.
4. The specificity as shown in equation (5), which is the true negative rate.
5. The AUC which is the area under the receiver operating characteristic (ROC) curve as shown in equation (6), provides a statistic to determine the perfection of the classifier based on the shape of the curve as a better identifier of the positive values.
6. The F1_Score as shown in equation (7), which is the harmonic mean of the precision and the recall (A metric describing how complete the results are).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{AUC} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{6}$$

$$\text{F1\_Score} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{7}$$

2.8. Data mining techniques in radio networks

Several studies have been done both for radio QoS and machine learning algorithms to enhance the radio optimization and capacity planning. Charoenlap and Uthansakul [55] proposed a method of tracking some of the common issues that are related to the radio part of the 3G network such as the interference and the coverage. While network testing procedures such as drive testing can help to detect specific areas with no network coverage, drive testing, however, cannot be performed over the whole network because it is expensive in terms of human and equipment resources. The data from the drive test was used to create a correlation between the interference, coverage and the data throughput. The main task was to improve the network QoS at a proportionately lower cost. The proposed model estimated the network throughput based on the data from the drive test, by capturing the comparison of energy per chip and noise (EcNo), received signal code power (RSCP) and the received signal strength indicator (RSSI) on the radio interfaces under specific geolocation coordinates obtained during the drive test. This was done with a view of identifying holes (Areas without coverage) and the areas with higher interferences on the network. From the drive test data, a model was then developed to determine the correlation between the radio conditions and the data throughput and was later used to predict the throughput knowing the input radio conditions.

The studies done by Octora and Iskandar [56] as well as Zhang and Yang [57] provided detailed methods for the optimization of the HSPA+ technology based on the measurement reports. The approaches focused on the radio condition metrics such as the RSCP, the EcNo and the KPIs related to the network capacity such as call setup success rate (CSSR) for both the voice and HSPA services, and the average downlink throughput per user. The authors in [56] and [57] were able to demonstrate that radio issues can have an impact on the QoS. They later suggested that by adjusting typical antenna parameters such as the antenna height, the tilt or the reconfiguration

of the transmitted power, the radio conditions could be optimised so as to achieve the QoS level required. Huang et al. [54] focused on the usage of machine learning techniques for radio optimization and proposed a method to mitigate the complexity growth of new base stations by creating a data-driven optimization framework for both offline and online modules. For the offline module, the study employed a dimension reduction method using hierarchical clustering analysis and the outcome was analysed using an ANN for optimal prediction of groups of base stations. For the online module, the K-medoids clustering algorithm was used to segment the groups of base stations for performance optimization. As such, to improve the planning and the optimization of the resources in the network. Lawal et al. [58] proposed a data traffic forecasting model using an ensemble of neural network models with firefly optimization and resilient propagation algorithms. A comparative analysis with other algorithms such as the support vector regression (SVR) and the group method of data handling (GMDH) based on abductive network methods was used for data traffic forecast, providing a better performance.

Researches based on the measurement reports analysis were conducted by Zhou et al. [59] and Moysen et al. [63]. The authors in [59] focused on the influence of multiple radio access bearers in relation to the high demand of packet-switched sessions as compared to the circuit-switched sessions. The measurements in the network were analysed using Adaboost, so as to derive reasons for potential call drops. The reduction of the number of call drop optimized the bad performance for both the voice and the data QoS. The authors in [63] instead, focused on the QoS prediction techniques so as to facilitate the planning for the network operators in heterogeneous networks. The measurement reports from the radio interfaces were collected and used to develop models based on correlational measurements. This was done with a view of improving the network planning from a QoS point of view. In essence, a comparison was done between different regression techniques for different types of features, while applying dimension reduction techniques as well.

Chen et al. [60] investigated diverse factors that impact the performance of a 3G cellular network on the RNC level from the packet loss rate and the round-trip time (RTT). The study applied a supervised machine learning with the RuleFit algorithm that combines both the decision tree and the linear regression. Roshdy et al. [64] proposed models based on correlation, clustering and regression. The cells in the network were classified according to their priorities and their QoS requirements to monitor the capacity of each group using the throughput limitation and to determine the load in order to be able to share the resources between cells in case of low usage.

2.9. Data mining techniques in core networks and traffic classifications

As the data traffic growth is largely influenced by video data usage, video traffic classification and categorization for QoS management or security have been a subject of intense researches. A web-service approach was presented by Kumar et al. [61] to predict the response time and the throughput based on the QoS parameters from the past data usage behaviour. A comparison of several machine learning algorithms, including bagging and SVM were applied to the dataset. The bagging and SVM performed better than the other algorithms. The model performance was assessed using the correlation coefficient, mean absolute error (MAE) and the root mean square error (RMSE). Nikravesh et al. [62] applied data analysis methods to maximize resource management for mobile service providers to avoid under-provisioning or over-provisioning of network capacity. Several machine learning techniques were applied which included multi-layer perceptron (MLP), multi-layer perceptron with weight decay (MLPWD) and SVM. The SVM technique was selected as the best method in handling multi-dimensional data traffic for that mobile system.

Garcia [65] applied unsupervised learning algorithms for clustering data such as K-means and density-based spatial clustering of applications with noise (DBSCAN) to analyse the customer traffic flows and the behaviour produced from the heavy-hitter segment of customers. From the data containing packets information and the DPI classification, four to six clusters were identified based on the flow behaviour and a subset of them represented the traffic that was not based on the video transfer. The research done by Trivedi and Patel [66] considered the unique byte patterns that are used within the DPI systems as the signature to detect different traffic applications that require regular maintenance through the updates of the DPI signatures. To reduce the laborious manual operations leading to errors during signatures update, they proposed an automated solution for the DPI signatures verification based on machine learning techniques. By using open source mobile automated tools for mobile application traffic generation, the signature patterns for undetected flows using well-known ports and machine learning algorithms reduced the time taken to do a signature update.

An analysis of the data flow between the client and the server to ensure efficient bandwidth usage was done by Kaoprakhon and Visoottiviseth [67]. The study provided a method of classifying non-encrypted audio and video traffic over hypertext transfer protocol (HTTP). The packet flow information was used to observe audio and video traffic. The method built flow profiles using the flow duration, the average received packet size and the server-client packets.

Huan [68] presented a video packet classification algorithm in the field of video surveillance and video management that requires fast video packet identification. The XOR and shift operations were used in combination with the linear hash function to filter video packets from the network traffic. The results of the experiment provided better performance for higher traffic network compared to simple rules of filtering methods.

For the efficient management of the network resource and the QoS required for video services, Dong et al. [69] proposed a fine-grained classification algorithm for video traffic on the internet using the hierarchical clustering technique based on a combination of statistical features from the QoS and the network resources requirements. The method provided better performances for the recall and the f-measure.

Within the domains of guaranteed QoS, Zai-Jian et al. [70] used the concept of QoS based flow aggregation that consisted of different QoS classes with features from the downstream and the upstream rates. Because of the sparsity of the multimedia QoS, the authors suggested a modified K-singular value decomposition (K-SVD) classification framework with an SVM classifier. The research demonstrated that the downstream and the upstream rates provided good features that could be considered for video traffic classification.

Extant literature such as the one by Dubin et al. [71] used machine learning algorithms to classify video titles of encrypted HTTP adaptive streams from popular videos based on Youtube video streams. The classification could handle long delays and high packet losses and proved to be accurate in its prediction. Although the video streaming such as Youtube are mostly over HTTPS, the study was able to derive the pattern of streaming.

Recent developments in classification done by Nossenson and Polacheck [72] proposed statistical classification for video traffic live and video on demand (VOD). Generally, such types of video transmission require different optimization techniques such as multicasting for live video streaming and the usage of cache for the VOD streaming. Thus, the internet service provider (ISP) and the content delivery network (CDN), may require online classification of the two types of video streaming for proper optimization of resources. By using of the packets size and the video traffic source, two classifiers were proposed to separate live video streaming from VOD streaming traffic. The prediction evaluation metric used for the classification and to assess the performance of the two methods was the accuracy.

Shi and Biswas [73] used traffic analysis methods to design a firewall framework to block the BitTorrent traffic while mixed with other types of traffic such as video streaming. The solution used 2-steps classifiers to detect BitTorrent traffic and to identify video streaming traffic under an encrypted tunnel. Another video traffic classification method was proposed by Tang et al. [74] based on the multi-fractal's theory. The approach uses the fractal characteristics from physical calculations rather than statistical features extraction from traditional analysis resulting in a better performance than the Bayes networks, the SVM, the hidden Markov model (HMM) and the decision tree.

There have been researches conducted to use the deep learning techniques in the telecommunication environment. These studies range from the physical, the data link and the network layers up to the packet flow identification and the intrusion detection systems [75].

From the physical and data link layers, Peng et al. [76] proposed a deep learning method for modulation classification in communication systems. AlexNet which is one of the variants of the CNN deep architecture was used for training and testing of the model. The model used constellation diagrams which were image representation of the modulated signals. The deep learning results were compared with traditional modulation classification based on cumulant and SVM showing a closer classification accuracy but without the need for the laborious task of manual feature selection in the case of deep learning. Xu et al. [77] proposed an automatic configuration model using deep reinforcement learning that adapts from the traffic conditions to reduce the delay in the network for the routing optimization. A simulated environment based on the OMNeT++ simulator was built to assess the delay under different variation of traffic and routing. The model based on deep reinforced learning provided a better performance with smaller network delay in respect to the benchmark setup.

Some studies focused on the network layer for routing optimization problems. Kato et al. [83] used a DNN architecture to improve traffic control in heterogenous networks. The study proposed an approach to properly characterized the input and the output of heterogenous network traffic with a supervised DNN. The method demonstrated a good performance for the throughput and the delay compared to the open shortest path first (OSPF) benchmark routing approach.

With the introduction of programmable software-defined routers, to reduce the cost of the packet processing through intelligent methods, the deep learning has been introduced in studies such as the one done by Mao et al. [82], where a simulation was conducted using the DBN architecture

to characterize the input and the output traffic patterns. the results of the proposed method outperformed the existing benchmark method based on performance indicators such as the delay and the throughput.

Due to the rapid growth of the internet traffic, numerous studies have been done on the usage of deep learning for traffic classification. Lotfollahi et al. [80] proposed an approach to investigate both the feature extraction and the classification tasks using the same platform. The major class of traffics such as the file transfer protocol (FTP), the P2P and the E2E applications such as BitTorrent and Skype were categorized using the CNN. Wang [81] focused on the challenge of the feature's identification within the flow of data. Since most of the systems used for traffic identification use features such as the port number, the application signature and other statistical characteristics; this study proposed a method based on the ANN and deep learning model using a stacked auto-encoder (SAE) architecture to extract the features and classify the traffic.

The deep learning techniques have also been used in some researches to improve the network security against attacks. Tang et al. [78] built a DNN model to mitigate against security threats in SDN. The model proposed an intrusion detection system that applied a deep learning approach on flow-based anomaly detection. Although the research results were not good enough to outperform the existing network intrusion systems, however, it demonstrated the potentiality of deep learning in SDN environments. Gao et al. [79] used a DBN architecture to address Big Data classification for intrusion detection. The DBN model learned high dimensional representation while performing efficiently the classification tasks compared to other models based on SVM and ANN.

2.10. Big Data in mobile networks

The term Big Data is often used for large data management which is computationally expensive and difficult to handle using the traditional database management tools. Previously Big Data was relying on 3-Vs which were the volume, the velocity and the variety of the data. Today the industry refers to 5-Vs by adding the value and the veracity as shown in Figure 18 [98].

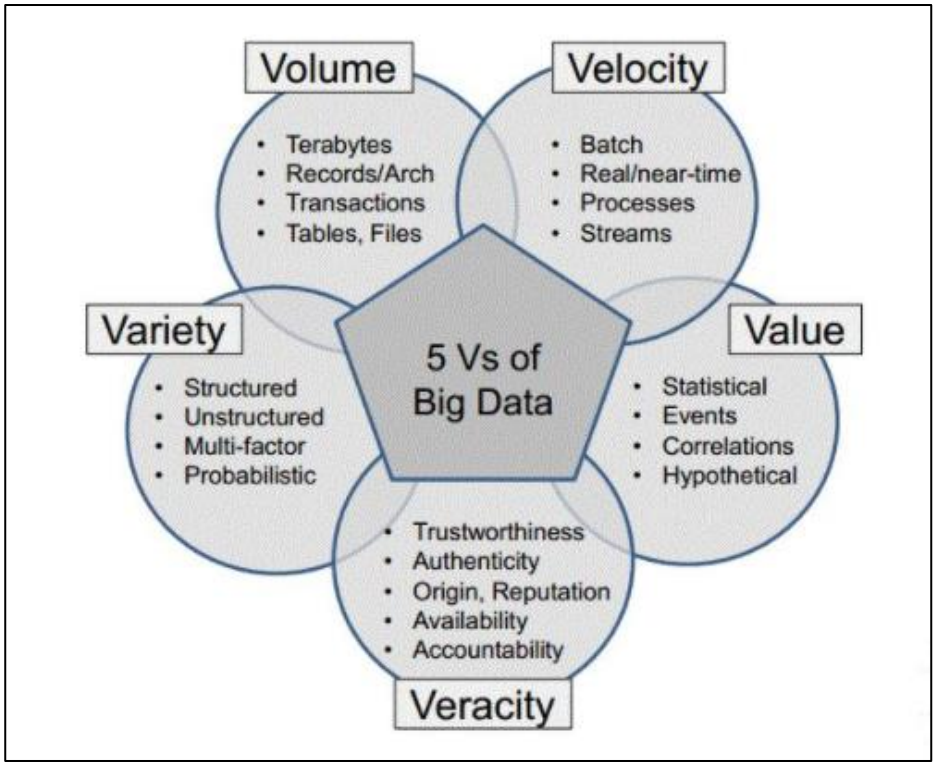


Figure 18: Big Data 5-Vs [98]

Some of the platforms handling Big Data are oracle DB2, EMC Greenplum, Vertica, Microsoft PDW, Teradata and Hadoop [84]. Hadoop is an open-source software platform implemented in Java programming language. It allows the store of large files on a single machine or in a cluster of computers for distributed processing of huge datasets. The main components of the Hadoop ecosystem are the Hadoop distributed file system (HDFS) and the MapReduce framework. The HDFS manages the storage of large files while the MapReduce is a technique used to distribute the tasks across several nodes by processing the input data and producing intermediate results in the Map-phase and merging the intermediate results having the same key in the Reduce-phase [85].

He et al. [21] proposed a unified data model for an architectural framework based on the random matrix theory and the application of machine learning techniques for Big Data analytics in mobile networks. The authors also illustrated examples of Big Data applications in mobile network such as data traffic, location, signalling, heterogeneous networks and radio waveforms. The research concluded with open research challenges of Big Data application in mobile networks such as data privacy, filtering and compression. Su et al. [20] proposed a Big Data platform to collect, process and analyse the large amount of data available in the mobile networks. A Hadoop-based

and a multiple parallel processing database architecture was used to achieve a unified management and storage system using the massive telecommunication data sources, to ensure proper maintenance and network optimization. The results of this study demonstrated a better performance of the Big Data platform in terms of data loading and analysis compared to the traditional data warehouse, providing the benefits of a Big Data infrastructure.

To enable the CSPs to manage the network resources in an effective and efficient way while supporting a better QoS, Si et al. [85] developed a Big Data analysis platform to analyse the mobile network data traffic patterns for the management of the resource usage of the network elements. two datasets were used with Apache Hadoop for the storage and Mahout for the machine learning algorithms. The algorithms included essentially the K-means and Fuzzy K-means for clustering. The results focused on improving the execution time by changing the Hadoop cluster parameters. Jun et al. [86] instead, collected core network data from a CSP's core network and proposed Zipf-like models to analyse the traffic volume, the exchanged requests between service providers and the subscribers' usage to characterise their distributions. The model essentially solved a time-series unsupervised clustering challenge by identifying the traffic patterns. The results of the study highlighted the users' behaviours leading to the traffic patterns and the service categories used.

Çelebi et al. [84] used a Big Data approach to analyse inter-radio access technologies handovers from the 3G to 2G networks. The study proposed an analysis of the A interface signalling messages between the base station subsystem (BSS) and the mobile switching centre (MSC). Due to the large amount of the signalling messages, a Hadoop platform was used to load the data into the HDFS and to run the queries using the Apache Hive to transforms structured query language (SQL) queries into MapReduce functions. The results provided visibility on 3G service holes (Areas with service discontinuity), outperforming the base station KPIs analysis approach based on the accuracy. Jie et al. [87] used a distributed computing Hadoop system to analyse high-speed network traffic from the massive data captured from a 3G network. The internet traffic from the smartphones were analysed to leverage a MapReduce parallel programming model with the objective to understand the usage patterns and the forecast growths of the network traffic. The data were collected using a traffic monitoring system deployed at the Gn interface between the SGSN and the GGSN. The results of this research provided flow characteristics of different smartphone operating systems and their related traffic, which could be useful for CSPs to anticipate the fast traffic growth in the network.

2.11. Root cause analysis (RCA) in mobile networks

The full automation of the processes in mobile networks management will still take time and therefore the support of human expertise is still needed. The mobile technologies knowledge plays a very important role into finding and solving problems in the network. But the evolution of technologies and the proliferation of handsets and services create a huge number of errors and faults in the network while increasing the scale of complexity for the incident management and the RCA.

Botta et al. [88] proposed an intelligent customer service assurance platform for mobile broadband network. To enhance advanced operations support systems (OSS), an architecture based on probes was used to improve the bit rate, to correlate the control and the user plane including a multidimensional and an RCA model. The result of the research was used in a real network to provide benefits on mobility and session management as well as transmission control protocol (TCP) connections.

Keeney et al. [89] proposed a recommendation system to assist the NOC operational team to manage incident occurring in the network. The approach consisted of a collection of telecommunication data from the OSS In an intelligent way to correlate them and add prediction for proactive maintenance. Kingsley and Dahj [90] proposed a tree-based SQM approach for efficient low-cost service management with a particular interest on the over-the-top (OTT) applications. The SQM-tree had four levels consisted of the 3G services classes i.e.: streaming, interactive, etc. The lower level of protocols and applications were available by drilling down from the service classes. The system connected to a cloud application to provide reporting throughput SparkSQL and to query the stored data in a Big Data framework allowing investigation of worst cells and subscribers. Fiadino et al. [91] also focused on the OTT internet services and developed a framework called RCATool. The RCAtool used the DNS protocol to detect and diagnose the traffic anomalies. The diagnostic features such as the device information, the error codes and the host name were used for the investigation of the root cause. The RCATool essentially employed two methods. The first one was applied to the entropy of the diagnostic features while the second one considered the statistical distribution of features such as the traffic.

Miyazawa and Nishimura [92] proposed an RCA approach to investigate services failures in a converged (fixed and mobile) network. The approach used alarms classification and a

hierarchical alarm data model on different types of alarms such as the resource alarms and the service alarms to pinpoint the causes of the failures and potential correlation between the alarms.

Chapter 3: Methodology

3.1. Introduction

The chapter describes the research methodology that was employed. It further outlines the model architecture and the data collection. The methods used to process and analyse the data are also described, including the approach implemented to validate and evaluate the performance of the system.

3.2. System Architecture

The physical architecture of the system consisted of a physical laptop (Computer1) running RStudio [93] and a virtual machine (VM) based on VMware [94] running a single node Cloudera platform with Hadoop as shown in Figure 19.

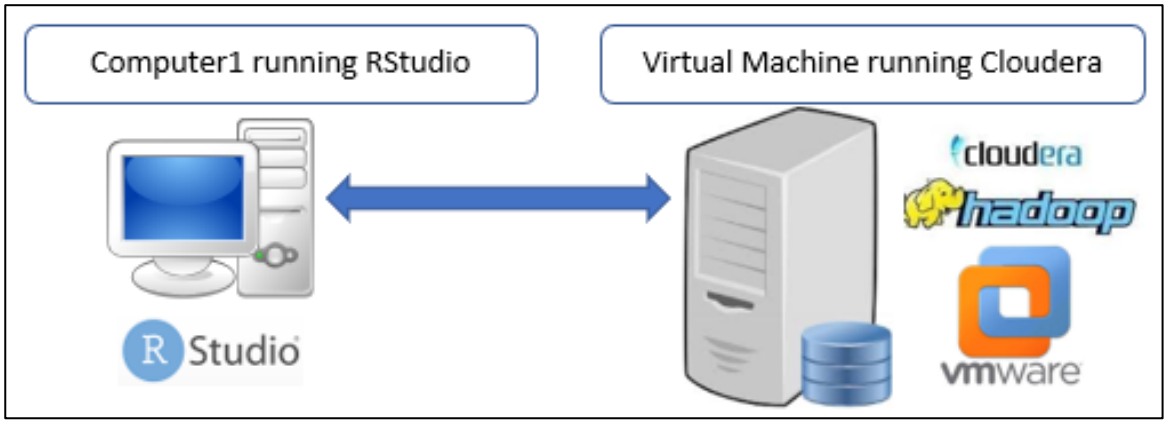


Figure 19: Physical architecture

Computer1 was used to prepare scripts and queries using R programming language, and to connect via a Cloudera Impala connector to the VM which contained the HDFS with the stored dataset files. Cloudera Impala [95] is a massively parallel processing (MPP) SQL query engine for Apache Hadoop, released via an Apache license to provide an open-source system.

The logical architecture of the system as shown in Figure 20 is made of:

1. Hadoop platform: The Hadoop platform via the HDFS was used to store the data in a format optimized for Big Data.
2. A parsing and aggregating module: This module was used to collect, parse and aggregate the data before loading them into the Hadoop platform.
3. Core engine: The core engine was used to design different models and implement the algorithms used for this research.
4. API using R programming language: the API was used to ensure the communication between the core engine and the Hadoop platform.

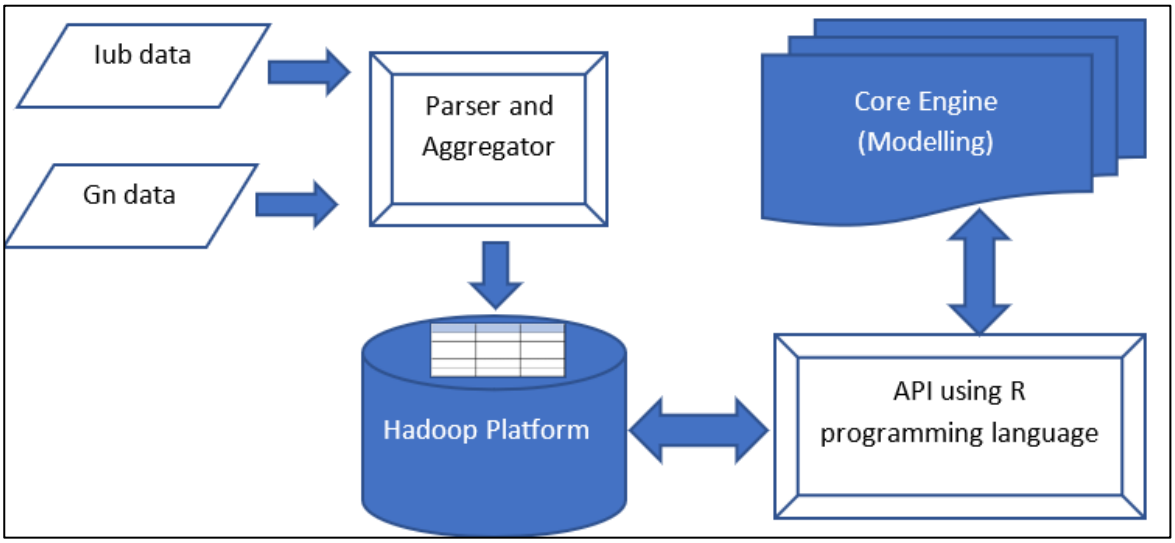


Figure 20: Logical architecture

3.3. Data Collection

We collected transactions data from the Iub and the Gn interfaces of a real mobile network operator. The raw data was parsed and stored in the Hadoop platform and based on the pre-processing requirements, aggregated into different datasets with a limited number of records to allow the training and testing of the models.

3.3.1. Iub Interface data collection

The data from the Iub interface was gathered from a full day collection of subscribers' related radio transactions on one RNC from an urban area. The details of the data collected from the Iub interface are shown in Table 4.

Table 4: Iub interface transaction data collection

field name	Description	Example
Timestamp	The time the transaction occurred in Unix format. This time is used to reference the transaction chronologically	1538025250
subscriber	The subscriber ID based on the International Mobile Subscriber Identity (IMSI). This is a unique identifier of a subscriber responsible of the transaction	******13992
handset	The device type based on the Type Approval Code (TAC). This is a unique identifier of mobile device used during the transaction	35195507
cell	The cell ID. This is unique identifier of the cell (network element) used by a subscriber for a specific transaction	******467
EcNo	The energy per chip and noise to provide information about the quality of the radio signal	-9 dB
RSCP	The received signal code power to provide the radio signal strength	-85 dBm
propagation delay	The distance at which the radio activity initiated	1 Km
Termination type of the packet-switched calls	The termination type provides details if the packet-switched transaction completed normally (Normal) or with a failure (Drop)	Drop

3.3.2. Gn Interface data collection

For the core networks, the raw data was collected for a single day on the Gn interface between the SGSN and the GGSN for both the uplink and the downlink traffic. The details of the data collected from the Gn interface are shown in Table 5.

Table 5: Gn interface transaction data collection

field name	Description	Example
Timestamp	The time the transaction occurred in Unix format. This time is used to reference the transaction chronologically	1538025197
service	The types of services used by the subscriber during the transaction (browsing, video, etc.)	browsing
subscriber	The subscriber ID based on the International Mobile Subscriber Identity (IMSI). This is a unique identifier of a subscriber responsible of the transaction	******13992
handset	The device type based on the Type Approval Code (TAC). This is a unique identifier of mobile device used during the transaction	35195507
cell	The cell ID. This is unique identifier of the cell (network element) used by a subscriber for a specific transaction	******467
dlsec_bins	String of second-based bytes usage on the downlink (60 values)	20854,1046,2359, 727,776,…,1698
ulsec_bins	String of second-based bytes usage on the uplink (60 values)	1251,25946,1202, 955,624,…,704
sec_dl	The active second on the downlink	60
retransbytes_dl	The bytes retransmission on the downlink	655594
bytes_dl	The bytes transmitted on the downlink	2451800
dns_successful	The number of success full DNS transaction	80
dns_failure	The number of failed full DNS transaction	0
latency_dl	The latency on the downlink	518ms

3.4. Facebook video traffic classification

While there are solutions known as the DPIs that dig into the packet-level to either understand the well-known ports, the IP addresses or signature patterns, the classification of services remains challenging. Social network platforms such as Facebook were usually classified as social media. But with the introduction of new features oriented to browsing and video services such as Facebook live and Facebook video streaming, there is a need for different QoS

management compared to the classical social media application. Adding to this, another challenging front in data classification due to the implication of encrypted traffic which could leave several sessions unlabelled or unknown. To enable proper classification of traffic, in this section we proposed a classification approach based on ANN for Facebook video traffic using second-based bytes usage on both downlink and uplink directions. From a sample of 4210 observations and 121 attributes, using different ANN topologies and varying the number of hidden neurons, the best model was selected based on the accuracy.

3.4.1. Data preparation

For every data session, transactions were produced every minute to generate an output of 121 attributes. The first attribute was the protocol name and the rest consisted of the byte's usage. The data was cleaned based on the protocol names to focus only on Facebook transactions and discard other protocols from the dataset. The selected 4210 observations based on Facebook transactions were further classified according an embedded DPI into chat (36%), browsing (37%) and video (27%) as shown in Figure 21.

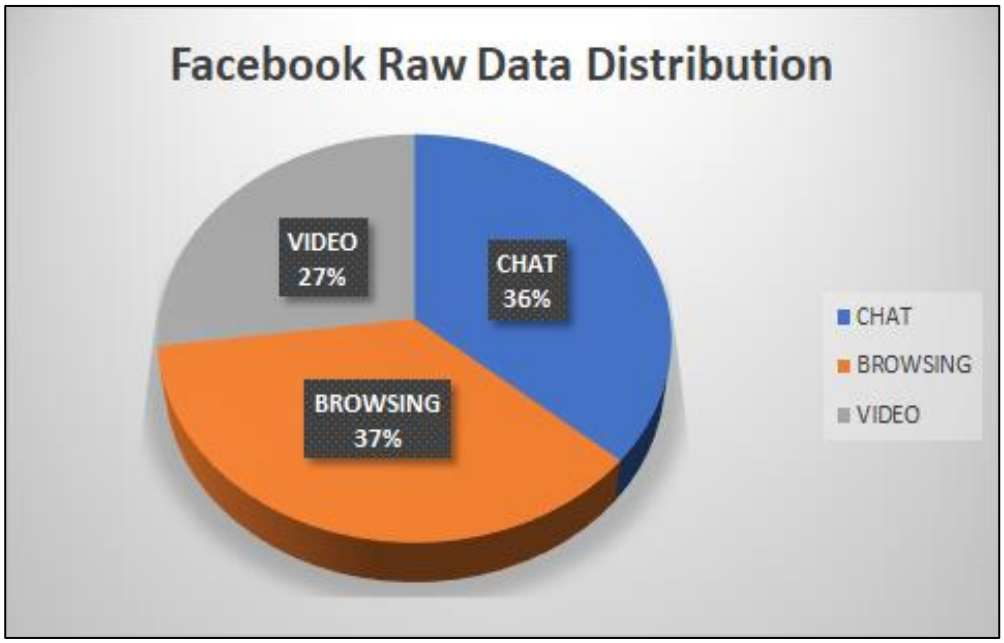


Figure 21: Facebook raw data distribution

Generally, over a relatively long period, video traffic is easy to classify since the average tends to be higher than the rest of the protocols. But for short period (less than 1 minute), classification by average can produce errors, due to outliers from other protocols, especially those from the

audio streaming and the browsing categories. Figure 22 shows some of the outliers that were observed from the collected data. This was done by checking the first byte's usage pattern with "dl" and "ul" representing the volume of the bytes used on the downlink and on the uplink respectively.

The ANN was selected as the machine learning technique to train the model for the classification of traffic, as it works well in modelling very complex patterns. As ANN works well with continuous variables rather than categorical values, we created two new outputs for the categorical variable using one-hot encoding with values of 0 for no and 1 for yes. It must be noted that having 2 outputs instead of 1 will not improve the results in case of binary classification but we have chosen this approach to enable dynamic configuration using one-hot encoding to support multiclass cases for the future. The full details of the final Attributes used for training and testing are shown in Table 6.

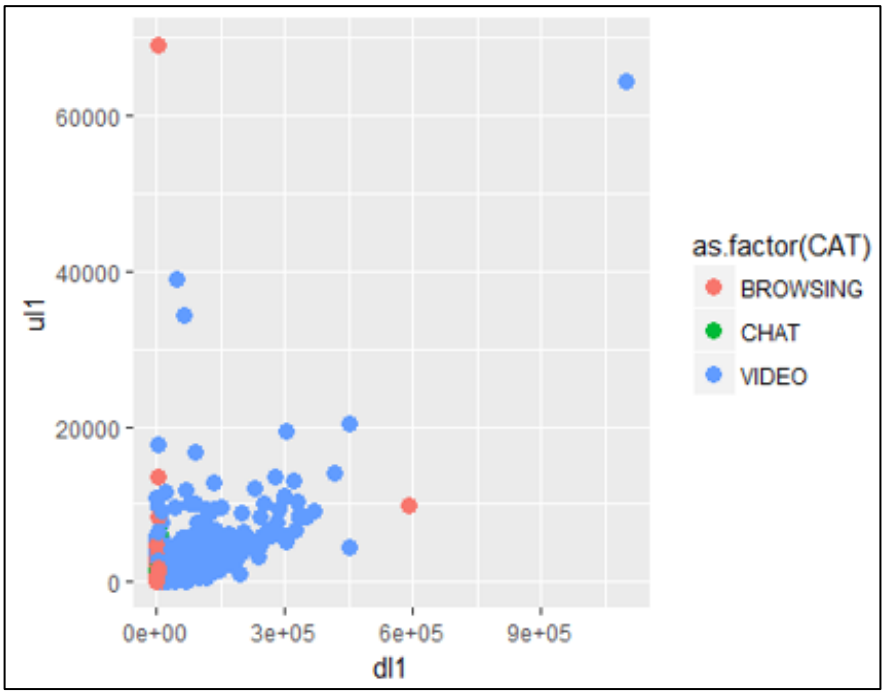


Figure 22: First- and second-bytes usage pattern

Table 6: Facebook video traffic classification attributes details

Attributes name	Description	Example
X1	Bytes down sec1	0
X2	Bytes down sec2	0
X3	Bytes down sec3	50
.	.	.
.	.	.
.	.	.
X60	Bytes down sec60	50
X61	Bytes up sec1	0
X62	Bytes up sec2	0
X63	Bytes up sec3	0
.	.	.
.	.	.
.	.	.
X120	Bytes up sec60	12
X121	Is Facebook-Others traffic? (1 for yes and 0 for no)	0
X122	Is Facebook-Video traffic? (1 for yes and 0 for no)	1

To ensure that the numerical values of the attributes are within the activation boundaries, while training neural networks, the normalization was applied to all the attributes except the targets output of X121 and X122 which were already designated as binary 0 or 1 based on hot encoding. The formula used to normalize is shown in equation (8):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

Where x_{norm} is the new value of variables after normalization, x_{min} and x_{max} are respectively the minimum and maximum values of the attributes in the dataset.

3.4.2. Model design

The dataset was divided into two new datasets with a ratio of 80% for the training dataset (3,368 Observations) and 20% for the validation dataset (842 Observations).

The ANN was used to train and compare four topologies based on the accuracy. The four topologies were multilayer-feedforward networks that used the logistic sigmoid function. The training algorithm of the ANN topologies was the resilient backpropagation algorithm with an error function as the cross-entropy. While all the topologies had I_N = 120 inputs nodes, O_N = 2 output nodes and 1 hidden layer, the main difference between them resides on number of neurons H_N calculated as follows:

1. Topology 1 was based on the arbitrary function as proposed by [50] Which resulted in H_N = 4 hidden neurons:

$$H_N = \frac{4{I_N}^2 + 3}{{I_N}^2 - 8} \tag{9}$$

2. Topology 2 was based on the arbitrary function as proposed by [52] Which resulted in HN = 5 hidden neurons:

$$H_N = \log_2(I_N + 1) - O_N \tag{10}$$

3. Topology 3 was based on the arbitrary function as proposed by [49]:

$$H_N = \frac{\sqrt{(1 + 8I_N)} - 1}{2} \tag{11}$$

And the one proposed by [51]:

$$H_N = \sqrt{I_N O_N} \tag{12}$$

Which both resulted in H_N = 15 hidden neurons.

4. Topology 4 was based on the thumb's rule as proposed by [48]:

$$H_N = \left(\frac{2}{3}\right)(I_N + O_N) \quad (13)$$

Which resulted in H_N = 81 hidden neurons.

For each of these topologies, a confusion matrix was generated to calculate the accuracy that was the chosen metric to select the best model.

3.5. A quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data

The network optimization and the incident management determine the level of maturity the CSPs since the reduction of the MTTR has a direct impact on the revenue, especially the OPEX. A fast RCA mechanism is therefore crucial to improve the efficiency of the operational team within the CSPs. This section proposed a quadri-dimensional (service, subscribers, handsets and cells) approach to build an SQM tree in a Big Data platform to speed up the RCA and prioritize the elements impacting the performance of the network. Two algorithms have been proposed to normalize the performance indicators and to build the SQM tree by aggregating the performance indicators for different dimensions and services to allow ranking and detection of tree paths with the worst performance.

3.5.1. Data preparation

With the objective to optimize the RCA which takes a lot of time, especially in the case of Big Data and the complexity of the services and other network related element, the system model is an SQM-tree approach where each node held information to enable sorting and prioritization of tree paths to understand which service, dimension and KPI is negatively influencing the performance of network.

From the Gn interface transaction data collection, an SQM file was built and stored in the HDFS containing 11 columns with 1 Million records aggregated based on 4 keys and 7 core network performance indicators. The 4 keys were the service, the subscriber based on the IMSI, the handset based on the TAC and the cell based on the cell-id The 7 core network performance indicators were the total number of events (events), the total time of data connection (sec_dl), the total bytes retransmitted on the downlink (retransbytes_dl), the total bytes transmitted on the

downlink (bytes_dl), the number of successful DNS transactions (dns_successful), the number of unsuccessful DNS transactions (dns_failure) and the latency from the core to the user equipment (latency_dl). The details of the SQM file is shown in Table 7.

Table 7 : SQM file fields description

field name	Description	Example
service	The types of services used by the subscriber during the transaction (browsing, video, etc.)	browsing
subscriber	The subscriber ID based on the International Mobile Subscriber Identity (IMSI). This is a unique identifier of a subscriber responsible of the transaction	******13992
handset	The device type based on the Type Approval Code (TAC). This is a unique identifier of mobile device used during the transaction	35195507
cell	The cell ID. This is unique identifier of the cell (network element) used by a subscriber for a specific transaction	******467
events	The number of events for every aggregation	3
sec_dl	The active second on the downlink	148
retransbytes_dl	The bytes retransmission on the downlink	655594
bytes_dl	The bytes transmitted on the downlink	2451800
dns_successful	The number of success full DNS transaction	80
dns_failure	The number of failed full DNS transaction	0
latency_dl	The latency on the downlink	518

3.5.2. Model design

3.5.2.1. Quadri-dimensional approach

To implement the quadri-dimensional approach, an SQM-tree was built based on four dimensions which were the service, the subscriber, the handset and the cell with four levels representing the depth of the tree nodes as shown in Figure 23. From the top to the bottom, we had the global level which is the highest aggregation providing visibility of the performance of the whole network, the service dimension consisting of the SQI for the services: browsing, video,

facebook, peer-to-peer (p2p) and others; the other dimensions SQI level: the subscriber, the handset and cell; and the KPI level: the round-trip time on the downlink (rtt_dl), the retransmission rate on the downlink (rtx_dl), the DNS success rate (dns_sr) and the throughput on the downlink (thp_dl).

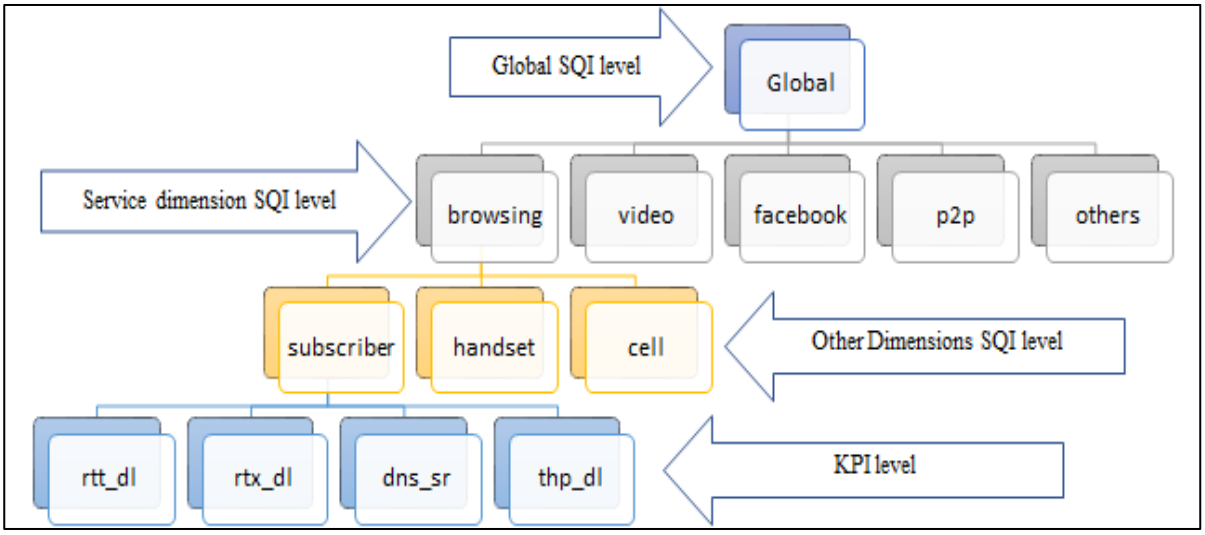


Figure 23: SQM Tree-based approach

To make sure we have meaningful data transactions, we used a flag for transactions with bytes transmitted on the downlink above 1,5 MBytes and a connection time above 50 seconds and only records with a set flag were considered. The calculated KPIs are as follows:

$$thp\_dl = \begin{cases} \dfrac{8 * (\sum bytes\_dl)}{1024 * (\sum sec\_dl)}, & flag > 0 \\ N/A, & flag \leq 0 \end{cases} \quad (14)$$

$$rtx\_dl = \frac{100 * (\sum retransbytes\_dl)}{\sum bytes\_dl} \quad (15)$$

$$dns\_sr = \frac{100 * (\sum dns\_successful)}{\sum (dns\_successful + dns\_failure)} \quad (16)$$

$$rtt\_dl = \frac{\sum latency\_dl}{\sum events} \quad (17)$$

3.5.2.2. SQM-Tree construction

To build the SQM-Tree, we used two algorithms, the first one to normalize the KPI level and the second one to build and fill the SQM tree following a quadri-dimensional approach.

Since the KPIs such as the throughput (thp_dl) and the round-trip time (rtt_dl) are numbers that can range from 0 to several thousands, the first algorithm as shown in Algorithm 1 was used to normalize the KPI level by receiving the original KPI value Kk and returning a normalized value Kk' ranging between 0 and 100. Since the data used were based on a 3G packet-switched network, we considered a thp_dl value less than 500Kbps as worst and normalized at 0, a value ranging from 500Kbps to 1Mbps normalized between 0 and 100 while any value above 1Mbps is considered as the best case and normalized at 100. For the rtt_dl, we considered a value less than 500ms as best and normalized at 100, a value between 500ms and 1000ms normalized between 100 and 0 while any value above 1000ms is considered as the worst case and normalized at 0. For the rtx_dl, as it is in percentage with 0% as the best value, the normalized value was considered as the complementary value to 100. Finally, the dns_sr remained the same because it is already a percentage and 100% is the best value.

The second algorithm as shown in algorithm 2 was the algorithm used to construct the SQM-tree nodes based on a quadri-dimensional approach focusing on the four dimensions: service, subscriber, handset and cell and to dynamically design Big Data queries to fill in the tree with both the KPIs and the aggregated SQIs data. To provide a quality indicator that does not only consider the aggregated values of the KPIs but also the impact of the performance on each dimension. All the nodes in the tree had three types of information which were the value, the impact and the quality. The value was the weighted aggregation of different KPIs, the impact was the percentage of a dimension (subscriber, handset, etc) with better service performance (normalized KPIs >50) and the quality was the weighted aggregation of both the value and the impact. The Algorithm 2 received three sets of data and returned the built and filled SQM-tree following the quadri-dimensional approach. The sets of data received were the service set S defined as S = { "browsing", "video", "facebook", "p2p", "others"} where Si represented each

service with i ∈ {1, 2, 3, 4, 5}, the dimension set D defined as D = {"service", "subscriber", "handset", "cell"} where Dj represented each dimension with j ∈ {1, 2, 3, 4} and the KPI set K defined as K = { "rtt_dl", "rtx_dl", "dns_sr", "thp_dl"} where Kk represented each KPI with k ∈ {1,2,3,4}.

Algorithm 1 KPI level normalization

```
Input: Kk
Output: Kk'
if Kk == 'thp_dl' then
    if Kk>1024 then
        Kk'=100;
    else if Kk>500 then
        Kk'=100*(Kk-500)/500;
    else
        Kk'=0;
    end if
else if Kk == 'rtt_dl' then
    if Kk<500 then
        Kk' = 100;
    else if Kk<1000 then
        Kk' = 100*(1000-Kk)/Kk
    else
        Kk' = 0;
else if Kk == 'rtx_dl' then
    Kk' = 100-Kk;
else
    Kk' = Kk
end if
```

Algorithm2 SQM-tree Construction

```
1:  Input: D, S, K
2:  Output: SQM-tree
3:  initialise global = node. New("global");
4:  for each Si in S do
5:     serv = global.AddChild(Si);
6:     for each Dj in D do
7:        if Dj != "service" then
8:           dim = serv.AddChild(Dj);
9:           for each Kk in K do
10:             kpi = dim.AddChild(Kk);
11:             query1 = “SELECT Algorithm1(Kk) FROM SQM WHERE service = Si”;
12:             query2 = “SELECT COUNT(DISTINCT(Dj)) FROM  SQM WHERE service = Si AND Algorithm1(Kk)>50”;
13:             query3 = “SELECT COUNT(DISTINCT(Dj)) FROM  SQM WHERE service = Si”;
14:             global.Si.Dj.Kk.value= dbGetQuery(impala,query1);
15:             global.Si.Dj.Kk.impact= 100*dbGetQuery(impala,query2)/dbGetQuery(impala,query3);
16:             global.Si.Dj.Kk.quality = 0.5*global.Si.Dj.Kk.value + 0.5*global.Si.Dj.Kk.impact;
17:          end for
18:       global.Si.Dj.value = 0.25*global.Si.Dj.K1.value + 0.25*global.Si.Dj.K2.value +
                                0.25*global.Si.Dj.K3.value+ 0.25*global.Si.Dj.K4.value;
19:       global.Si.Dj.impact = 0.25*global.Si.Dj.K1.impact + 0.25*global.Si.Dj.K2.impact +
                                  0.25*global.Si.Dj.K3.impact + 0.25*global.Si.Dj.K4.impact;
20:       global.Si.Dj.quality = 0.5*global.Si.Dj.value + 0.5*global.Si.Dj.impact;
21:       end if
22:    global.Si.value = 0.5*global.Si.D2.value + 0.25*global.Si.D3.value+ 0.25*global.Si.D4.value;
23:    global.Si.impact = 0.5*global.Si.D2.impact + 0.25*global.Si.D3.impact + 0.25*global.Si.D4.impact;
24:    global.Si.quality = 0.5*global.Si.value + 0.5*global.Si.impact;
25:    end for
26: end for
27: global.value = 0.2*global.S1.value + 0.2*global.S2.value + 0.2*global.S3.value +
                     0.2*global.S4.value + 0.2*global.S5.value;
28: global.impact = 0.2*global.S1.impact + 0.2*global.S2.impact +
                      0.2*global.S3.impact + 0.2*global.S4.impact + 0.2*global.S5.impact;
29: global.quality = 0.5*global.value + 0.5*global.impact;
```

3.6. User-based QoS categorization from radio conditions

As the new data greedy services increase the complexity of service management and the level of customers' expectations on the QoS delivered, this section introduced machine learning techniques to predict user-based QoS category using radio conditions parameters. A comparative analysis of the accuracy performance of different data mining techniques was proposed.

3.6.1. Data preparation

The data collected from the Iub interface was aggregated per IMSI which is the unique identifier of a subscriber so as to reduce the number of records. To clean the data, transactions without the important fields such as IMSI, EcNo, RSCP and propagation delay were discarded and only subscribers with at least 10 packet-switched calls a day were selected to ensure the representativeness of the collected data.

The final dataset had a total number of 28,468 observations with each record representing an IMSI aggregated transaction for a day with different new computed metrics based on the three radio conditions information from the measurement reports. The new metrics were classes of radio conditions where a combination of EcNo and RSCP, propagation delay and the average of each radio condition were considered. The thresholds used for the radio conditions were:

1. Threshold EcNo = -15dB, any value below this is considered as poor radio signal quality.
2. Threshold RSCP = -100dBm, any value below this is considered as poor radio signal strength.
3. Threshold propagation delay = 6km, any value above this was considered as an overshooting distance for cells in an urban area as all the data were collected from urban areas.

The other attributes considered were the total number of packet-switched calls and the user-based QoS category. The user-based QoS category was derived from the packet-switched dropped call rate. The full list of attributes with their description is shown in Table 8.

Table 8: Packet-switched radio conditions QoS attributes details

Attributes name	Description	Example
X1	IMSI	******13992
X2	Percentage of transactions with good EcNo and good RSCP	0
X3	Percentage of transactions with bad EcNo and good RSCP	50
X4	Percentage of transactions with good EcNo and bad RSCP	0
X5	Percentage of transactions with bad EcNo and bad RSCP	50
X6	Percentage of transactions with long Propagation delay	0
X7	Average EcNo	-9 dB
X8	Average RSCP	-85 dBm
X9	Average propagation delay	1 Km
X10	Number of packet-switched calls	12
X11	User-based packet-switched QoS category	BAD

3.6.2. Model design

The system model is a user-centric one and focuses on the users' perception of the QoS from categorization between "BAD" users with a daily packet-switched call dropped rate higher than 20% and "GOOD" users with a daily packet-switched call dropped rate less than 20%. A common subscriber considers 2 packet-switched call drops out of ten to be acceptable while over that threshold, the feedback on the experience becomes unacceptable. To derive the categorization from the measurements of the radio conditions, three parameters were considered from the

measurement reports: the EcNo to provide information about the radio signal quality, the RSCP to provide the radio signal strength and the propagation delay for the indication of the distance at which the transaction was initiated respective to the network cell.

After comparing five different machine learning algorithms for training, the best model was selected based on the accuracy. The machine learning algorithms that were used for comparisons were the LDA, the KNN, the CART, the RF and the SVM.

3.6.2.1. Training and testing approach

The dataset was gathered from an urban 3G packet-switched network with a total of 28,468 observations that was divided into two subsets with a ratio of 80% for the training dataset (22,775 observations) and 20% for the testing dataset (5,693 observations). Although the original dataset was balanced, to avoid risk of over-fitting, during the training phase, the cross-validation method was applied to randomly divide the training dataset into further ten subsets. All the machine learning techniques used ten rounds and one subset was selected for validation (Testing fold) and the aggregation of the rest used as training set. The results of the ten rounds were then averaged to get the final result.

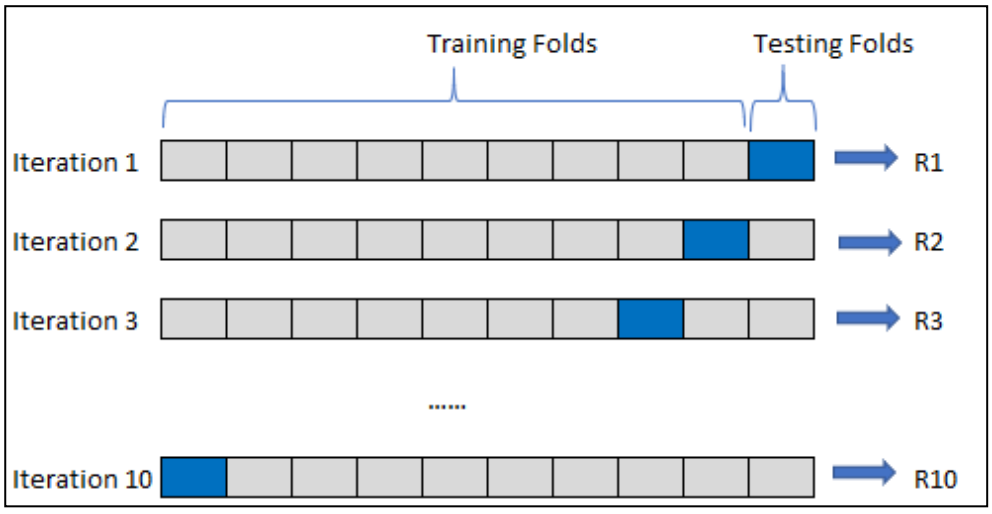


Figure 24: 10 folds cross-validation

Figure 24 shows the 10-fold cross-validation where each iteration provided a result R*i* with *i* as the number of iteration or number of folds.

$$\mathrm{R} = \frac{1}{10} \sum_{i=1}^{10} Ri \tag{18}$$

To ensure normal distribution of each of the numerical attribute, the data was transformed using the z-score:

$$\mathrm{Zscore} = \frac{X - \mu}{\sigma} \tag{19}$$

Where *X* is the value of the attribute to be transformed, μ as the mean and σ as the standard deviation:

$$\mu = \frac{\sum_{i=1}^{n} Xi}{\mathrm{n}} \tag{20}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (Xi - \mu)^2}{n}} \tag{21}$$

where *n* as the number of observations in the dataset. 5 different models which are LDA, KNN, CART, RF and SVM will be trained involving 10-fold cross-validation and hyperparameters tuning to select the best model during training based on the accuracy will be used for testing.

3.7. Poor data throughput root cause analysis (RCA)

This section proposed a deep learning approach based on a DNN architecture to train and evaluate a model for a poor throughput root cause analysis using both the radio and the core network performance indicators of a 3G packet-switched network as inputs. The approach is a user-centric one and is based on user perception of QoS from categorization between “BAD” for 3G data throughput less than 500Kbps and “GOOD” throughput higher than 500Kbps. To ensure end-to-end visibility, both the performance from the radio and the core network have been correlated based on subscribers as common keys. For the radio part, the radio conditions were derived from the measurement reports considering essentially two parameters: The energy per chip and noise (EcNo) to provide information about the quality of the radio signal and the received signal code power (RSCP) to provide the radio signal strength. For the core network part, seven

parameters were considered: the total time of data connection (act_dl), the total bytes transmitted on the downlink (bytes_dl), the total bytes retransmitted on the downlink (retransbytes_dl), the latency from the core to the user equipment (latency_dl), the total number of events (event), the number of successful DNS transactions (dns_successful) and number of unsuccessful DNS transactions (dns_failure).

3.7.1. Data preparation

The data was hourly aggregated and cleaned from unnecessary information. Each record represented an IMSI aggregated transaction for an hour with different new computed metrics based on the two radio conditions information from the measurement report and the seven core network parameters. The seven core network parameters reused the equations (14), (15), (16) and (17) to compute the four core network metrics which are the round-trip time on the downlink (rtt_dl), the retransmission rate on the downlink (rtx_dl), the DNS success rate (dns_sr) and the throughput on the downlink (thp_dl). Instead, the two radio parameters based on equations (22) and (23) computed new metrics with good EcNo (ecno_good), the sum of transactions with bad EcNo (ecno_bad), the sum of transactions with critical EcNo (ecno_critical), the sum of transactions with good RSCP (rscp_good), the sum of transactions with bad RSCP (rscp_bad), the sum of transactions with critical RSCP (rscp_critical).

$$ecno = \begin{cases} good, & \text{Threshold EcNo} > -10dB \\ bad, & -15dB < \text{Threshold EcNo} < -10dB \\ critical, & \text{Threshold EcNo} \leq -15dB \end{cases} \quad (22)$$

$$rscp = \begin{cases} good, & \text{Threshold RSCP} > -90dBm \\ bad, & -100dB < \text{Threshold RSCP} < -90dBm \\ critical, & \text{Threshold RSCP} \leq -100dBm \end{cases} \quad (23)$$

A total number of 44,711 observations that were later split for the training, validation and testing of the model considering the average thp_dl higher than 500Kbps as “GOOD” and less than 500Kbps as “BAD”. Table 9 provides the full list of attributes with their description.

Table 9: Poor data throughput attributes description

Attributes name	Description	Example
X1	IMSI	**********13992
X2	thp_dl	BAD
X3	rtx_dl	21%
X4	dns_sr	96%
X5	rtt_dl	111msec
X6	ecno_good	9
X7	ecno_bad	14
X8	ecno_critical	2
X9	rscp_good	7
X10	rscp_bad	13
X11	rscp_critical	5

3.7.2. Model design

3.7.2.1. Training and testing approach

The dataset has been divided into two new datasets with a ratio of 70% for the training and validation dataset (31,298 Observations) and 30% for the testing dataset (13,413 Observations). nine attributes were used as predictors: rtx_dl, dns_sr, rtt_dl, ecno_good, ecno_bad, ecno_critical, rscp_good, rscp_bad and rscp_critical. The response variable for training and testing was the thp_dl that was one-hot encoded to provide two outputs (first output for “BAD” throughput response and the second output for the “GOOD” throughput response) each with

binary possibility of 0 or 1 (0 for “No” and 1 for “Yes”). It must be noted that having two outputs instead of one will not improve the results in case of binary classification, but we have chosen this approach to enable dynamic configuration using one-hot encoding to support multiclass cases for the future. The normalization transformation based on equation (8) was applied to all the attributes except the targets outputs which were already binary 0 or 1 as result of one-hot encoding.

The DNN training has an objective of finding the neural network parameters that would minimize the loss or cost function while improving the performance. The proposed model was a sequential model as a linear stack of layers as shown in Figure 25.

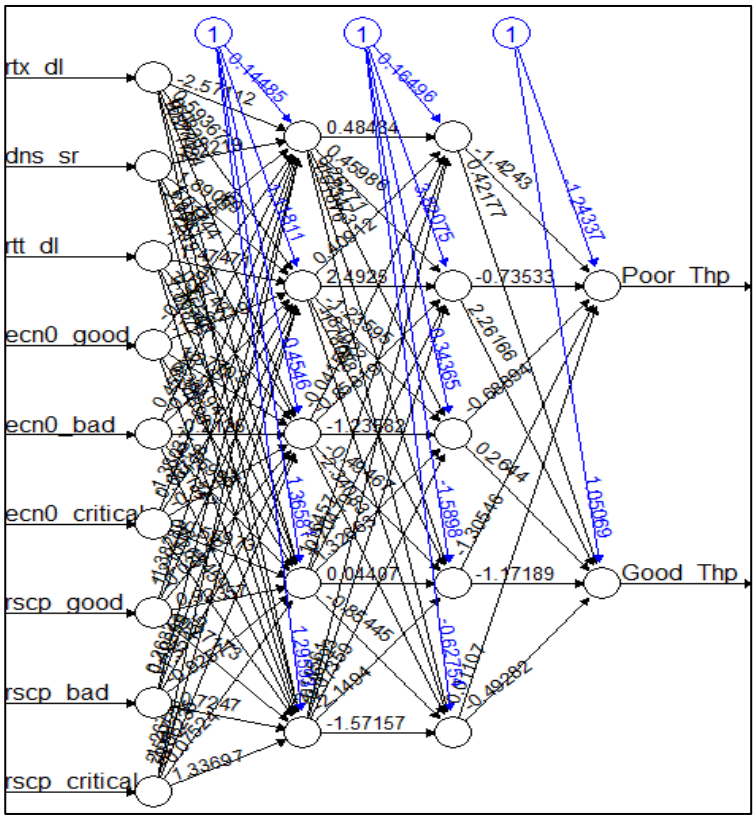


Figure 25: Proposed DNN architecture

The activation function used for the hidden layers was the rectified linear unit (ReLu) which is one of the common modern non-linear activation functions as shown in Figure 26 in comparison to traditional non-linear activation functions such as the Hyperbolic Tangent.

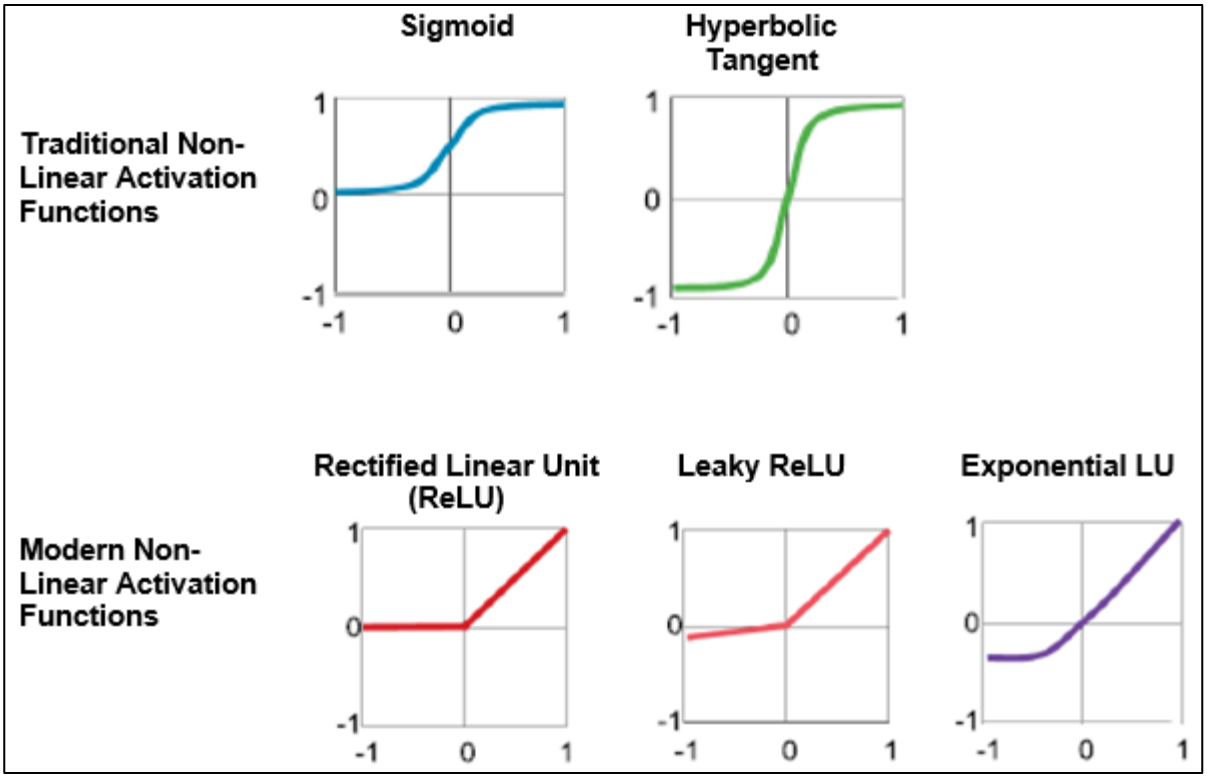


Figure 26: Non-linear activation functions

The output layer instead used the Softmax activation function which output the probability of an instance to belong to a specific class or category. The configuration summary of the model built is described in Table 10.

Table 10: Proposed deep neural network architecture

Layer type	**Number of units**	**Activation function**
Input layer	9	-
Hidden layer 1	5	ReLu
Hidden layer 2	5	ReLu
Output layer	2	Softmax

One of the difficult tasks in deep learning is the optimization of the learning rate, since small learning rate causes many iterations until convergence and trapping in local minima while large learning rate causes overshooting. Different methods are used to guess the learning rate when using the stochastic gradient descent (SGD), but most of them are time consuming. We used the approach proposed by Kingma and Ba [96] with the adaptive learning rate optimizer such as the “Adam” providing better performance as shown in Figure 27 [96].

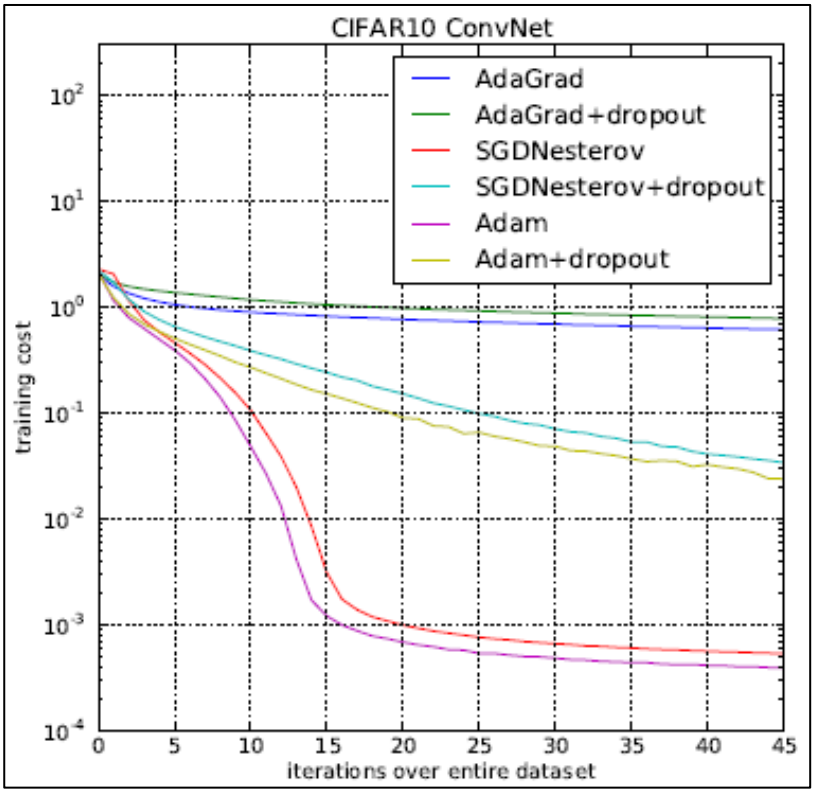


Figure 27: Convolutional neural networks training cost comparison [96]

With adaptive learning rate, the learning rate is no longer fixed and can be made larger or smaller depending on the size of the gradient, how fast learning is happening and the size of weights. We used the minibatch technique to fit model since it provides much accurate estimation of the gradient with smoother convergence allowing larger learning rates and therefore faster training. The parameters used to compile and fit the model are shown Table 11.

Table 11: Proposed deep neural network model parameters

Model parameters	**Value**
optimizer	Adam
loss function	Cross-entropy
Optimization metric	Accuracy
batch size	32
epochs	50
validation split	20%

The accuracy and the loss of both training and validation phase were used to tune the parameters while the F1_Score and the AUC were used to evaluate the final model.

3.7.2.2. Root cause analysis (RCA) approach

Very often the blackbox data mining techniques such as DNN are not used in the environment of root cause analysis because of lack of explanation from the results, but this study proposed a root cause analysis using the feature importance during the prediction phase to derive the predictors that impacted the most the final output of the throughput. The user-based root cause analysis required a different approach since the poor QoE of different users might have different reasons. To be able to understand single poor throughput reason, we used a R library based on the local interpretable model-agnostic explanations (LIME) technique which provides explanation of complex machine learning classifiers.

While the model could retrieve the features importance for any subscribers in the testing dataset, for demonstration, only the first four subscribers have been used to outline the influence of predictors focusing on the output response “BAD” for poor data throughput. For this, the interest was only on the top four features based on importance using a kernel width of 0.7.

Chapter 4: Experimental Results and Discussions

4.1. Introduction

This chapter presents an analysis of results of the various models that we developed. The results are presented in four section: Facebook video traffic classification, a quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data, user-based QoS categorization from radio conditions using machine learning techniques and poor data throughput RCA. The specification of the hardware used for the experiments and their configuration parameters are shown in Table 12.

Table 12: Hardware configuration

Parameter	**Value**
Computer1 Processor	Intel Core i7 (4 Processors)
Computer1 Memory	16 GB (8GB used for the VM)
Computer1 Storage	1 Terabytes HDD
Computer1 IP	198.168.68.1
VM Processor	Hosted on VMware in Computer1 (4 Virtual Processors)
VM Memory	8 GB
VM Storage	64 GB Dedicated Virtual HDD
VM IP	198.168.68.120

.

4.2. Facebook video traffic classification

4.2.1. Model implementation

The experiments were carried out on RStudio using the R programming language with the "neuralnet" library. During the experiment for Facebook video traffic classification, four topologies were built based on the variation of the number of hidden neurons. The topologies were trained with the training dataset and validated with the unseen data from the testing dataset. For each case, a topology was generated with the prediction confusion matrix to enable computation of the accuracy which was the metric used to benchmark the models.

Topology 1

Figure 28 shows the number of hidden neurons configured for the first topology with H_N = 4. Table 13 shows the results of the prediction confusion matrix that was used to compute the accuracy of this model.

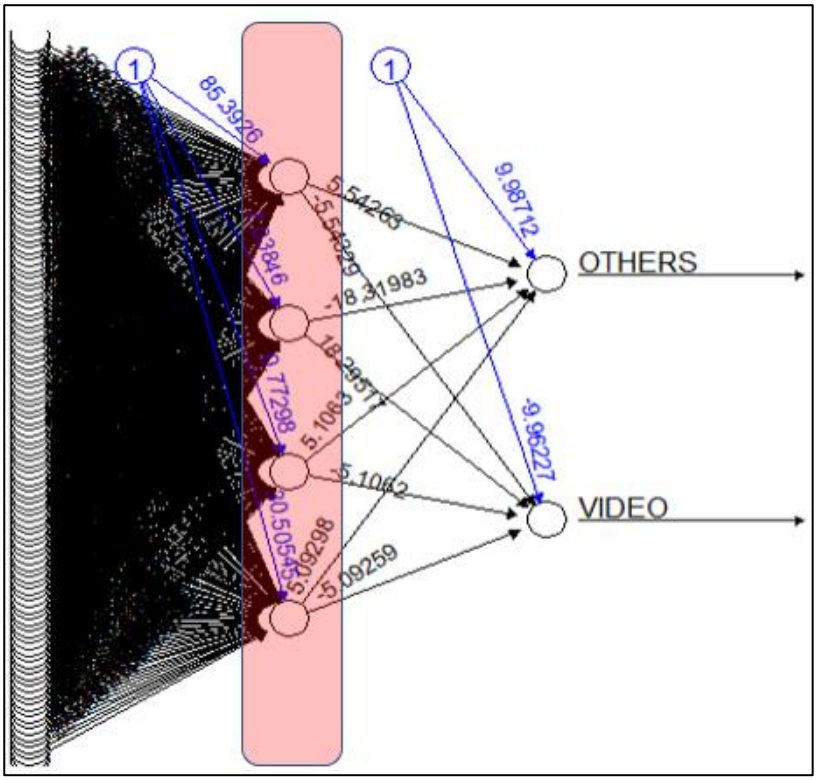


Figure 28: Topology 1 (4 Hidden Neurons)

Table 13: Topology 1 confusion matrix

	Reference	
Prediction	OTHERS	VIDEO
OTHERS	571	38
VIDEO	15	218

Topology 2

Figure 29 shows the number of hidden neurons configured for the second topology with $H_N = 5$. Table 14 shows the results of the prediction confusion matrix that was used to compute the accuracy of this model.

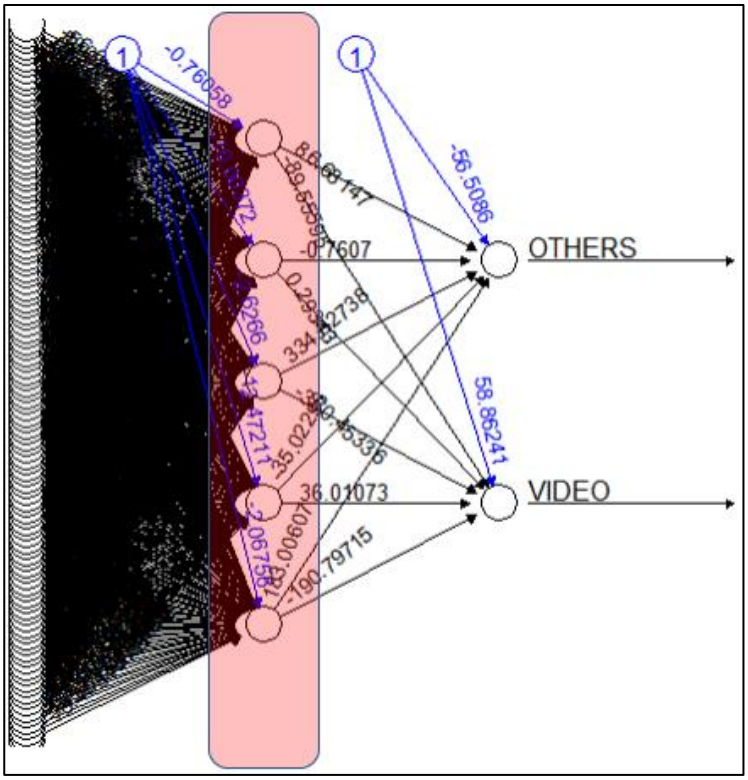


Figure 29: Topology 2 (5 Hidden Neurons)

Table 14: Topology 2 confusion matrix

	Reference	
Prediction	OTHERS	VIDEO
OTHERS	547	33
VIDEO	39	223

Topology 3

Figure 30 shows the number of hidden neurons configured for the third topology with H_N = 15. Table 15 shows the results of the prediction confusion matrix that was used to compute the accuracy of this model.

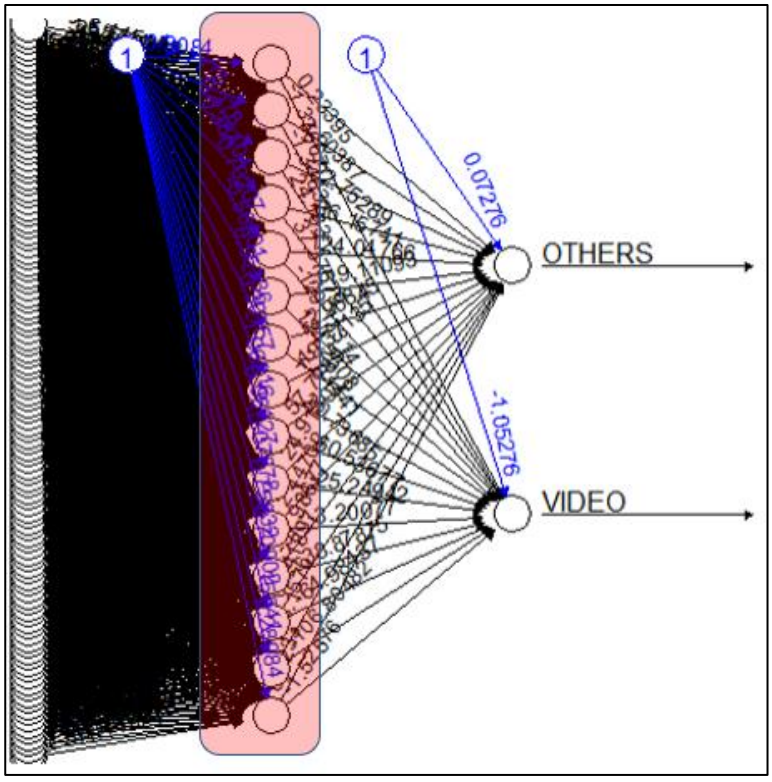


Figure 30: Topology 3 (15 Hidden Neurons)

Table 15: Topology 3 confusion matrix

	Reference	
Prediction	OTHERS	VIDEO
OTHERS	545	64
VIDEO	41	192

Topology 4

Figure 31 shows the number of hidden neurons configured for the fourth topology with H_N = 81. Table 16 shows the results of the prediction confusion matrix that was used to compute the accuracy of this model.

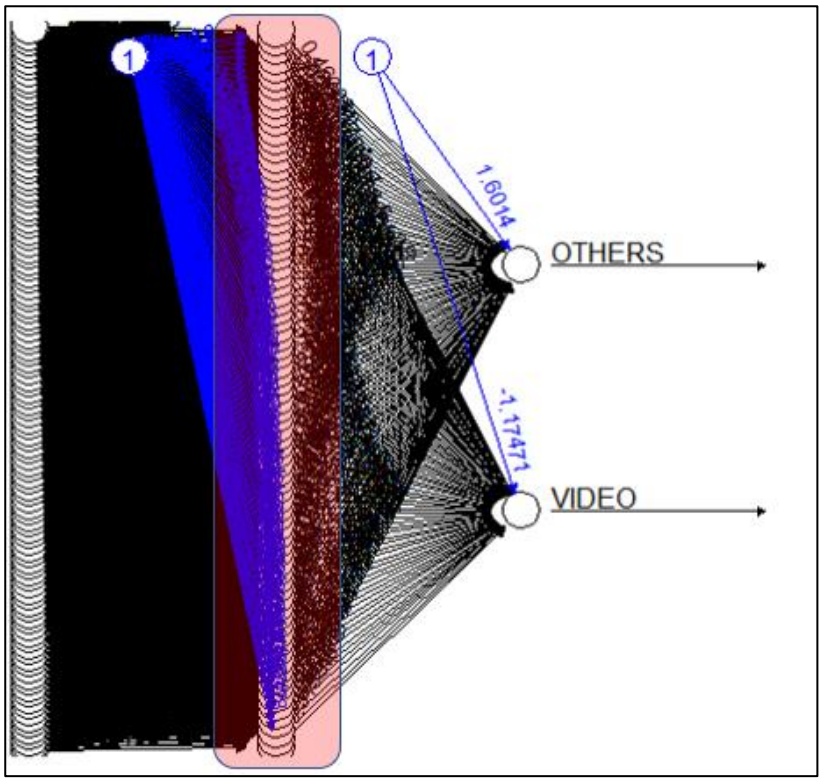


Figure 31: Topology 4 (81 Hidden Neurons)

Table 16: Topology 4 confusion matrix

	Reference	
Prediction	OTHERS	VIDEO
OTHERS	545	64
VIDEO	41	192

4.2.2. Prediction performance benchmark

The metric used to benchmark the models during the experiments was the accuracy. Although the results were quite closer to each other for all the built topologies, the best topology

considering the test prediction (accuracy), was topology 1 with 93.7% of prediction accuracy on unseen data as shown in Figure 32.

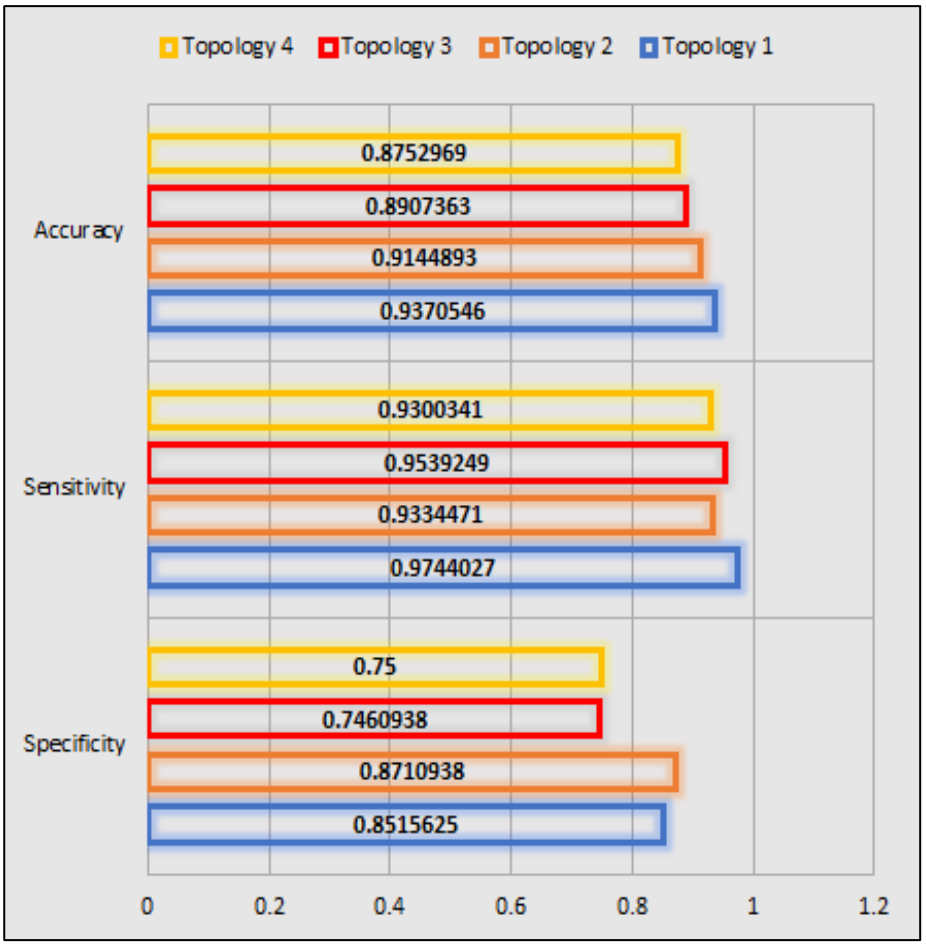


Figure 32: benchmark of prediction performances

The initial assumption was to increase the number of hidden nodes to improve the learning capability of the model, as more iterations provides the optimal weights of the neural networks. However, this experiment showed that after a certain threshold, in this case four hidden neurons, the model started to overfit by memorizing the training data rather than learning from the relationship between them. The prediction performance started to deteriorate rather than improving.

During training of ANN models, it is important to test different scenarios starting from a relatively small number of hidden nodes and gradually increase while benchmarking the accuracy of the models.

4.3. A quadri-dimensional approach for poor performance prioritization in mobile networks using Big Data

4.3.1. Model implementation

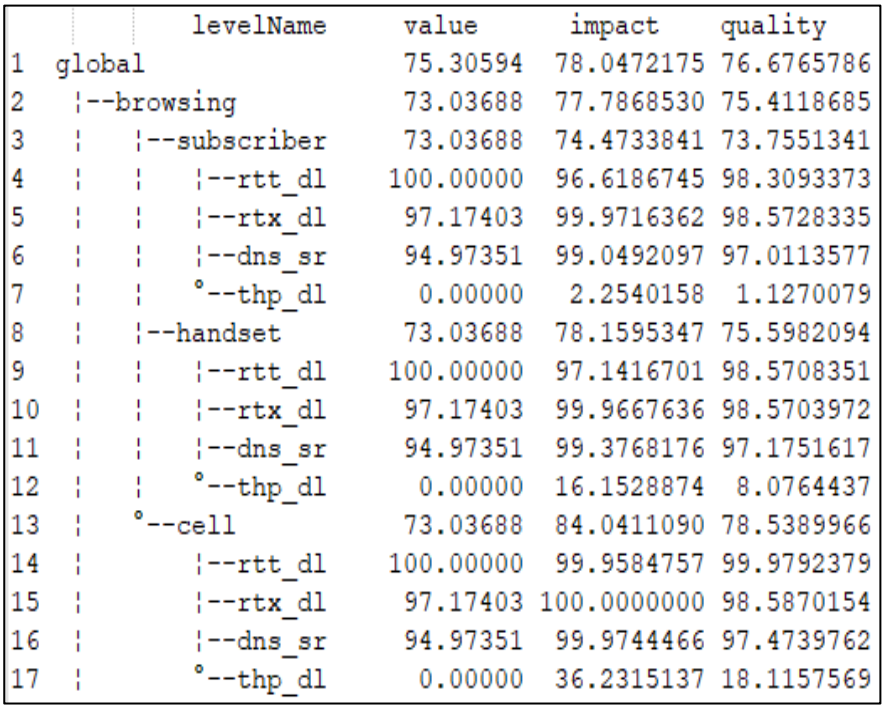
During the experiment for quadri-dimensional approach was conducted running Rstudio on the hosting machine and the Cloudera platform on the virtual machine. To demonstrate the benefits of the Big Data platform we run a parallel experience using MySQL to compare the results. Table 17 shows the average performance comparison between MySQL and Cloudera Impala using the queries from Algorithm 2. The results show that using a Big Data platform, even on a single machine has a performance three times better than the traditional MySQL.

Table 17: MySQL and Big Data performance comparison

Test	**MySQL execution time**	**Cloudera Impala execution time**
Query1	2.2 sec	0.7 sec
Query2	6.1 sec	2.3 sec

4.3.2. SQM-tree results

From the SQM-tree output of algorithm 2, we had the value, the impact and the quality, all ranging from 0 to 100 for all the nodes in the tree represented by the “levelName” as shown in Figure 33.



	levelName	value	impact	quality
1	global	75.30594	78.0472175	76.6765786
2	¦--browsing	73.03688	77.7868530	75.4118685
3	¦ ¦--subscriber	73.03688	74.4733841	73.7551341
4	¦ ¦ ¦--rtt_dl	100.00000	96.6186745	98.3093373
5	¦ ¦ ¦--rtx_dl	97.17403	99.9716362	98.5728335
6	¦ ¦ ¦--dns_sr	94.97351	99.0492097	97.0113577
7	¦ ¦ °--thp_dl	0.00000	2.2540158	1.1270079
8	¦ ¦--handset	73.03688	78.1595347	75.5982094
9	¦ ¦ ¦--rtt_dl	100.00000	97.1416701	98.5708351
10	¦ ¦ ¦--rtx_dl	97.17403	99.9667636	98.5703972
11	¦ ¦ ¦--dns_sr	94.97351	99.3768176	97.1751617
12	¦ ¦ °--thp_dl	0.00000	16.1528874	8.0764437
13	¦ °--cell	73.03688	84.0411090	78.5389966
14	¦ ¦--rtt_dl	100.00000	99.9584757	99.9792379
15	¦ ¦--rtx_dl	97.17403	100.0000000	98.5870154
16	¦ ¦--dns_sr	94.97351	99.9744466	97.4739762
17	¦ °--thp_dl	0.00000	36.2315137	18.1157569

Figure 33: SQM-tree result screenshot

The benefit of the quadri-dimensional approach being the fast troubleshooting and root cause analysis capability, it is therefore possible from the SQM-tree, to sort the specific “levelName” by quality and identify the worst paths instead of running multiple Big Data queries and KPIs analysis as it is done in most of the NOC/SOC. Since the SQM-tree paths were build following the quadri-dimensional approach, the worst paths provide also information about which service, dimension and KPIs have the most impacted the network quality, the QoS and the QoE.

4.3.3. Worst SQM-Tree paths

Figure 34 shows a screenshot of the ten worst paths ranked by the performance quality. From the list of the worst paths and based on the KPIs, the worst path is linked to the throughput on the downlink. This in essence, affects several network dimensions and services. Of this most impacted is the Facebook network service. An investigation and troubleshooting can then be performed by prioritizing the paths with poor performance so as to reduce the mean time to detection and the MTTR. This will improve the efficiency of the CSPs operation team.

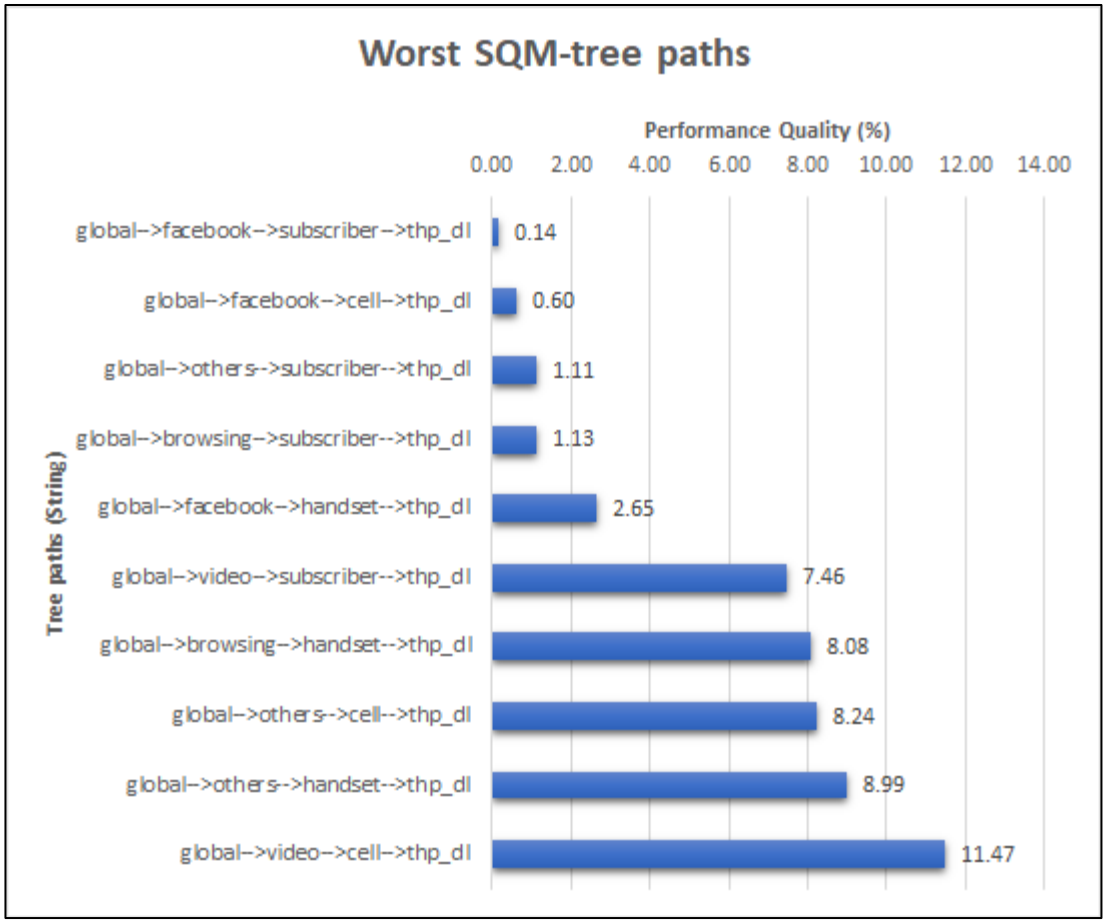


Figure 34: 10 Worst SQM-Tree paths ranked by performance quality (%)

4.4. User-based QoS categorization from radio conditions

4.4.1. Model implementation

The training and testing of the model were implemented using R programming language with the library “caret”, providing several data processing tools and machine learning models. The models trained and compared were based on the following algorithms: LDA, KNN, CART, RF and SVM. The best model based on accuracy turned out to be the RF as shown in Figure 35.

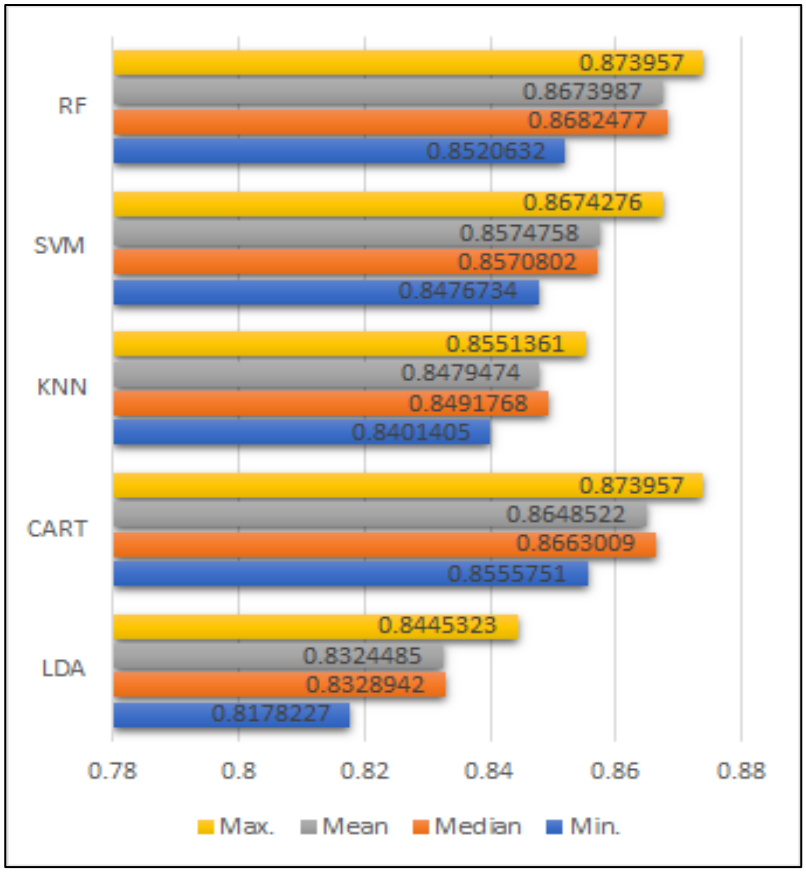


Figure 35: Models Accuracy metrics

Based on cross-validation, the resampling results focused on tuning the accuracy as the chosen evaluation metric by varying the mtry which is number of available variables to split at each tree node. The optimal result was achieved with mtry = 2 with an accuracy mean of 86.74% during training process as shown in Table 18.

Table 18: Random Forests mtry vs accuracy

mtry	Accuracy
2	0.867399
5	0.854709
9	0.847728

One of the reasons that could justify a better performance of the RF model is the fact that it can handle extremely large datasets where other models tend to underperform due to "curse of dimensionality". Models based on the RF also prone less to overfitting and can handle very noisy data while selecting only the most important features.

4.4.2. Prediction using the best model

Prediction was done using the testing dataset which contained 5,693 observations where the best model during training phase was the RF. The prediction accuracy was 86.35% with a precision of 95.7% considering the class “BAD” as the positive class as shown in the confusion matrix in Table 19.

Table 19: Random Forests prediction confusion matrix

	Reference	
Prediction	BAD	GOOD
BAD	1980	89
GOOD	688	2936

Table 20 provides details of other metrics after prediction:

Table 20: Detailed prediction metrics

Metrics	**Values**
Accuracy	0.8635
Sensitivity	0.7421
Specificity	0.9706
Prediction of "BAD" class	0.957
Prediction of "GOOD" class	0.8102

4.5. Poor data throughput root cause analysis (RCA)

4.5.1. Model implementation

The training and testing of the model was implemented in R programming language using the library “Keras”, which is a high-level neural network API for deep learning based on TensorFlow developed by Google. The training and validation of the model was done using 70% of the data with 31,298 observations from which the model trained on 25,038 samples and validated on 6,260 samples in different epochs. Figure 36 shows the summary of the performance of the training data against cross-validation over time showing the training accuracy (acc), the training loss (loss), the validation accuracy (val_acc) and the validation loss (val_loss).

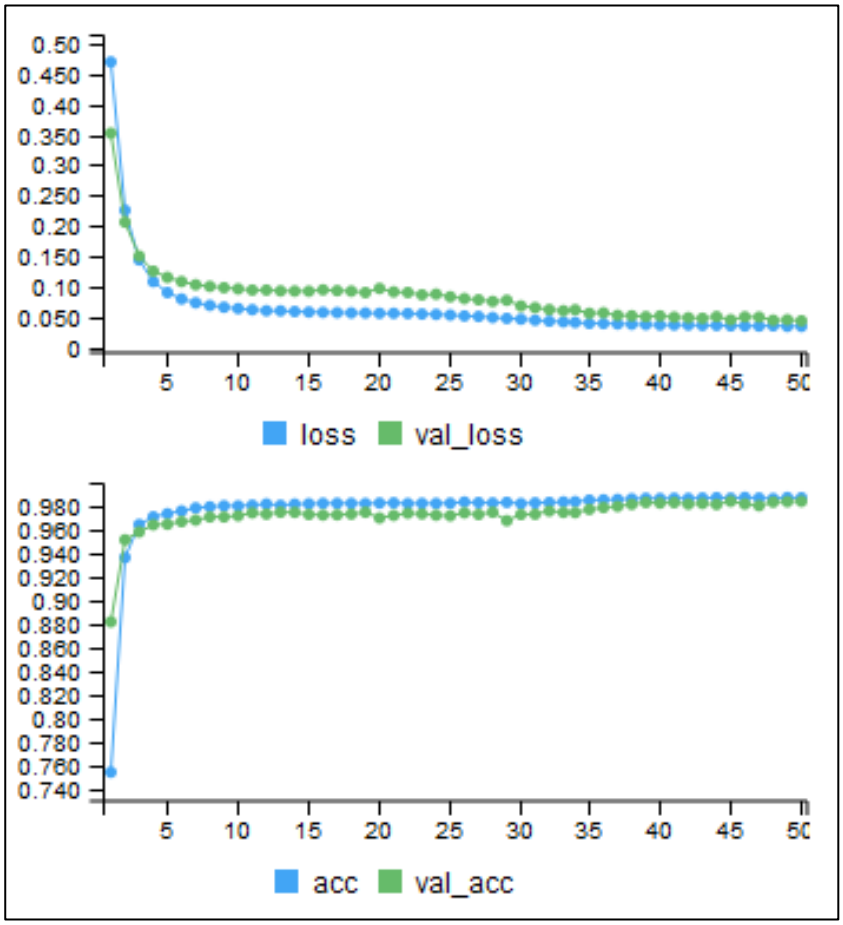


Figure 36: DNN training and validation metrics

The final performance during the training and validation phase are shown Table 21.

Table 21: DNN training evaluation metrics

Training evaluation metrics	**Value**
training loss	0.03522
validation loss	0.04641
training accuracy	0.9886
validation accuracy	0.9856

4.5.2. Prediction using the testing dataset

Prediction was done using the testing dataset which contained 13,413 unseen data. The prediction accuracy on unseen data was 98.9% with an AUC of 99.87% and an F1-Score of 99.23%. Table 22 shows the confusion matrix of the results after the prediction.

Table 22: DNN prediction confusion matrix

	Reference	
Prediction	BAD	GOOD
BAD	3753	134
GOOD	13	9513

4.5.3. Global RCA

The global RCA was done through the throughput correlation analysis showing the features with negative correlation that prevented a better throughput. Those features for the overall network view were the rtt_dl, the ecno_critical and the rscp_critical as shown in Figure 37.

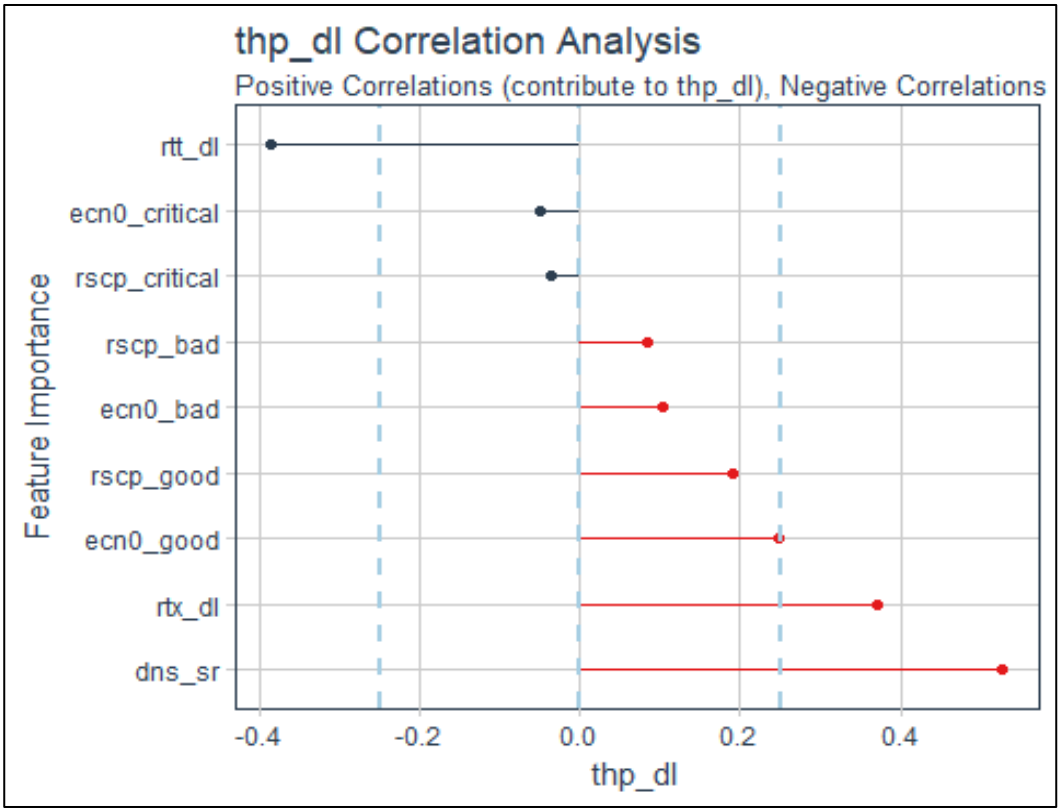


Figure 37: thp_dl correlation analysis

4.5.4. User-based RCA

Figure 38 represents detailed views of the RCA focusing on the first four subscribers that were labelled as “BAD” by the model. The green colour represents the features that contribute to the poor throughput performance as follows:

1. The 1st case shows a poor performing subscriber with 100% probability. The rtt_dl is the feature influencing the most the poor QoE, followed by the rscp_critical.
2. The 2nd case shows a poor performing subscriber with 85% probability. The rscp_bad is the feature influencing the most the poor QoE.
3. The 3rd case shows a poor performing subscriber with 99% probability. The rtt_dl is the feature influencing the most the poor QoE, followed by the ecno_critical.
4. The 4th case shows a poor performing subscriber with 100% probability. The rtt_dl is the feature influencing the most the poor QoE.

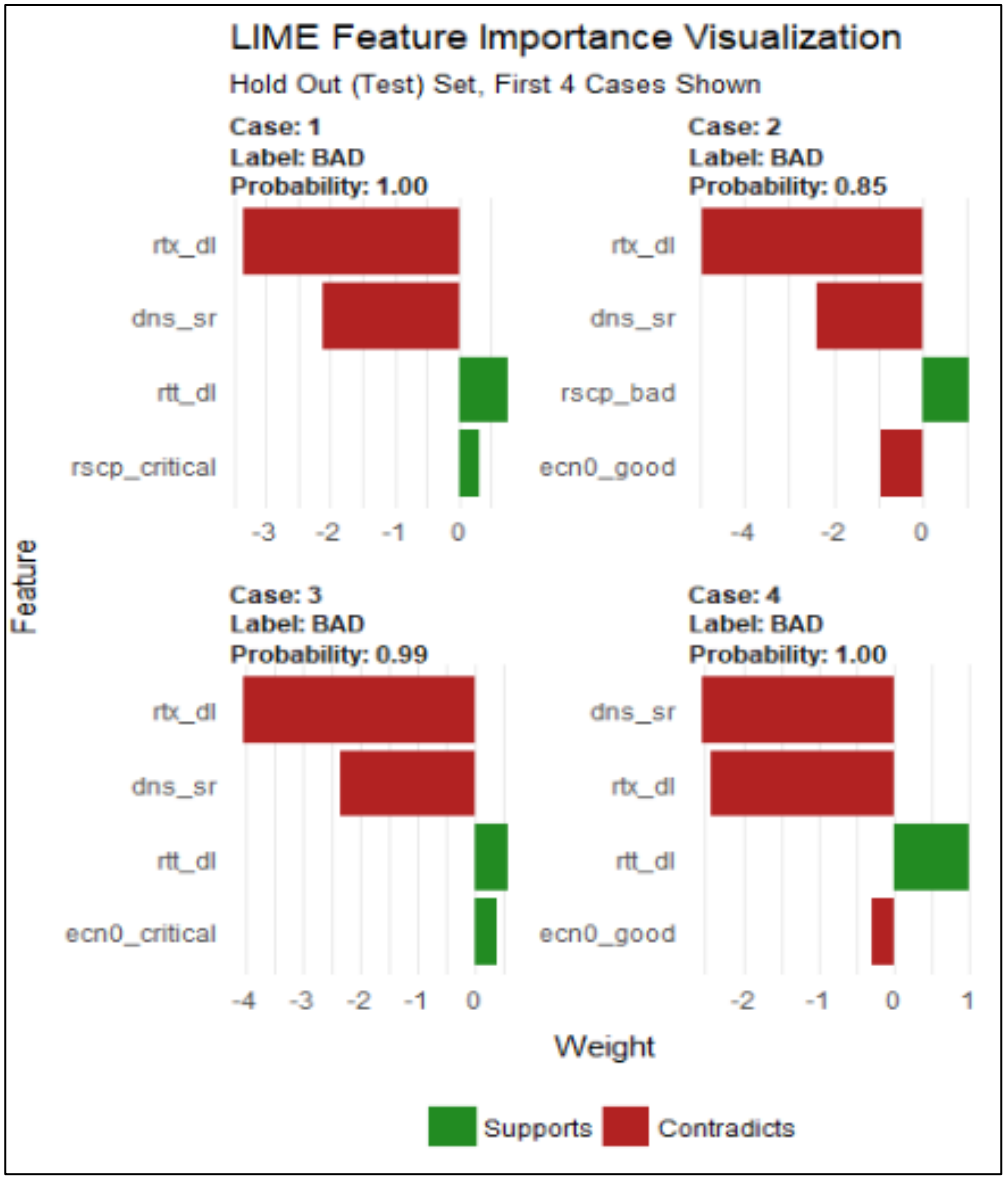


Figure 38: LIME feature importance

Although the throughput in the network is globally impacted by the rtt_dl, the ecno_critical and the rscp_critical; the reasons for poor performance could be unique for different users. This is why in a user-centric quality improvement, a global network view is not enough and a look at either a segment of customers or a view of single customer investigation is important to determine the root cause of poor QoE.

Chapter 5: Conclusion

5.1. Conclusion

In this research, we have proposed an SQM design approach considering the four dimensions in the mobile networks (service, subscriber, handset and the cell). The SQM designed followed a tree approach designed based on a KPI normalization algorithm and an SQM-tree construction algorithm dynamically preparing Big Data queries essential for the tree node weights. The tree nodes hold values not only from KPI aggregation but also considered the impact of the KPIs on the mobile network's dimensions. The final tree results could then be sorted to provide faster RCA and prioritization to manage first the issues affecting the most the network.

To solve the challenge of the Facebook video traffic misclassification, we used an ANN to build, train and test four different models for a prediction of traffic classification. As a trend, Facebook data usage is essentially made of video and social networking, we considered classification of Facebook video traffic among Facebook traffics such as chatting and browsing. While several methods in the literature have used the packet transmission patterns, the well-known ports and the IP addresses to classify the traffic, we used a new approach based on the subscriber's bytes usage of every second for both uplink and downlink traffic. Different ANN topologies were built by varying the number of hidden neurons using arbitrary functions suggested by previous literature and the best model was selected based on the prediction accuracy. This research not only demonstrated the robustness of ANN to predict large sets of data based on different attributes but also the fact that a higher number of hidden neurons does not necessary implies better performance. Therefore, several tests and comparison studies should be done while training neural networks to achieve an optimal result.

This research also proposed an approach to show the relationship between the radio conditions and the user QoS perception on real 3G packet-switched network from an urban area. We have trained and compared five different models from which the best model, based on accuracy, was selected. The approach demonstrated the relationship between the user-based QoS on packet-switched based on drop and the radio conditions while predicting an 86.35% of accuracy on the QoS perceived by the users. The perception was based on two categories: i.e.: "BAD" and "GOOD"; and took into consideration attributes such as the EcNo, the RSCP and the propagation delay collected from events in the measurement reports.

Finally, the last step of the research was to use a DNN architecture to train a model predicting poor QoE based on the poor throughput. The final proposed model not only predicted with higher accuracy but was also able to extract the features importance to derive the factors from either the radio or the core network of a 3G packet-switched network that influence the poor QoE.

5.2. Recommendation and future works

For future research, the following topics need to be subjected for further research:

1. Training and comparison of different deep learning methods to improve the performance of the models.
2. Training and comparison of models based on GPUs methods and performance evaluation.
3. Extension of the Big Data platform with either clusters of VMs or multiple physical nodes.
4. Usage of latest Real-time analytics technologies such as Kafka for data collection and processing.
5. Consideration of billing information and correlation with the QoS information to enhance the evaluation of the QoE.
6. Consideration of social media and surveys information through NLP and correlation with the QoS information to enhance the evaluation of the QoE.

References

[1] D. Istrefi, B. Cico, "Mobile payment through integrated NFC module on smartphones: New shopping experience with the use of software agents", Mediterranean Conference on Embedded Computing, pp. 66-69, June 2012.

[2] V.P. Kafle, Y. Fukushima, H. Harai, "ID-based communication for realizing IoT and M2M in future heterogeneous mobile networks", 2015 International Conference on Recent Advances in Internet of Things (RIoT) 7-9 April 2015, pp. 1-6, May 2015.

[3] Z. Kljaić, P. Škorput, N. Amin, "The challenge of cellular cooperative ITS services based on 5G communications technology", 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) May 30, 2016-June 3 2016, pp. 587-594, July 2016.

[4] N. Banović-Ćurguz, D. Ilišević, "Moving from network-centric toward customer-centric CSPs in bosnia and herzegovina", 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 696-701, May 30, 2016-June 3 2016.

[5] C. Zhang, Z. Wen-an, C. Jian, G. Hai-sheng, "A Method to Design CEM Metrics for Telecommunication Services Based on Customer Touchpoints Analysis", International Conference on Computer Science and Service System (CSSS), pp. 942-946, August 2012.

[6] C. Jian, L. Wen-wang, Z. Wen-an, G. Hai-sheng, C. Zhang, M. Shao-fu, "Customer Experience Oriented Service Quality Management", IEEE Symposium on Robotics and Applications(ISRA), pp. 298-301, June 2012.

[7] J.F. Monserrat, I Alepuz, J. Cabrejas, V. Osa, J. López, R. García, M.J. Domenech, V. Soler, "Towards user-centric operation in 5G networks", EURASIP Journal on Wireless Communications and Networking (2016) 2016:6, pp. 1-7.

[8] K.M. Shabana, J. Wilson, "A novel method for automatic discovery, annotation and interactive visualization of prominent clusters in mobile subscriber datasets", 2015 IEEE

9th International Conference on Research Challenges in Information Science (RCIS) 13-15 May 2015, pp. 127-132, June 2015.

[9] K. Peffers, T. Tuunanen, M.A. Rothenberger, S. Chatterjee, "A Design Science Research Methodology for Information Systems Research", Journal of Management Information Systems, Volume 24 Issue 3, pp. 45-78, December 2007.

[10] K. Peffers, T. Tuunanen, C.E. Gengler, M. Rossi, W. Hui, V. Virtanen, J. Virtanen, "The design science research process: a model for producing and presenting information systems research", Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006), pp. 83-106, February 2006.

[11] T. Mshvidobadze, "Evolution Mobile Wireless Communication And LTE Networks", 6th International Conference on Application of Information and Communication Technologies (AICT), Georgia, pp 1-7, 2012.

[12] K. Yaici, J. M. Kulinska, "Sub-Saharan Africa Telecoms Market: Interim Forecast Update 2016-2021", Analysys mason, 2017

[13] G. Cheng, L. Liu, X. Qiang and Y. Liu, "Industry 4.0 Development and Application of Intelligent Manufacturing," 2016 International Conference on Information System and Artificial Intelligence (ISAI), Hong Kong, 2016, pp. 407-410.

[14] A. Botta, A. Pescape, C. Guerrini, M. Mangri, "A customer service assurance platform for mobile broadband networks", IEEE Communications Magazine, 2011, Issue 10 Volume 49, pp 101-109.

[15] V. Menkovski, "Computational Inference and Control of Quality in Multimedia Services", Philips Research, Springer International Publishing Switzerland 2015.

[16] A. Schwind; M. Seufert; O. Alay; P. Casas; P. Tran-Gia; F. Wamser, "Concept and implementation of video QoE measurements in a mobile broadband testbed" Network Traffic Measurement and Analysis Conference (TMA), pp. 1-6, 2017.

[17] ITU-T Recommendation G.1000, “Communications Quality of Service: A Framework and Definitions,” 2001.

[18] T. V. Staden, "Enterprise telecoms survey: operators must do more to overcome customer dissatisfaction", Analysys mason, 2017.

[19] S. Sale, "Challenger operators receive high customer satisfaction ratings in Sub-Saharan Africa", Analysys mason, 2017.

[20] F. Su, Y. Peng, X. Mao, X. Cheng and W. Chen, "The research of Big Data architecture on telecom industry," 2016 16th International Symposium on Communications and Information Technologies (ISCIT), Qingdao, 2016, pp. 280-284.

[21] Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao and R. C. Qiu, "Big Data Analytics in Mobile Cellular Networks," in IEEE Access, vol. 4, pp. 1985-1996, 2016.

[22] S. Bokun, H. He, A. Rao, "A SOC evolves from a cost centre to a revenue centre for some CSPs”, Analysys Mason, March 2016.

[23] Y. Ouyang; M. H. Fallah, "A performance analysis for UMTS packet switched network based on multivariate KPIs", 2010 Wireless Telecommunications Symposium (WTS), pp. 1-10, 2010.

[24] 3GPP, "TS 25.430 V10.1.0", 2011.

[25] J. Laiho, A. Wacker, T. Novosad, "Radio Network Planning and Optimisation for UMTS", Willey UK, 2006.

[26] R. Kreher, T. Rudebusch, "UMTS Signaling", Willey UK, 2007

[27] H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, V. Niemi, "UMTS Networks Architecture, Mobility and services", Willey UK, 2001.

[28] M. Sauter, "From GSM to LTE-Advanced", Willey UK, 2014.

[29] Shen Qingguo, Shen Ruisong and Wang Li, "QoS guaranteeing during UMTS packet-domain handover," Proceedings of the Fourth International Conference on Parallel and

Distributed Computing, Applications and Technologies, Chengdu, China, 2003, pp. 387-390.

[30] C. Rawal; R. Gupta, "A research on point to point QOS in 3G using OPNET MODELER", 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 796-799, 2016.

[31] P. Casas, M. Seufert, F. Wamser, B. Gardlo, A. Sackl, R. Schatz, "Next to You: Monitoring Quality of Experience in Cellular Networks From the End-Devices", IEEE Transactions on Network and Service Management, Volume 13, Issue 2, pp. 181-196, 2016.

[32] B. Upadhyaya, Y. Zou, I. Keivanloo, Joanna Ng, "Quality of Experience: User's Perception about Web Services", IEEE Transactions on Services Computing, Volume 8, Issue 3, pp. 410-421, 2015.

[33] M. Fiedler, K. D. Moor, H. Ravuri, P. Tanneedi, M. Chandiri, "Users on the Move: On Relationships Between QoE Ratings, Data Volumes and Intentions to Churn", 2017 IEEE 42nd Conference on Local Computer Networks Workshops (LCN Workshops), pp. 97-102, 2017.

[34] D. T. Larose, C. D. Larose, "Data Mining and Predictive Analytics", Wiley USA, 2015.

[35] B. Lantz, "Machine Learning with R", Packt Publishing UK, 2015.

[36] G. James, D. Witten, T. Hastie, R. Tibshirani, "An Introduction to Statistical Learning with Applications in R", Springer, 2013.

[37] J. Zhou, J. N. Kee-Yin, "A selector method for providing mobile location estimation services within a radio cellular network", First International Conference on Availability, Reliability and Security (ARES'06), pp. 89-96, 2006.

[38] O.S. Eluyode and Dipo Theophilus Akomolafe, "Comparative study of biological and artificial neural networks", European Journal of Applied Engineering and Scientific Research, pp. 36 – 40, 2013.

[39] L. Deng, D. Yu, "Deep Learning: Methods and Applications", Foundations and Trends in Signal Processing, vol. 7, nos. 3–4, pp. 197–387, 2013.

[40] I. Ni'mah, R. Sadikin, "Deep architectures for super-symmetric particle classification with noise labelling", 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pp. 169-174, 2016.

[41] L. F. Maimo, A. L. P. Gomez, F. J. G. Clemente, M. G. Perez, G. M. Perez, "A Self-Adaptive Deep Learning-Based System for Anomaly Detection in 5G Networks", IEEE Access, Volume: 6, pp. 7700-7712, 2018.

[42] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, R. Atkinson, "Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey", eprint arXiv:1701.02145, Jan. 2017.

[43] Z. M. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, K. Mizutani, "State-of-the-Art Deep Learning: Evolving Machine Intelligence Toward Tomorrow's Intelligent Network Traffic Control Systems", IEEE Communications Surveys & Tutorials, Volume: 19, Issue: 4, pp. 2432-2455, 2017.

[44] V. Sze, Y. Chen, T. Yang, J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey", Proceedings of the IEEE, Volume: 105, Issue: 12, pp. 2295-2329, 2017.

[45] V. Kumar, M. L. Garg, "Deep learning in predictive analytics: A survey", 2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT), pp. 1-6, 2017.

[46] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, J. Lloret, "Network traffic classifier with convolutional and recurrent neural networks for Internet of Things", IEEE Access, vol. 5, pp. 18042-18050, 2017.

[47] A. Oussidi, A. Elhassouny, "Deep generative models: Survey", 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), pp. 1-8, 2018.

[48] J. Heaton, "Introduction to Neural Networks with Java". Heaton Research Inc, 2005.

[49] J. Y. Li, T. W. S. Chow, Y. L. Yu, "The estimation theory and optimization algorithm for the number of hidden units in the higher-order feedforward neural network", Proceedings of IEEE International Conference on Neural Networks Volume: 3, pp. 1229-1233, 1995.

[50] K. G. Sheela, S. N. Deepa, "Review on Methods to Fix Number of Hidden Neurons in Neural Networks", Mathematical Problems in Engineering, 2013.

[51] K. Shibata, Y. Ikeda, "Effect of number of hidden neurons on learning in large-scale layered neural networks", Proceedings of the ICCAS-SICE International Joint Conference, pp. 5008-5013, 2009.

[52] D. Hunter, H. Yu, M. S. Pukish III, J. Kolbusz, B. M. Wilamowski, "Selection of Proper Neural Network Sizes and Architectures-A Comparative Study", IEEE Transactions on Industrial Informatics Volume: 8, Issue: 2, pp. 228-240, 2012.

[53] T. Vujicic, T. Matijevic, J. Ljucovic, A. Balota, Z. Sevarac, "Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network", 27th Central European Conference on Information and Intelligent Systems (CECIIS 2016), pp. 219-223, 2016.

[54] S. Huang, Q. Liu, T. Han, N. Ansari, "Data-Driven Network Optimization in Ultra-Dense Radio Access Networks", GLOBECOM 2017 IEEE Global Communications Conference, pp. 1-6, 2017.

[55] S. Charoenlap, P. Uthansakul, "Prediction of interference areas for 3G network based on drive test and throughput data", 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1-5, 2016.

[56] R. D. Octora, Iskandar, "Optimization technique and analysis in HSPA+ network", 9th International Conference on Telecommunication Systems Services and Applications (TSSA), pp. 1-6, 2015.

[57] K. Zhang, P. Yang, "WCDMA ratio network optimization approach based on measurement report", IET International Conference on Communication Technology and Application (ICCTA 2011), pp. 379-383, 2011.

[58] I. A. Lawal, S. A. Abdulkarim, M. K. Hassan, J. M. Sadiq, "Improving HSDPA Traffic Forecasting Using Ensemble of Neural Networks", 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 308-313, 2016.

[59] S. Zhou, J. Yang, D. Xu, G. Li, Y. Jin, Z. Ge, M. B. Kosseifi, R. Doverspike, Y. Chen, L. Ying,"Proactive call drop avoidance in UMTS networks",Proceedings IEEE INFOCOM, pp. 425-429, 2013.

[60] Y. Chen, N. Duffield, P. Haffner, W. Hsu, G. Jacobson, Y. Jin, S. Sen, S. Venkataraman, Z. Zhang, "Understanding the complexity of 3G UMTS network performance", IFIP Networking Conference, pp. 1-9, 2013.

[61] S. Kumar, M. Kumar Pandey, A. Nath, K. Subbiah, M. Kumar Singh, "Comparative study on machine learning techniques in predicting the QoS-values for web-services recommendations", International Conference on Computing, Communication & Automation, pp. 161-167. 2015.

[62] A. Y. Nikravesh, S. A. Ajila, C. Lung, W. Ding, "Mobile Network Traffic Prediction Using MLP, MLPWD, and SVM", IEEE International Congress on Big Data (BigData Congress), pp. 402-409, 2016.

[63] J. Moysen, L. Giupponi, J. Mangues-Bafalluy, "On the potential of ensemble regression techniques for future mobile network planning", IEEE Symposium on Computers and Communication (ISCC), pp. 477-483, 2016.

[64] A. Roshdy, A. Gaber, M. Khairy, "Radio Resources Dimensioning Using Machine Learning with case study in Live Network", International Conference on Computer and Applications (ICCA), pp. 67-73, 2017.

[65] J. Garcia, "A clustering-based analysis of DPI-labeled video flow characteristics in cellular networks", IFIP/IEEE Symposium on Integrated Network and Service Management (IM), pp. 991-994, 2017.

[66] U. Trivedi, M. Patel, "A fully automated deep packet inspection verification system with machine learning", IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1-6, 2016.

[67] S. Kaoprakhon, V. Visoottiviseth, "Classification of audio and video traffic over HTTP protocol", 9th International Symposium on Communications and Information Technology, pp. 1534-1539, 2009.

[68] Z. Huan, "Fast Packet Classification Algorithm for Network Video Stream", International Conference on Computer Science and Service System, pp. 614-616, 2012.

[69] Y. Dong, L. Yao, H. Shi, "Fine grained classification of Internet video traffics", 21st Asia-Pacific Conference on Communications (APCC), pp. 580-584, 2015.

[70] W. Zai-jian; Y. Dong; H. Shi; Y. Lingyun; T. Pingping, "Internet video traffic classification using QoS features", International Conference on Computing, Networking and Communications (ICNC), pp. 1-5, 2016.

[71] R. Dubin, A. Dvir, O. Pele, O. Hadar, "I Know What You Saw Last Minute—Encrypted HTTP Adaptive Video Streaming Title Classification", IEEE Transactions on Information Forensics and Security Volume: 12, Issue: 12, pp. 3039-3049, 2017.

[72] R. Nossenson, S. Polacheck, "On-Line Flows Classification of Video Streaming Applications", IEEE 14th International Symposium on Network Computing and Applications, pp. 251-258, 2015.

[73] Y. Shi, S. Biswas, "Using traffic analysis for simultaneous detection of BitTorrent and streaming video traffic sources", 9th International Conference on Communication Systems and Networks (COMSNETS), pp. 79-86, 2017.

[74] P. Tang, Y. Dong, Z. Wang, L. Yang, "Classification of Internet video traffic using multi-fractals", 17th International Symposium on Communications and Information Technologies (ISCIT), pp. 1-6, 2017.

[75] Q. Mao, F. Hu, Q. Hao, "Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey, IEEE Communications Surveys & Tutorials, (Early Access), Pages: 1-27, 2018.

[76] S. Peng, H. Jiang, H. Wang, H. Alwageed, and Y. Yao, "Modulation classification using convolutional neural network based deep learning model", in Proc. 26th Wireless and Optical Communication Conference (WOCC 2017), Newark, NJ, USA, pp. 1-5, 2017.

[77] Z. Xu, Y. Wang, J. Tang, J. Wang, M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs", 2017 IEEE International Conference on Communications (ICC), pp. 1-6, 2017.

[78] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, M. Ghogho, "Deep learning approach for Network Intrusion Detection in Software Defined Networking", 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM), pp. 258-263, 2016.

[79] N. Gao, L. Gao, Q. Gao, H. Wang, "An Intrusion Detection Model Based on Deep Belief Networks", 2014 Second International Conference on Advanced Cloud and Big Data, pp. 247-252, 2014.

[80] M. Lotfollahi, R. S. H. Zade, M. J. Siavoshani, M. Saberian, "Deep Packet: A Novel Approach For Encrypted Traffic Classification Using Deep Learning", eprint arXiv:1709.02656, Sep. 2017.

[81] Z. Wang, "The Applications of Deep Learning on Traffic Identification", Available online: https://www.blackhat.com/docs/us-15/materials/us-15-Wang-TheApplications-Of-Deep-Learning-On-Traffic-Identification-wp.pdf.

[82] B. Mao, Z. M. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, K. Mizutani, "Routing or Computing? The Paradigm Shift Towards Intelligent Computer Network Packet Transmission Based on Deep Learning", IEEE Transactions on Computers, Volume: 66, Issue: 11, pp. 1946-1960, 2017.

[83] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, K. Mizutani, "The Deep Learning Vision for Heterogeneous Network Traffic Control: Proposal, Challenges, and Future Perspective", IEEE Wireless Communications, Volume: 24, Issue: 3, pp. 146-153, 2017.

[84] Ö. F. Çelebi et al., "On use of Big Data for enhancing network coverage analysis," ICT 2013, Casablanca, pp. 1-5, 2013.

[85] M. Si, C. Lung, S. Ajila and W. Ding, "An Empirical Investigation of Mobile Network Traffic Data for Resource Management," 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, pp. 291-298, 2016.

[86] L. Jun, L. Tingting, C. Gang, Y. Hua and L. Zhenming, "Mining and modelling the dynamic patterns of service providers in cellular data network based on Big Data analysis," in China Communications, vol. 10, no. 12, pp. 25-36, Dec. 2013.

[87] Y. Jie, Z. Shuo, Z. Xinyu, L. Jun, C. Gang, "Characterizing smartphone traffic with MapReduce", International Symposium on Wireless Personal Multimedia Communications, WPMC, pp. 1-5, 2013.

[88] A. Botta, A. Pescape, C. Guerrini and M. Mangri, "A customer service assurance platform for mobile broadband networks," in IEEE Communications Magazine, vol. 49, no. 10, pp. 101-109, Oct. 2011.

[89] J. Keeney, S. van der Meer and G. Hogan, "A recommender-system for telecommunications network management actions," 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), Ghent, 2013, pp. 760-763.

[90] O. A. Kingsley, J. N. Dahj, "Modeling Of An Efficient Low Cost, Tree Based Data Service Quality Management For Mobile Operators Using In-Memory Big Data Processing And Business Intelligence Use Cases", 2018 International Conference on Advances in Big Data, Computing and Data communication Systems (icABCD)", At Uhlanga, Durban. South Africa, 2018.

[91] P. Fiadino, A. DAlconzo, M. Schiavone and P. Casas, "RCATool - A Framework for Detecting and Diagnosing Anomalies in Cellular Networks," 2015 27th International Teletraffic Congress, Ghent, 2015, pp. 194-202.

[92] M. Miyazawa and K. Nishimura, "Scalable root cause analysis assisted by classified alarm information model-based algorithm," 2011 7th International Conference on Network and Service Management, Paris, 2011, pp. 1-4.

[93] J. Verzani, "Getting Started with RStudio", O'REILLY, 2011.

[94] VMware, "Virtualization Overview", 2006. Available online at https://www.vmware.com/pdf/virtualization.pdf.

[95] M. Frampton, "Big Data Made Easy, A Working Guide to the Complete Hadoop Toolset", APRESS, 2015.

[96] D. P. Kingma, J. Ba, "Adam: A Method for Stochastic Optimization", eprint arXiv:1412.6980, Jan. 2015.

[97] K. Willems, "Keras Tutorial: Deep Learning in Python", 2017. Available online at https://www.datacamp.com/community/tutorials/deep-learning-python

[98] Y. Demchenko, C. Ngo, C. De Laat, P. Membrey, D. Gordijenko, "Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure". Proc. Secure Data Management (SDM'13) Workshop. Part of VLDB2013 conference. Italy, 2013.

[99] A. Alexiou, C. Bouras and E. Rekkas, "A Power Control Scheme for Efficient Radio Bearer Selection in MBMS," 2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, Espoo, Finland, 2007, pp. 1-8.

[100] M. Muntjir, A. T. Siddiqui, M. Rahul, "An Improved Data Mining Technique and Online Mining for Energy Efficiency in Wireless Sensor Networks: A Comparative Review", International Journal of Computer Science and Telecommunications, Volume 7, Issue 7, October 2016. pp. 19-24.

[101] A. Ananthram, "Deep Learning For Beginners Using Transfer Learning In Keras", 2018. Available online at https://towardsdatascience.com/keras-transfer-learning-for-beginners-6c9b8b7143e