Machine Learning Based QoE Estimation Model for Video Streaming over UMTS Network

BY: DIGIS WELDU

ADVISER: YIHENEW WONDIE (PHD)

A Thesis submitted to

School of Electrical and Computer Engineering

Addis Ababa Institute of Technology

In Partial Fulfillment of the Requirements for the Degree of Master of Science

(Telecommunication Engineering)

February, 2020

Declaration

I, the undersigned, declare that the thesis comprises of my own work and compliance with internationally accepted practices. I have fully acknowledged and referred all materials used in this thesis work.

Digis Weldu

Name	Signature

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Signed by the Examining Committee:

Internal Examiner	____________________	Signature	__________	Date	__________
External Examiner	____________________	Signature	__________	Date	__________
Adviser	Yihenew Wondie (PhD)	Signature	__________	Date	__________
Director of Postgraduate Program	____________________	Signature	__________	Date	__________

____

Dean, School of Electrical and Computer Engineering

ABSTRACT

The advent of data-intensive services needs quality Internet services. This in turn, makes Quality of Experience (QoE) gain prominent recognition in the telecommunications industry. Ethio telecom uses network Quality of Service (QoS) monitoring data obtained from Network Management Systems (NMS) tools to comprehend its network performances. However, as QoS measurement refers to network performances, this method does not generally give QoE data as perceived by the user. Therefore, QoE estimation models are proposed as solutions in the literature, recently.

This study focuses on developing QoE estimation models using QoS features of round-trip time (RTT), jitter, loss rate (LR) and throughput, and QoE scores collected using Application for prediCting QUality of experience at Interne Access (ACQUA)-based crowdsourcing in Universal Mobile Telecommunication Systems (UMTS) networks in a real-time basis. Data preparations techniques such as data cleaning and dataset imbalance corrections have been applied to the collected datasets. Machine Learning (ML) algorithms of Artificial Neural Network (ANN), K-Nearest Neighbor (KNN) and Random Forest (RF) are selected based on their suitability for multi-label problems. After training these models developed, they are evaluated using commonly used performance metrics such as accuracy, Root Mean Square Error (RMSE) and Receiver Operating Characteristics (ROC).

Experimentation results exhibit that RF with an accuracy of 98.39%, is the best model while KNN and ANN achieve 87.47% and 77.59% overall accuracy, respectively. As a conclusion, all three models achieve acceptable performances. As a conclusion, our QoE estimation models if implemented can help Telecommunications Service Providers (TSP) in estimating user QoE in real-time.

KEYWORDS

Universal Mobile Telecommunication Systems, Quality of Service, Quality of Experience, Supervised Machine Learning, Quality of Experience Estimation Models

ACKNOWLEDGMENTS

First, I would like to thank God for giving me the courage to complete my study. My special gratitude goes to my advisor, Yihenew Wondie (PhD) for his genuine pieces of advice throughout the thesis work and his suggestions are valuable and helped me shape my work. I would like to thank my examiners Beneyam Berhanu (PhD) and Dereje Hailemariam (PhD) for their constructive comments and feedbacks during my thesis progress seminars and final evaluations.

Last but not least, my special thanks goes to my wife, Shewit Hadgu and my daughter, Eldana. I would also like to express my gratitude to my family and friends for their encouragement and support throughout my study.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACQUA Application for prediCting QUality of experience at Interne Access

ANN Artificial Neural Network

ATM Asynchronous Transfer Mode

API Application Programming Interface

AUC Area Under the Curve

BTS Base Transceiver Station

CN Core Network

CS Circuit Switched

CV Cross-Validation

FAN Fixed Access Network

GSMC Global System for Mobile Communication

GGSN Gateway GPRS Support Node

GPRS General Packet Radio Service

GSM Global System for Mobile Communications

HLR Home Location Register

HSPA High-Speed Packet Access

HSPA+ enhanced HSPA

HSDPA High-Speed Downlink Packet Access

HSUPA High-Speed Uplink Packet Access

IP Internet Protocol

IPTV Internet Protocol Television

IQR Inter-Quartile Range

ITU International Telecommunications Union

KQI Key Quality Indicators

KNN K-Nearest Neighbor

LR loss rate

LTE Long-Term Evolution

MAE Mean Absolute Error

ML Machine Learning

MLP Multilayer Perceptron

MOS Mean Opinion Score

Mbps Megabits per second

ME Mobile Equipment

MSC Mobile Switching Center

MS Mobile Station

ms Millisecond

2G Second-Generation

3G Third-Generation

NMS Network Management Systems

NNOC National Network Operation Center

OM Operation and Maintenance

PLR Packet Loss Rate

PRR Packet Reorder Rate

PCC Pearson's Correlation Coefficient

PS Packet Switched

QoE Quality of Experience

QoS Quality of Service

RF Random Forest

RMSE Root Mean Square Error

ROC Receiver Operating Characteristics

RNC Radio Network Controllers

RTT round-trip time

SDU Service Data Unit

SGSN Serving GPRS Support Node

SMC Service Management Center

SMS Short Message Services

SMOTE Synthetic Minority Oversampling Technique

SRNS Serving Radio Network System

SVM Support Vector Machine

TSP Telecommunications Service Providers

UE User Equipment

UMTS Universal Mobile Telecommunication Systems

USIM UMTS Subscriber Identity Module

UTRAN UMTS Terrestrial Radio Access Network

VBR Video Bit Rate

VLR Visitor Location Register

VoIP Voice over Internet Protocol

WEKA Waikato Environment for Knowledge Analysis

WiMAX Worldwide Interoperability for Microwave Access

1

INTRODUCTION

High data rates become essential for Internet services. Users' preference of data-intensive services such as multimedia access, online streaming, online Internet gaming and video conferencing have led to the generation of huge data traffic and it will only get bigger in the future, where everything is believed to be interconnected. To suport it, mobile video traffic as forecast by Ericsson will account for 74% of all mobile data traffic in 2024 [1]. AS found in a report by Cisco [2], mobile traffic will represent 20% of the total Internet Protocol (IP) traffic and smartphones will surpass 90% of all mobile data traffic by 2022. Moreover, as predicted in [3], total mobile subscribers across the globe are expected to surpass 5.5 billion in 2022, as shown in Figure 1.1.

Therefore, TSPs are moving from the existing QoS-centric based quality managements to the more end-user-centric QoE-based quality management approaches. Since QoE-based quality management practices focus on network performances of a telecommunications services, it not been successful, QoE will overtake this approach. QoE approaches are now recommended in the literature to improve Internet services quality. QoE's prominence is largely due to its user focus rather than the services themselves. QoE unlike QoS, is a subjective metric concerned with human dimensions involving user perception, expectations and experiences [4].

International Telecommunications Union (ITU) defined QoS as *"The totality of characteristics of a telecommunication service that bear on its ability to satisfy the stated and implied needs of the user of a service"* [5]. On the other hand, ITU defined QoE as an *" Overall acceptability of an application or service, as perceived subjectively by the end-user".* More convenient QoE definition by ITU seems, *"The degree of delight or annoyance of a user of an application or service"* [5]. Moreover, [6] and [7] defined QoE as *"An overall user perception about a product or service".* The authors in [8], also defined QoE as *"The assessment of human expectations, feelings, perceptions, cognition and satisfaction with respect to a particular product, service or application".*

To monitor and ensure Internet quality at the user level, the concept of QoE is more appropriate than QoS. This is due to QoE's inclusion of various factors which are not included in the QoS approach such as expectations, perceptions and feelings of the individual user [9]. Evidently, QoS places more focus on the technical aspects of telecommunications networks; whereas, QoE

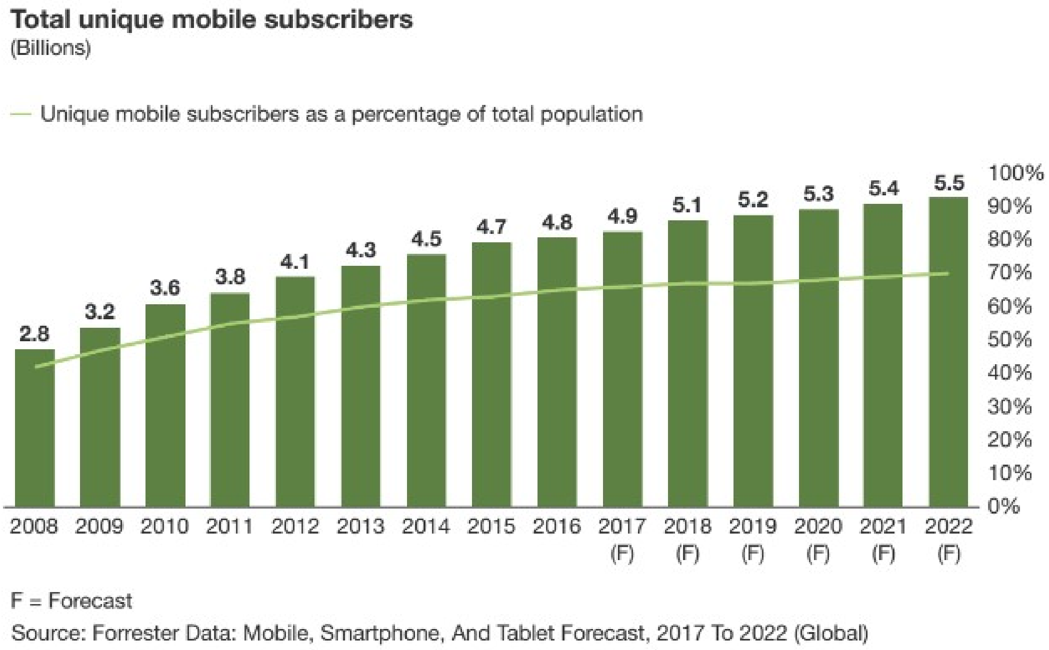


Figure 1.1: The Growth of Mobile Subscribers [3]

focuses more on end-user satisfaction. Internet video streaming services like other services' are mainly affected by network QoS parameters or network QoS features consisting of a delay, throughput, jitter, packet loss, bit-rate, bandwidth and signal strength [10]. However, researches are still attempting to identify the most influencing techniques used to measure QoE as accurate as the users of a service.

Telecommunications systems are communication infrastructures which can basically be divided into core, distribution, access and/or application domains. Quality and consistent network is important to ensure quality in UMTS networks. One thing that needs a note here is; however, quality of Internet services especially video streaming services can be affected by various network-dependent, application-specific, content-based, business and context-oriented factors [11] and [12]. Therefore, for multimedia service providers, understanding the degree of influence of various QoS factors on user QoE is a priority.

1.1 *BACKGROUND OF ETHIO TELECOM*

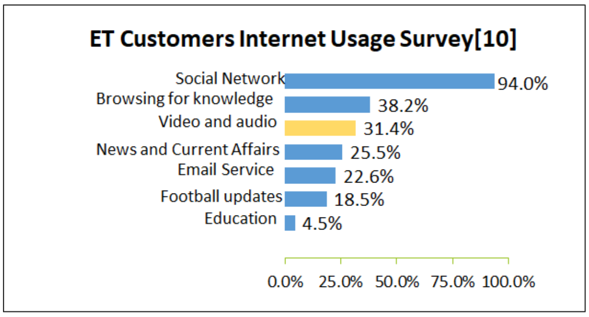
Ethio telecom is a state-owned and sole telecom operator in Ethiopia. Its customer base is growing fast that Ethio telecom becomes the largest mobile operator in Africa in 2017 in terms of subscriptions [13]. As of June, 2019, Ethio telecom has approximately around 35.94 Million mobile customers, out of which around 20% are Internet users [14]. Because of the tremendous demand for data-intensive services, Ethio telecom faces issues such as service capacity, availability and accessibly problems. Sometimes, it is evidently difficult to access Internet data using Third-Generation (3G) networks in Addis Ababa, the capital city of Ethiopia; where, data collection used for our experimentation is done.

In Ethio telecom, Key Quality Indicators (KQI) metrics used to analyze network performances are poor connectivity signal, low video starting success, video play interruption, frequent video stalling (delaying) or frequent play disconnection during online streaming. Most of these are what customers experience as end-users in defining quality of Internet services. User QoE depends not only on network QoS factors, but also by other issues such as type of application, equipment used, service type or contextual things. For example, network quality might be good for someone who uses a laptop to watch YouTube video online, but it might not be as good for someone who uses her/his mobile for Facebook video streaming service.

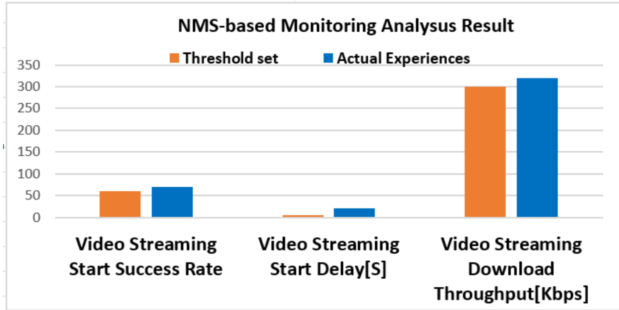
In assessing end-user QoE, users are the perfect quality measurement means, because they are the ultimate witnesses of any product or service. In Ethio telecom, mobile Internet services quality are monitored and analyzed using KQI performance data collected from NMS tools. These techniques as stated in [9] are focused on network performance from the access point to the core network. In other words, NMS tools do not indicate quality conditions between access network and the end-users' applications. So, performance information collected by NMS does not reflect the very end-users' satisfaction level and the crucial point i.e. QoE is missed out.

Therefore, QoE approach helps to look at how users are perceiving quality to the advantage of improving both users expectations and operators' better decision making. This is because, making better decision needs getting reliable and accurate end-user QoE information. Thus, there must be a new approach to capture end-users QoE perceptions subjectively [10]. Figure 1.2a depicts Internet service satisfaction survey results conducted by Ethio telecom's Marketing Research and Intelligence Department in June, 2018. Figure 1.2a shows Internet services popularity among customers and online video streaming takes 31.4% of all the services included in the survey. Figure 1.2b shows that Internet browsing quality-related information obtained from NMS for the same month of June 2018.

In contrast to survey results, NMS-based KQI monitoring analysis results show that overall video quality is around or even sometimes above the threshold set by Ethio telecom for video streaming services as shown in Figure 1.2b. This indicates, most users experience acceptable online video streaming services using UMTS networks in Ethio telecom. The thresholds for acceptable



(a) Survey Results per Service Type



(b) Network Performance from NMS

Figure 1.2: Customer Internet QoE Survey [16] vs Network Performance obtained from NMS [17]

network performances are set by Ethio telecom together with its vendors like Huawei Technologies Co., Ltd. However, according to the user survey analysis results mentioned above [16], only 14.2% of the participants are satisfied by the Internet services they get [17].

The gap can occur in any TSP globally. This can be due to fact that NMS tools emphasize on the network QoS performances. A survey conducted on 362 TSPs worldwide yielded that 80% of TSPs believed, they offer superior customer experience looking at their network performances. However, their customers agreed only 8% of them were really delivering [18]. This shows that existing network performance-based quality management approaches may not be effective in capturing user experiences.

1.2 *STATEMENT OF THE PROBLEM*

Poor Internet service results in degraded user QoE and high dissatisfaction. This in turn, may result revenue losses in the TSP side. As solutions, Ethio telecom currently uses both NMS tools-based network monitoring and user surveys. However, the existing approaches drawbacks are:

- **NMS Tools** - Uses network performance KQI data obtained from NMS. It indicates more of network performance, but not user experiences. Thus, it is difficult to estimate end-user QoE and map it to quantifiable scaled QoE numbers.

- **User Surveys** - Used to collect actual user QoE of a service. However, surveys are exhaustive, too expensive and time-consuming. Additionally, they often involve a handful of users, making it difficult to determine on the total population.

To solve this, ML prediction models can objectively estimate end-user QoE using network QoS conditions. Therefore, if implemented, our solutions may effectively capture user QoE. According to [15], existing state-of-the-art QoE estimation solutions usually use synthetic datasets collected from experimental setups or software simulations. However, our solution will be built using UMTS traffic collected from actual actual YouTube video streaming experiences in UMTS networks.

1.3 OBJECTIVES

1.3.1 General Objective

The research work has an objective of proposing ML-based estimation models that can predict service user QoE for Youtube-based streaming in UMTS network.

1.3.2 Specific Objectives

The specific objectives of the study are:

- To identify QoS factors impacting Internet video streaming service user experience;
- To build dataset using suitable data collection techniques;
- To develop QoE estimation models using ML-based techniques;
- To analyze the performance of these estimation models;
- To finally recommend the most accurate QoE estimation models based on our findings.

1.4 SCOPE AND LIMITATION OF THE STUDY

The study aims to provide end-user QoE estimation solution for video streaming services using selected UMTS QoS features. Though there exist some Internet video streaming applications, our crowdsourcing technique is limited to ACQUA-based YouTube streaming services. Secondly, though different factors can affect user QoE, the study uses only selected network level downlink measurements.

1.5 CONTRIBUTION OF THE STUDY

Our solutions if implemented may improve the way quality is monitored in TSPs and our findings may be used as inputs to the research areas community, because:

- The correlation results between QoS factors and QoE may help in understanding factors that are more influential for Internet quality degradation in streaming services.
- The proposed QoE estimation solutions may potentially be applicable for real-time QoE network monitoring and assessment in more accurate and efficient ways.
- Our solutions can serve as components for the monitoring and control building blocks of the larger QoE management frameworks consisting of the monitoring, control and manager blocks.
- In the future, the methodology used and subsequent findings of our work may contribute in identifying which ML algorithms can perform more accurately in the case of imbalanced dataset.

1.6 LITERATURE REVIEW

For service providers, it is important to quantify and measure QoE with accuracy. Quantifying QoE means translating user perceptions and performances into interpretative values. Measuring and analyzing users QoE is challenging because of the complexities involved in capturing users' perceived experiences. As stated in [15], the subjective QoE is presented through MOS labels, which are a five-point Likert scale (5=Excellent, 4=Good, 3=Fair, 2=Poor and 1=Bad).

A contemporary survey was conducted in [4] to analyze the impacts of network QoS factors over user-perceived quality. For data collection, the authors simulated wireless test-bed, where, a short video was streamed from a server to a client computer. Users watched the video and gave their quality perceptions using MOS rates. Using a small dataset, they found out that network QoS parameters of Packet Loss Rate (PLR) and Packet Reorder Rate (PRR) have exponentially degrading scatter plots, but Video Bit Rate (VBR) produced a logarithmic plot. When PLR and PRR increase, the perceived video quality (QoE of the users) decreases or vice versa. However, when VBR increases, user QoE increases or vice versa. Similar findings were obtained in [8], who are the first to use Rough Set Theory (RST) for quantitative assessment of the collected datasets.

Coming to ML techniques to develop prediction models, a work by [12] implemented and verified a solution using network delay, jitter and LR features labeled by MOS rates for Long-Term Evolution (LTE). For data collection, they built video streaming network simulators and users were able to watch and rate their MOS perceptions. Corresponding network QoS measurements were also captured to build the dataset in real-time. Then, they used the feed-forward ANN algorithm to evaluate their predictor models in Python. The authors used mean square error to validate their proposed prediction solution. Their findings showed that the performances of ANN was very good having a mean square error value of 0.22 which is less than the acceptable value of 0.25 [12].

QoE prediction models for Software Defined Network (SDN) was proposed in [19]. The K-fold Cross-Validation (CV) ML technique was used to train models in Waikato Environment for Knowledge Analysis (WEKA) workbench. Four ML algorithms, namely ANN, decision tree, KNN and RF were used. The authors performed some experimentation by varying K values for the K-fold CV. They found out that the estimation accuracy of ANN was worse than the other algorithms while RF was the best predictor model. The final performance results were achieved by experimenting the K-fold CV varying K-Values from one to ten. The estimation accuracy of RF was close to that of M5P, but this performance was for RF, at K=9, whereas, M5P was at K=6. Since larger K-value indicates better model [19], the authors concluded RF at k=9 was the best prediction model.

Furthermore, [20] did an ML-based QoE prediction. The objective of their work is to find out the impacts of class imbalance on prediction performances of selected ML algorithms. For this purpose, the authors conducted two different experimentation techniques. One with the imbalanced datasets and secondly, with balanced datasets collected from Internet Protocol Television (IPTV) users. Their findings indicated that ANN's performance accuracy was a lot improved for the balanced datasets in comparison to Support Vector Machine (SVM) and decision tree algorithms. On the other hand, ANN performances are more affected by data imbalances than SVM and decision tree. The authors also stated QoE prediction models can effectively be used as real Internet QoE prediction solutions.

Authors in [21] built ML-based QoE prediction/estimation model from QoS features of throughput, packet loss, jitter and delay for Worldwide Interoperability for Microwave Access (WiMAX) networks using an ANN algorithm. These network QoS attributes are the ones used for QoE prediction in this work. In comparison to the other reviewed papers, the authors of this paper [21] used relatively larger datasets totalling to 600 instances/data points to evaluate their QoE prediction models. The prediction model performances were in agreement with that of [12] who stated that ANN prediction models performed very well.

As a conclusion, the reviewed works used synthetically generated datasets obtained from either controlled experimental setup or software simulations. Moreover, the size of dataset used for experimentation was relatively small, ranging from 24 to 600 instances. In this work, we used real datasets obtained from UMTS customers using crowdsourcing and with relatively larger datasets in comparison to the ones in the literature reviewed.

1.7 *RESEARCH METHODOLOGY*

The methodologies we followed are briefly listed out as follows:

- First, selected literature, papers, books and electronic resources help us to identify and shape the objectives as well as to design the methodologies described as follows.
- Subjective crowd-sourcing data collection methodologies are used to build our datasets required for experimentation. These techniques will be discussed in detail in Section 4.3.2.
- Data preparation and pre-processing techniques like data cleaning, inconsistent datasets removal and correcting data imbalances are then performed.
- Three supervised ML algorithms namely: ANN, KNN and RF are chosen based on their suitability for multi-class problems and their prediction accuracy in RStudio and WEKA data mining tools.
- Then, the developed ML models performances are evaluated using performance metrics like accuracy, RMSE, precision, recall, F-measure and ROC.
- Finally, results and findings are discussed and recommendations are provided.

1.8 *THESIS ORGANIZATION*

The remaining parts of the paper are organized into five chapters. Chapter 2 discusses overview of UMTS technologies, its network architecture, UMTS quality attributes and QoS classes. A brief description of the existing QoS and QoE approaches in Ethio telecom are also included here. Chapter 3 introduces us to the concept of ML and discusses ANN, KNN and RF algorithms in detail. Data collection, preprocessing techniques and models evaluation metrics are covered in Chapter 4. Chapter 5 summarizes the results and findings of the work. Finally Chapter 6 consists of conclusions and recommendations of our thesis work.

2

OVERVIEW OF THE UMTS NETWORK

2.1 INTRODUCTION TO UMTS NETWORKS

UMTS is a 3G mobile network evolved from the Second-Generation (2G) systems of Global System for Mobile Communications (GSM) and General Packet Radio Service (GPRS). Due to limited capacity to support high-speed data in GSM and GPRS, 3G has emerged to support higher data rates than GSM and GPRS. When we say UMTS, we refer to the widely accessible groups of 3G networks. There are two UMTS technologies: High-Speed Packet Access (HSPA) and its enhanced HSPA (HSPA+) and both technologies are widely available in Ethiopia. There is also the LTE technology deployed in the capital city, Addis Ababa [22]. HSPA is a standard for wireless network communication in the 3G family. The HSPA family of network protocols consists of the High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA) for the down-link and up-link communications, respectively.

HSPA uses HSDPA for download traffic as it supports theoretically maximum data rates between 1.8 Megabits per second (Mbps) to 14.4 Mbps in comparison to the 384 Kilobits per second (Kbps) maximum data rate in the original 3G. When introduced, HSDPA provided such a significant speed improvement over older ordinary 3G that HSDPA based networks are referred to as 3.5G or Super 3G [23]. HSUPA supports data rates up to 5.7 Mbps and by design, HSUPA offers lower data rates than HSDPA. Like in all other TSPs, in Ethio telecom, HSDPA is used for the down-link streaming because the majority of network capacities are provided for down-links to match the usage patterns of cellphone users. The evolved HSPA+ has also been deployed by Ethio telecom to support the huge growth of mobile broadband services in a better way. HSPA+ is the fastest 3G protocol supporting data rates of 42, 84 and sometimes 168 Mbps for downloads and up to 22 Mbps for uploads.

2.2 OVERVIEW OF THE UMTS NETWORK ARCHITECTURE

As UMTS is evolved from GPRS by replacing the radio access networks, there is much similarity in their architecture [24] and [25]. The UMTS network architecture is described in Figure 2.1.

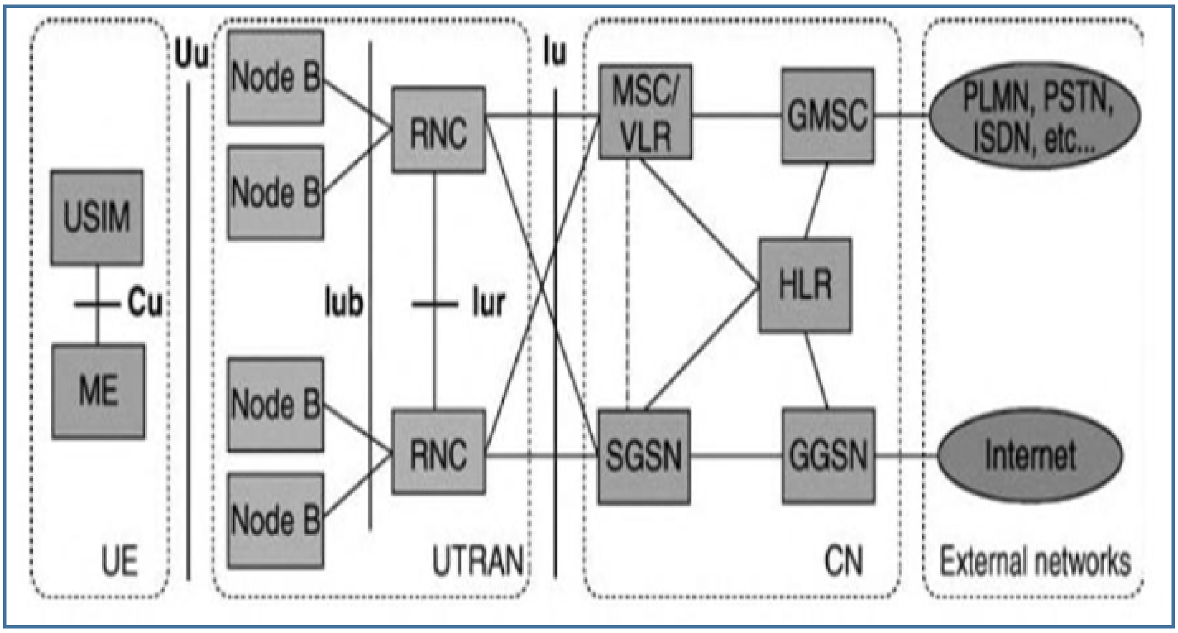


Figure 2.1: UMTS Network Architecture [24]

User Equipment (UE) consists of two parts: Mobile Equipment (ME), the 3G term for Mobile Station (MS) and UMTS Subscriber Identity Module (USIM). ME is used for radio communication with UMTS Terrestrial Radio Access Network (UTRAN) whereas, the USIM is a smartcard which holds subscriber information and authentication information. The UE connects with Node Bs through the radio interface Uu based on the Wide-band Code Division Multiple Access Code Division Multiple Access (WCDMA) technology. Three operation modes are defined for UMTS-based UE as stated in [24]:

- Packet Switched (PS)/Circuit Switched (CS) mode that UE is equivalent to GPRS Class A MS.
- PS mode that UE is equivalent to GPRS Class C MS.
- CS mode that UE can only attach to the CS domain.

Each part's descriptions can be further referred in [25]. UMTS consists of Node Bs (the 3G term for Base Transceiver Station (BTS)) and Radio Network Controllers (RNC) connected by an Asynchronous Transfer Mode (ATM) network. ATM is a protocol commonly deployed for

UMTS systems because of its low latency characteristics and QoS capabilities. The RNC and Node B serving an MS are called the Serving Radio Network System (SRNS), and it owns and controls radio resources in its domain. In UMTS, every Node B is connected to an RNC through the Iub interface. Every RNC is connected to an Serving GPRS Support Node (SGSN) through the Iu-ps interface, and to an Mobile Switching Center (MSC) through the Iu-cs interface. The RNC may be connected to several other RNCs through the Iur interfaces.

Core Network (CN) as detailed in [25] consists of Home Location Register (HLR), MSC, Global System for Mobile Communication (GSMC), Visitor Location Register (VLR), SGSN and Gateway GPRS Support Node (GGSN). HLR is a database which consists of a permanent profile of subscribers including information on permitted and forbidden services. MSC and VLR are switches and a temporary database for a copy of UE's location for services in CS services respectively. When UE needs to connect to external CS networks, the functionality is handled by GSMC. SGSN is similar in functionality to MSC/VLR of CS but is dedicated for PS services. The functionality of the GGSN is in line with GSMC though it is applicable only for the PS service.

The external networks are divided into two parts: the CS networks and PS networks. Connections like telephony or voice services to external networks are routed across the external CS network while PS services like the Internet are forwarded through the external PS network.

2.3 NETWORK QUALITY OF SERVICE ATTRIBUTES IN UMTS

General requirements to define the set of attributes characterizing a network QoS are covered in [26]. Negotiation between UE and CN gateway node for QoS attributes should be possible as well as renegotiating the QoS for active sessions. The UE and CN gateway node should be able to indicate the QoS properties to the application layer. Interoperability with previously existing QoS schemes should be assured and the overall complexity generated by the QoS mechanisms should also be lower. Mapping between the application QoS attributes and the UMTS services are done by the QoS mechanisms. The QoS mechanisms should assure different levels of QoS using the UMTS mechanisms independent of QoS mechanisms in other networks.

In UMTS, it should be possible to have different QoS attributes for multiple streams of a session. A session is considered to be a progression of events devoted to a particular activity [26]. A streaming service provided to a session is a distinct service with its own QoS attributes. For example, for a given session, simultaneous voice and data transfer should be possible. Each of the different streams should be provided with different QoS.

2.3.1 UMTS QoS Classes

Asymmetric bearers (with different QoS for up-link and down-link) should be supported. In order to better control the QoS mechanisms, Third-Generation Partnership Project(3GPP) demands application traffic differentiation into four profiles of services, named as QoS classes. According to [26], the differentiation among different QoS classes is mainly done considering the delay sensitiveness of the information to be carried.

1. **Conversational Class**: As the name implies, conversational class provides conversational services and comprises of real-time symmetric services such as Voice over Internet Protocol (VoIP) or video telephoning. Human perception of the maximum transfer delay defines the characteristics of this traffic class. So, it is suggested that fixed resources should be allocated in the network for conversational class services.

2. **Streaming Class**: Comprises typically one-way real-time services used by a human destination. Examples of such services include video downloading, news streaming, web-radio etc. For these services, low delay is not a stringent requirement due to application-level buffering in UE and UTRAN and due to the fact that buffering offers the appearance of real-time service to end-user.

3. **Interactive Class**: Provides an asymmetric non-real time service with more capacity for the down-link than for the up-link services. Interactive Web and database retrievals are examples of interactive services. If packet error happens, re-transmissions increase the delay; thus, diminishing the QoS. The low bit error rate is essential for this class.

4. **Background Class**: Background class services are characterized by the fact that the destination is not expecting the service to arrive within a certain time. Examples of such services include background delivery of e-mails, files or Short Message Services (SMS) messages. These classes require that the packets should be transmitted with a low bit error rate.

As discussed in [26], the main challenges that QoS in UMTS needs to overcome are service differentiation based on a set of traffic classes. This needs a simple and reliable translation mechanism between the different domains involved.

2.3.2 Mapping UMTS Attributes to QoS Classes

Telecommunication networks of any type should be monitored and managed to assure the implementation of the user agreements. Negotiation and modification of the QoS available from the network should be possible. End-to-end QoS has two dimensions. (1) A vertical one which refers to the mapping of high-level bearer service attributes into lower-level bearer service parameters and, (2) A horizontal one which implies translation of QoS attributes and QoS management mechanisms between different domains.

In the context of vertical mapping, it is important for the UMTS service bearer to meet the extent to which the standards elucidate the mapping towards the underlying bearer services. The mechanisms to map the UMTS service classes to attributes typical for IP based bearer services are summarized in Table 2.1 below.

Table 2.1: UMTS Bearer Attributes Defined for each Traffic Class [27]

UMTS QoS Attributes	**QoS Traffic Classes**			
	Conversational	**Streaming**	**Interactive**	**Background**
Maximum bit rate (kbps)	x	x	x	x
Delivery order (y/n)	x	x	x	x
Guaranteed bit rate (Kbps)	x	x		
Maximum Service Data Unit (SDU) Size	x	x	x	x
SDU format information (bits)	x	x		
SDU error ratio	x	x	x	x
Residual bit error ratio	x	x	x	x
Delivery of erroneous SDUs (y/n)	x	x	x	x
Transfer delay (ms)	x	x		
Traffic handling priority			x	
Allocation/Retention Priority	x	x	x	x
Source statistics descriptor	x	x		
Signaling indication			x	

SDU represents the payload of user data and the delivery order specifies if the UMTS bearer has to deliver the SDU in order or not. The allocation/retention priority is used to distinguish between bearers when allocating or retaining resources. Source statistics descriptor optimizes the service provided to a source with statistical properties, like conversational speech. The other QoS attribute names are self-explanatory and they can be referred at [27]. As it will be discussed more in the mentioned in Section 4.1, here, we consider RTT, jitter, LR and throughput as our QoS metrics and these features were used in [12] for LTE and [21] for WiMAX technologies, respectively.

2.4 *HIERARCHY OF QUALITY MANAGEMENT LEVELS IN UMTS*

Telecom operators monitor, assess or evaluate the performance of their network services to know what their customers feel on the services they offer. Ethio telecom currently evaluates its networks and service performances using KQI data collected from NMSs and occasional user surveys in cooperation with other survey expert institutes.

From personal observation and what is written in the literature, these measurements might not be enough to capture the actual experiences of customers [28]. Globally TSPs and their customers do not agree when they talk about QoS. The authors in [18] studied the gap between TSPs and their customers regarding the telecommunication service performances. As it has been mentioned in Section 1.1, a survey on 362 TSPs worldwide shows that 80% of the TSPs believed that they offer superior customer experience, but their customers agreed only 8% of them were really delivering [18]. The gap is so big that many pieces of research are dealing to close it so that both TSPs and their customers will come to the same terms when talking about end-user QoE.

Figure 2.2 shows the hierarchy of quality assessment indicators practiced in the telecommunications sector. QoE is at the top of the hierarchy showing the most perfect way of ensuring quality when TSPs reach at this point of the pyramid. Currently Ethio telecom has reached the KQI of the hierarchy showing that it still needs to move to the top of the pyramid in Figure 2.2(QoE level). Thus, at that point, both Ethio telecom and users will have the same quality perception for any given service.

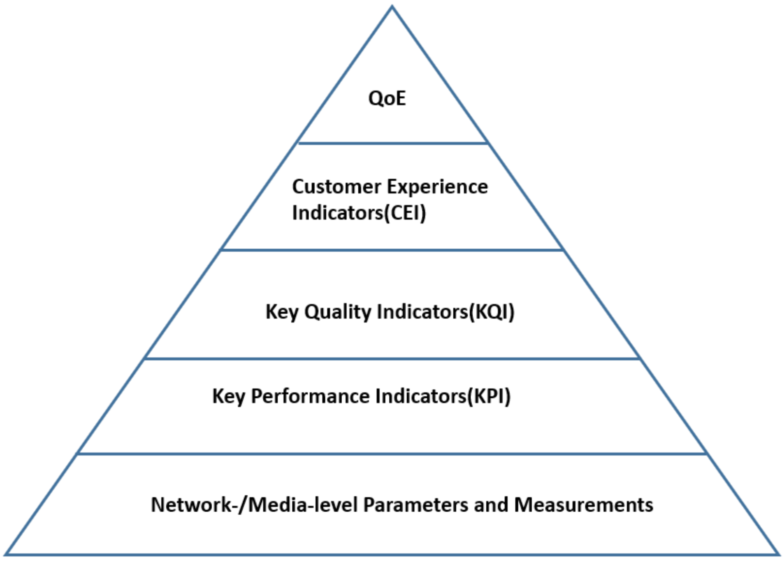


Figure 2.2: Hierarchy of Quality Assessment Indicators [29]

2.5 MANAGING SERVICE QUALITY IN ETHIO TELECOM

Existing organizational structure in Ethio telecom shows that the Customer Service and Network Division takes care of complaints coming from its customers. The Service Management Center (SMC) Section is responsible for ensuring end-user service quality through the network performance monitoring tools such as NMS tools. However, according to [22], Ethio telecom has no unique process for handling UMTS data service complaints. For example, if a customer complains about low down-link throughput when accessing mobile Internet service, Ethio telecom can looks at the network monitoring results obtained from Smart-Care NMS tools, but these tools do not capture the exact experiences of the end users.

There should be a clear processes not only for Internet services but also for all voice, SMS an other services. This would improve customer care by taking proactive measures and actions before complaints are received. The proposed ML-based QoE estimation models may mainly be used to proactively monitor, assess and manage end-user complaints. Generally, the drawbacks of the current practices in Ethio telecom regarding Internet quality are summarized as follows in [22].

- Internet quality complaints for fixed broadband network are handled using complainants handling processes, but cellular networks Internet service complaints are not handled properly.

- There are no clear methods for continuous monitoring and follow up of mobile Internet service quality-related submitted complaints. Such complaints are often handled through public, management or quality circle meetings.
- On the other hand, UMTS data service quality issues are occasionally handled in an informal way.

If formal communication with customers is established, complaints like UMTS data service speed degradation, low-speed throughput and delay in accessing websites can be properly managed using the structure depicted in Figure 2.3. The communication flow involves the Customer Service, SMC, Operation and Maintenance (OM), Fixed Access Network (FAN), National Network Operation Center (NNOC), Engineering and Vendor Support sections and departments.

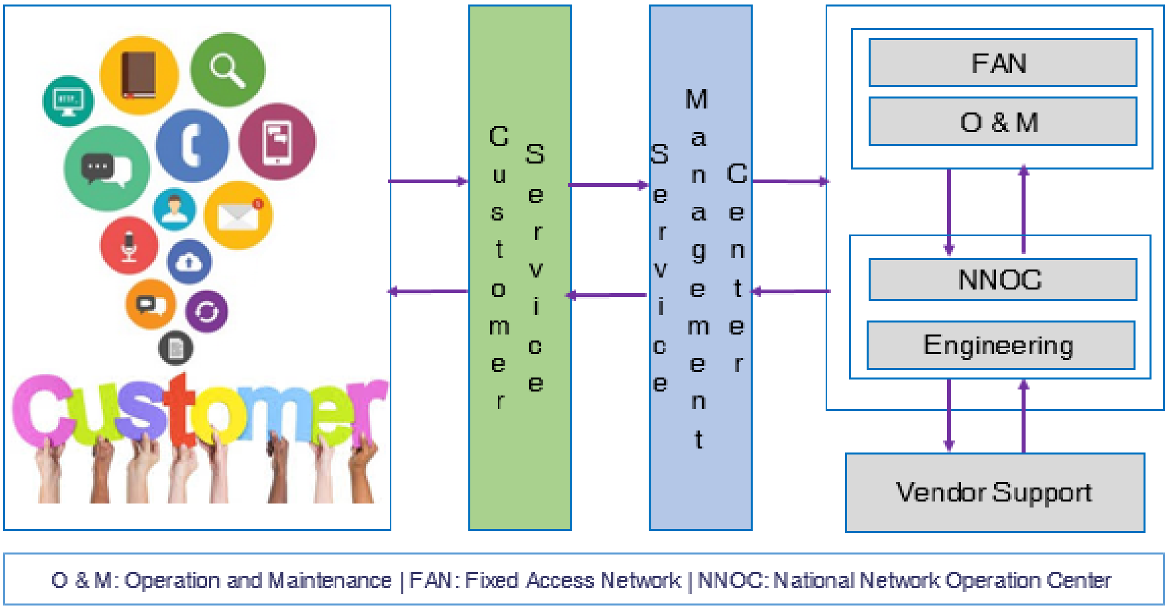


Figure 2.3: Organizational Structure to Handle Internet Quality Complaints in Ethio telecom [22]

There must be communications with customers when there is mobile data service problems or complaints. Customer Service is the interface for the customers and issues related to UMTS data service which cannot be resolved by Customer Service will be communicated to SMC. SMC can also communicate with other departments of the Network Division to resolve the complaints received. In addition to this, if there are problems which cannot be resolved by the departments under Network Division, there will be communication with Vendors for further support and maintenance. Finally, as the communication is bi-directional, customers have to be notified through Customer Service for better customer satisfaction after addressing the problems.

3

MACHINE LEARNING TECHNIQUES

ML is the science of making computers learn and act like humans by feeding data and information without being explicitly programmed [30]. It is the study of algorithms that automatically improve their performance with experience enriching their decisions through learning, which is attained by an iterative process. As it can be seen in Figure 3.1, the first one has the data regarding the identified problem. Algorithms and tools are chosen based on the behavior of the problem and data. These datasets are then fed to the algorithms and tools and the systems learn data patterns and can now analyze when new data is fed to them. That means, ML algorithms make decisions and predictions based on past data and what has been learned in the training stage.

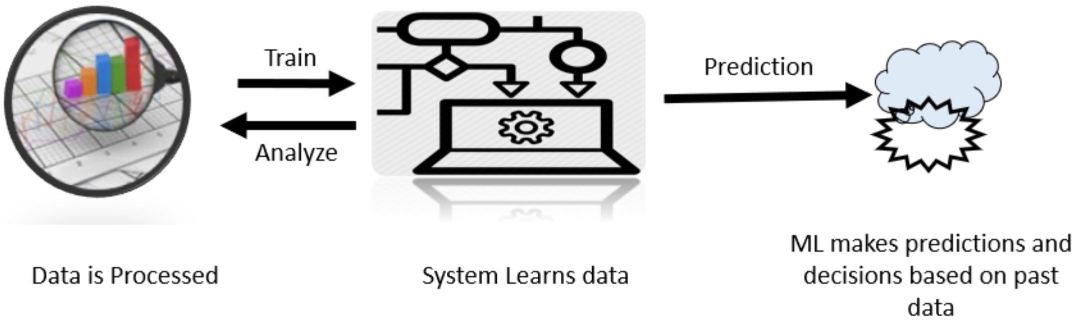


Figure 3.1: Machine Learning Working Principle [30]

ML provides mechanisms large data that are difficult to analyze using human computing capabilities to be automatically analyzed. There are several applications of ML, the most common of which is prediction, also called estimation depending on the type of solution required.

3.1 MACHINE LEARNING ALGORITHM TYPES

Generally, there are four categories of ML algorithms. They are supervised, unsupervised , semi-supervised and reinforcement learning.

3.1.1 *Supervised Learning Algorithms*

In supervised or predictive learning, the goal is to predict an event or estimate the values of a continuous numeric attribute. In supervised learning, there are input fields or attributes and outputs or target fields. Input fields are also called predictors because they are used by the algorithms to identify a prediction function for the output or class field. Supervised models can be described as learning a function f(x) = y, where y is the label (also called class) of the data and x denotes the attributes of these examples (also called features). We can think of input parameters as the X part of the function and the output field as the Y part or the outcome [31]. Supervised learning models are trained with data that have been pre-classified or labeled.

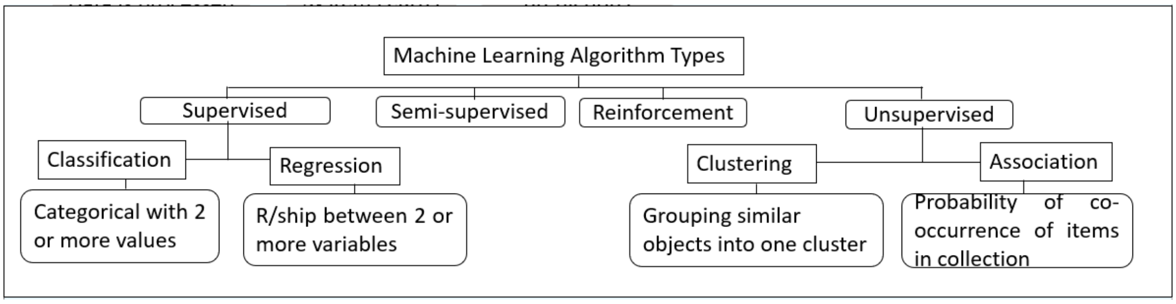


Figure 3.2: Machine Learning Types [30]

There are two main categories of supervised ML methods [32]: (1) classification and (2) regression. Classification uses data that has labels with two or more categories. This thesis uses classifications with five labels or MOS classes. They are the QoE or MOS values of bad(1), poor(2), fair(3), good(4) and excellent(5). Regression finds relationships between two or more variables. For example, when one variable increases the other variable may also increase or decrease, or vice versa. Based on this, there might be positive or negative relationships among the variables. In ML, examples of input-output functionality are referred to as the training data. Supervised learning is used when pre-classified training datasets are found. Some common supervised algorithms are logistic regression, ANN, decision tree types, gradient boosting machines, Naive Bayes, RF, SVM and KNN.

3.1.2 *Unsupervised Learning Algorithms*

In unsupervised learning, also called undirected learning, there is no output field or no label is given in the training data where instances are not named. According to the authors of the book in [32], the pattern recognition is un-directed or it is not guided by a specific target attribute. The aim of unsupervised learning is to identify patterns in the data that extend the knowledge and understanding of the domain that the data reflects. The goals of such ML models are to uncover data patterns in the set of input fields. Unsupervised ML algorithms are further classified as clustering and association as shown in Figure 3.2 above.

Clustering is the grouping of similar objects into one group or cluster. In these models, the groups are not known in advance. Instead, clustering needs the algorithms to analyze the input data patterns and identify the natural groupings of records or cases. When new cases are scored by the generated cluster model they are assigned to one of the revealed clusters [32]. Associations are used to show the probability of co-occurrence of items in datasets. They do not involve the direct prediction of a single field. Association models detect associations between discrete events, products, or attributes. The most famous unsupervised learning methods include k-means clustering, hierarchical clustering, and Self-Organizing Map (SOM).

3.1.3 *Semi-Supervised Learning Algorithms*

Semi-supervised learning is an ML method where a mixture of labeled and unlabeled data are used. This combination of classified and unclassified data is used in generating an appropriate model for the classification of data [33]. In semi-supervised, the labeled of the data can be used to aid the learning of the unlabeled part. Semi-supervised learning lends itself to most processes in nature and more closely emulates how humans develop their skills [30]. Semi-supervised is commonly used in artificial intelligence.

3.1.4 *Reinforcement Learning*

Reinforcement is a type of learning which is based on agents in a different environment. The agent learns how to behave in an environment by performing actions and reinforcements done based on the results. According to [30], the agent attempts to take a sequence of actions that may maximize a cumulative reward such as winning a game of checkers, for instance.

3.2 *SUPERVISED ML ALGORITHMS*

3.2.1 *Artificial Neural Network*

A neural network is an algorithm that is based on how the human brain works even though neural networks are not as complex as the brain [34]. This is because, there are two key similarities between biological neural networks and ANN. First, the building blocks of both networks are simple computational components that are highly interconnected. Secondly, the connections between neurons determine the function of the network. The neural network builds supervised prediction or estimation models by learning the patterns in historical data. The neural network is a collection of layered elements of neurons also called nodes connected with dendrites. Each node processes a small part of the task.

The most common type of neural network is called MLP, where the nodes are organized in layers linked with weighted connections [19] and [34]. The first layer is called the input layer, the outermost layer is termed as the output layer and between these two comes one or more layers which are called hidden layers. Each of the layers is interconnected by modifiable weights, which are represented by the links between the layers.

$$y_j = \sum_{i=1}^{n} (x_i.w_{ij} + B) \tag{3.1}$$

where x_i's can be the input features (RTT, jitter, LR and throughput) in our case, w_{ij} are the weights from node i to node j, B is the bias node, y_j is the output for that neuron and f(x) is the activation function.

$$f(x) = \frac{1}{1 + e^{-y_j}} \tag{3.2}$$

Equation 3.2, if *f(x)* is greater than the threshold values, the perceptron fires an output 1 else 0 (it does not fire). Training the perceptron aims at determining the optimal weights and bias values at which the perceptron fires. Most of the time, activation functions and intermediate outputs are included implicitly in the nodes and weights in the arcs (connections) between nodes. What an artificial neuron does when simply put is, it calculates a weighted sum of its inputs, adds a bias as shown in Equation 3.1. Then decides whether it should fire a signal or not as shown in Equation 3.2.

Figure 3.3 depicts a fully connected feed-forward MLP algorithm. The name feed-forward is used because, ANN completes as to the arcs (arrows) between the layers i.e. there exist all possible arrows from each node of a layer to the nodes of the following layer but there are no arrows between the nodes of the same layer. However, there are no lateral arcs (arrows) between the nodes of the same layer in feed-forward MLP networks. An MLP is an ANN with more than a single layer. It has an input layer that connects to the input variables, one or more hidden layers, and an output layer that produces the output variables.

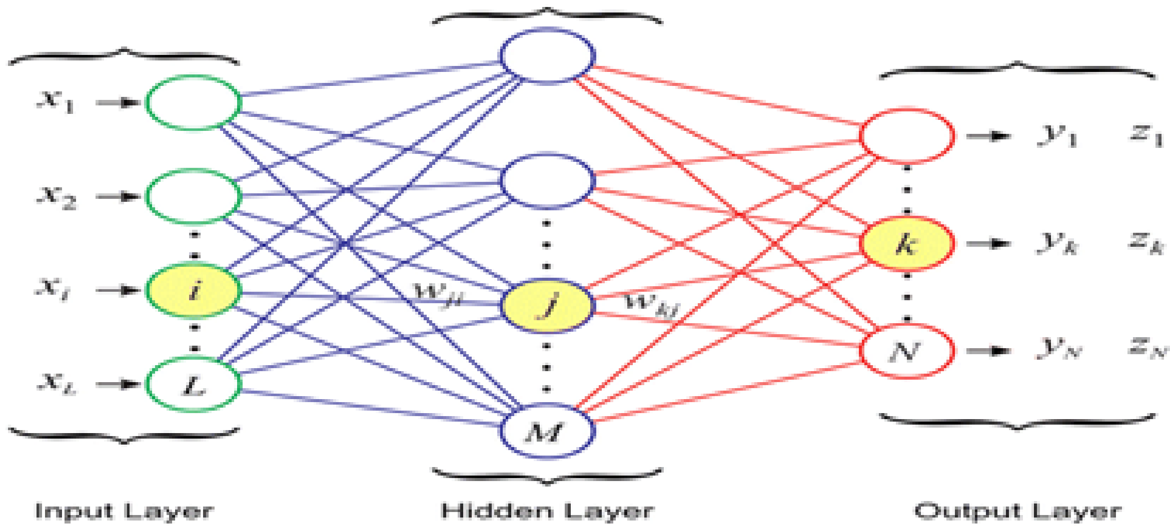


Figure 3.3: A Fully Connected MLP [19]

Bias nodes are added to feedforward neural networks to help ANN networks learn patterns. Bias nodes function like input nodes that always produce one or other constants. Because of this property, they are not connected to the input layer. The constants (B1, B2, B3) in Figure 3.3 above are the bias nodes but not all neural networks have bias nodes.

Another important unit in the ANN structure is the activation function, also called a threshold function or a transfer function. There are different types of activation functions such as linear function, sigmoid function, Hyperbolic Tangent(tanh), Rectified Linear Unit (reLU) etc. The most commonly used activation functions are the sigmoid function [35].

3.2.2 *K-Nearest Neighbor*

KNN is a supervised learning algorithm based on the underlying principle of "Tell me who your friends are, and I will tell you who you are" [19]. KNN makes use of neighbors' information to decide for new instances and it is one of the simplest and commonly used ML algorithms. KNN uses databases in which the data points are separated into several classes to predict the classification of a new sample.

KNN is considered a lazy learning technique, because the algorithm does not build a model using the training set until a query of a new data is performed [19]. The only calculations it makes are when it is asked to poll the data point's neighbors. This makes KNN very easy to implement for data mining. Supervised learning is done at run-time by observing the new data instance's closest neighbors. Each time, a prediction is done for a new instance, the algorithm is repeated and a search for new friends is performed [19].

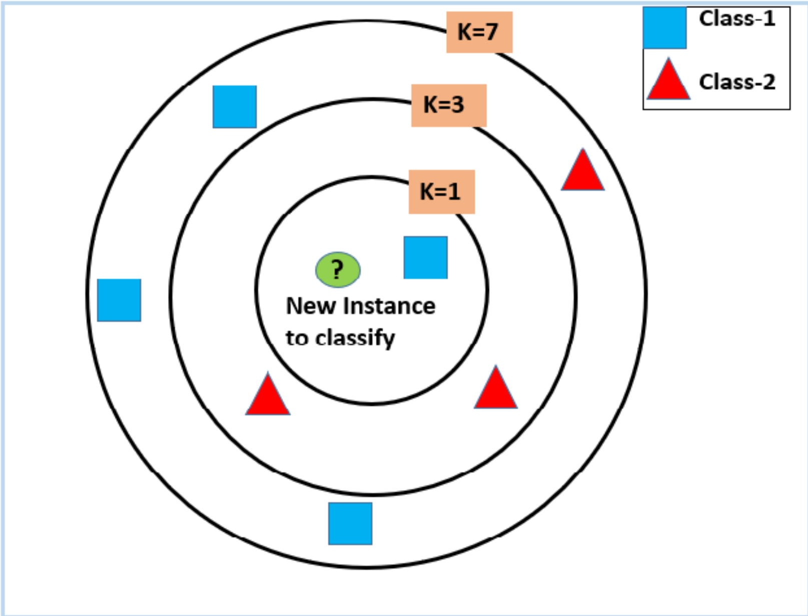


Figure 3.4: KNN Classification Example [36]

Figure 3.4 shows an example of a KNN based prediction model. The test sample (inside the circle) should be classified either to the first class of blue squares or to the second class of red triangles. For instance, if K=1, the new example is classified as Class 1 (blue rectangle) because there is only one neighbor which is the blue rectangle. Nevertheless, if K=3 (outside circle), the new example is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. So, class label decisions are determined by the majority label votes.

There are some neighbor distance calculation techniques used by KNN algorithms. The most common ones according to [37], are Euclidean, Manhattan, Minkowski and Chebyshev distances calculation methods. According to [19], the Euclidean distance is suitable for numerical class label types while the Manhattan distance is suitable for categorical label problems. Here, the Euclidean distance is used to calculate the distance from new data samples to the nearest neighbors in KNN algorithm. The Euclidean distance calculation from a new data to the neighbors has the form shown in Equation 3.3 below.

$$D((y_1...y_p),(u_1...u_p)) = \sqrt{\sum_{i=1}^{p}(y_i - u_i)^2} \tag{3.3}$$

where $(y_1 ... y_p)$ denotes the selected neighbors' class labels and $(u_1 ... u_p)$ represents new example for which neighbors are to be determined.

as stated in [19], after the calculation in Equation 3.3 for a new observation (x,y), the nearest neighbor $(x_{(1)}, y_{(1)})$ in the sample learning is determined by:

$$D(y, y_{(1)}) = mini_i(D(y, y_i)) \tag{3.4}$$

After experimenting K values from 1 to 10. In other words, the best accuracy performance has been achieved at 1-NN.

3.2.3 Random Forest

RF is a another type of supervised learning algorithm. As the name suggests, RF creates the forest from several trees. RF is a combination of multiple decision tree models and these kinds of models are called ensemble models. Other examples are boosting and bagging. RF is one of the most popular ensemble classifier relying on multiple decision tree prediction models [38]. RF uses majority votes among individual decision tree models. This potentially leads to much more robust and accurate models than learning using a single model.

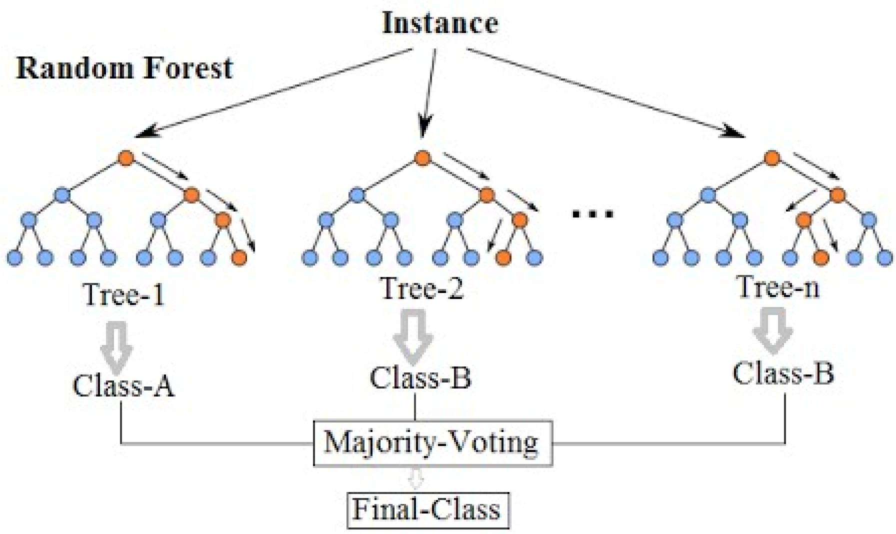


Figure 3.5: The RF Algorithm [39]

As shown in Figure 3.5, a tree includes one root node, several internal and leaf nodes. The leaf nodes correspond to decision results and the other nodes correspond to attributions test. The final model of a random forest is decided by the majority of votes produced by all individual decision trees. Each decision tree has a decision to label any testing data and each tree is built by classifying a random sample of the input data using a tree algorithm. Finally, RF model decides the classification results of the testing data after collecting the votes of all the tree models. For a given dataset D, how the trees are formed in RF are described as follows. First an entropy is computed as in Equation 3.5.

$$E(D) = -\sum_{i=1}^{c} P_i \log_2 P_i \tag{3.5}$$

where P_i is the probability of class c_i in the dataset, D.

Entropy is used as a measure of information in a tree. If the attribute A_i with v values, is made to be the root of the current tree, this will partition the total dataset, D into ‘v‘ subsets of D1, D2 … Dv. The expected entropy if A_i is used as the current root is shown in Equation 3.6.

$$E_{Ai}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \cdot E(D_j) \tag{3.6}$$

The information gained by selecting attribute A_i to branch or to partition the data, D is calculated using information gain by combining Equation 3.5 and Equation 3.6.

$$G(D, A_i) = E(D) - E_{Ai}(D) \tag{3.7}$$

RF works efficiently on relatively large datasets. RF also balances error and maintains accuracy by estimating missing data when a large proportion of the data in unbalanced datasets [38].

4

DATA COLLECTION AND PREPARATION

Here, descriptions of the selected QoS features and methodology (system model) are first presented. Then, data collection and data preprocessing techniques are explained in detail. Finally, some models performance evaluation metrics are described.

4.1 *FEATURES SELECTION*

In a telecommunications environment, network traffic passes through different devices like BTS, MSC, gateways, etc. In the meantime, traffic is disturbed or degraded because of end to end delay, packet reordering, packet loss, and/or packet errors in transmission from the source device to the destination in the network. Packets passing through these network devices facing long waiting in queues might be discarded due to errors and other related issues [40]. There are many important QoS attributes required for UMTS cellular networks as detailed in [41]. Some examples are maximum bit rate, delivery order, guaranteed bit rate, SDU format information, SDU bit error ratio, delivery of erroneous SDU and transfer delay.

In this research, a practical approach is taken that provides models to map network traffic QoS metrics to QoE directly. More importantly, for video streaming QoE estimation, network QoS parameters approach help to maximize the usage of all network traffic measurements which can be collected from smartphones, independently of the specific applications used [45]. Nevertheless, application-level QoE estimation is generally more cumbersome. This is because in most cases, not every application allows communication protocols or Application Programming Interface (API) to access relevant parameters, and device root access must be granted to perform measurements deeply into applications, hindering large-scale passive monitoring [45].

The number of features in prediction model development should be optimal. If less number of features are used, it becomes easy to interpret the results, but it may result in low prediction accuracy. However, if the number of features selected is larger, high prediction accuracy can be achieved. However, it is difficult to interpret and the resulting models are more likely to over-fit. Capturing accurate QoE requires measurements collected at multiple levels of the communications stack like the physical, network, application and device layers. The goal of the research is mapping network QoS measurements to user QoE directly and it is achieved using four QoS features described as follows.

A *Round Trip Time*

RTT, also called ping is the time required to transmit data packets from source to the destination and receive replies across a network. RTT is the time it takes for traffic to go both ways or it is the time it takes for a signal to traverse from point A to point B and back to A. On other hand, RTT is a two way trip time as shown in Equation 4.1. RTT may be impacted by the failure or overload of any element in the cellular network chain which is used to transmit data.

$$\text{Round Trip Time} = \text{Time Packet Received} - \text{Time Packet Sent} \tag{4.1}$$

where *Time Packet Received* is the time when a packet is received and *Time Packet Sent* is time when a packet is sent.

RTT is different from delay, because delay is only one-way time for a packet to be transmitted from source to destination. RTT can cause apparent loss of data in real-time communication flows such as in VoIP and online streaming services. RTT can also cause high network congestion in the case of reliable transmissions (TCP connection) caused by repeated re-transmissions or data losses when unreliable connections (UDP is used).

B *Jitter*

Jitter is an inter-packet arrival delay or it is the variance in delay between data packets over a network measured in a time unit. Jitter comes from a disruption in the normal sequence of arrival of data packets or from inconsistency in delay among packets of a message. Jitter like RTT can be a considerable problem for real-time and near-real-time communications including IP telephony, video conferencing, and virtual desktop infrastructure.

Jitter is an important QoS aspect that contributes to video quality degradation and in turn the user QoE. Jitter is characterized as having varying delays that could cause out-of-order video artifacts. The same as RTT, jitter can cause apparent loss of data in real-time flows such as VoIP and video streaming services. An application might be able to handle delay and jitter by using an appropriate buffer size. However, jitter might still be more difficult to deal with at the application layer and hence may cause significant QoE degradation [4].

C Loss Rate

Packet LR reflects the number of packets lost per the number of packets sent by an electronic host due to network impairments. The PLR represents the ratio of packets lost to the total number of packets sent. Each packet has a deadline or time to live before which must be executed. LR is often described as the number of packets lost per 100 packets sent as stated in Equation 4.2 below.

$$\text{Loss Rate} = \frac{\text{Packets Lost}}{\text{100 Packets Sent}} \tag{4.2}$$

where *Packets Lost* is packets lost in the communication network and *100 Packets Lost* is latest 100 packets transmitted from the application.

LR can be caused by a variety of factors such as network congestion, network element failure, inadequate signal strength, lower layer bit error rate, excessive system noise, hardware failure and software corruption [4]. In UMTS cellular video streaming, LR creates the artifacts in the video sequences; thus, negatively impacting the user's QoE.

D Throughput

Throughput is defined as the amount of data being sent or received in a unit of time. Throughput is the measure of how much data packets do actually travel through the network successfully. The amount of data packets are being actually transferred can be affected by many factors including devices capacity, latency, the protocols used etc. Throughput is different from bandwidth since bandwidth is the theoretical maximum units of data packets per unit of time; whereas, throughput is the actual units of data packets per unit of time.

$$\text{Throughput} = \frac{\text{Data Transferred}}{\text{Transfer Completion Time - Transfer Start Time}} \tag{4.3}$$

where *Transfer Completion Time* is the time data transfer is completed successfully and *Transfer Start Time* is the time data transfer starts.

4.2 SYSTEM MODEL

Streaming is an asymmetric one-way real-time service and the down-link QoS performance are more important than the up-link network performances. So, only the download network measurements and their corresponding MOS labels are used to prepare our datasets. In other words, most telecommunications users' activities are attached to watching or downloading videos than uploading their own videos.

The system model presents the model building methodology required for a multi-dimensional MOS prediction.

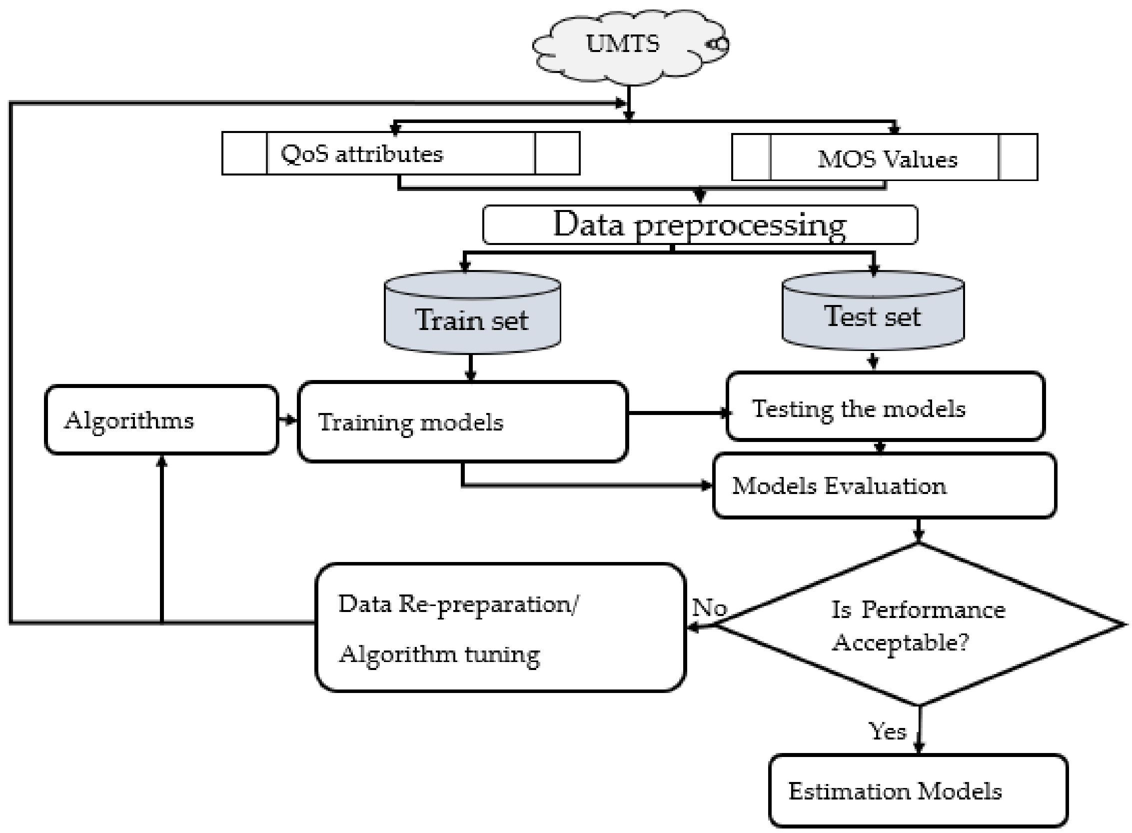


Figure 4.1: System Model

As shown in Figure 4.1, the system model begins with the collection of both QoS and QoE measurements from 3G Internet real-time streaming users. Each collected instance is labeled as either 'Bad', 'Poor', 'Fair, 'Good' or 'Excellent' during data crowd-sourcing. These QoE labels are equivalent to the MOS rates of 1, 2, 3, 4 or 5 respectively. The output of the experimentation is MOS estimation models built using ML algorithms. The remaining activities in the system models are described briefly as follows:

- Data is collected from UMTS telecommunications network using ACQUA, a subjective crowd-sourcing Android tool.
- Then, datasets are cleaned and preprocessed using some ML tools and techniques.
- The preprocessed datasets are transformed into Comma Separated (.CSV) values and then to Attribute Relation File Format (.ARFF) ready for experimentation.
- The dataset is divided into two separate sets of databases: training and test sets. Training sets are used to train the ML predictors. The test set is not involved in training, but used to validate the final prediction models.
- Next comes model training using ANN, KNN and RF followed by testing the developed models using the separated test set.
- Then, estimation models performance is analyzed using metrics such as accuracy, RMSE and ROC, F-measure etc.
- If the training and test performance results are acceptable, then the models become the final estimation models. Otherwise, the steps above are repeated some adjustments to the datasets and algorithm parameters until the desired level of performances is achieved.
- The final models are saved with their detailed statistics of prediction and prediction errors.

4.3 SAMPLING DESIGN AND DATA COLLECTION

4.3.1 Sampling Design

A population comprises of all the possible cases (people, objects or events etc.) in a study. Population constitutes a known whole of all the subjects one wants to study. In most cases, it is not feasible to include everyone in the population of interest and samples are used because they are considered to be true representatives of the whole population. Sampling is the process of selecting a group of subjects for a study in such a way that the individuals represent the larger group from which they have been selected. Sampling helps researchers to reduce the time and cost of contacting every member of the population, but with an acceptable range of data collection accuracy.

It is important that samples provide a representative cross-section of the population they supposedly represent. Otherwise, the study results using samples will be misleading when applied to the population as a whole. Yount's "Rule of Thumb" detailed in [42] is a sample size design technique that guides researchers on how to choose their sample size from a given population. The rule is based on the assumption that if the population is less than 100, then the rule guides you to include all of them as your samples as shown in Table 4.1,. However, when the survey population gets larger and larger, small representatives of the population are taken as samples of the population.

Table 4.1: Yount's Rule of Thumb [42]

Rule of Thumb	**Range of Population Size(N)**	**Sample Size as % of Population**
RT-1	0 - 100	100%
RT-2	101 - 1,000	10%
RT-3	1,001 - 5,000	5%
RT-4	5,001 - 10,000	3%
RT-5	Above 10,000	1%

Currently, it is believed that there are more than 4 million 3G active users in Addis Ababa and Yount's rule of thumb gives a sample size of 40,000 people as shown in Equation 4.4 below.

$$Sample - size = 4,000,000 * 1\% = 40,000 \tag{4.4}$$

Studies in [43] and [42] state that good sample size is between 100 and 1000 subjects for any population size, in which the accuracy of results stabilize regardless of how big or how small the sample size is. For this work, we found it important to balance between data accuracy and data collection costs. Therefore, a sample size of 300 participants is designed in this study, in total 230 people with success rate of 76.67% participated in the data collection survey. To minimize sampling error, crowdsourcing participants are randomized by including people in all corners of life such as men, women, students, professionals, businesspersons and homeworkers. Taking the trade-off between sampling bias, and the financial and time constraints, we included people in the mix of both within and without our convenient reach.

4.3.2 Data Collection

ACQUA is an open-source Android application which can be freely downloaded and installed on any Android smartphones, tablets or other similar devices. The detailed instructions on how to install the ACQUA App is depicted in the Appendix, Section A.4. The ACQUA tool is developed by [44] in 2017 and it has already been used by researchers in [45] and [46]. The tool measures and collects user-level network traffic conditions as well as providing QoE feedback rating capabilities while watching real-time YouTube videos in real-time. ACQUA is supported by a project in Antipolis, France. It presents a new way for the evaluation of the performance of Internet access starting from a network and device level measurements like signal strength, download and upload bandwidth, RTT, download and upload LR, download and upload jitters etc.

According to the ACQUA developers [44], ACQUA targets the estimated QoE related to the applications of interest without even the need to run them (e.g., estimated Skype quality, estimated YouTube video streaming quality). The crowd-sourcing participants (ACQUA users) have the luxury of watching any video of interest using YouTube 720P (High Definition). When users submit their MOS feedbacks, corresponding network traffic measurements for both the download and upload traffics and MOS rates are stored at a multitude of servers located at Antipolis, France.

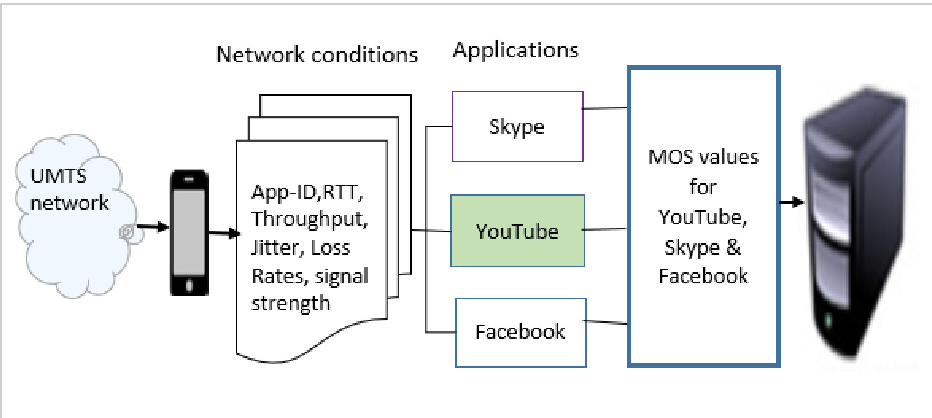


Figure 4.2: ACQUA Working Principles [44]

According to the developers the software [44], ACQUA uses supervised ML techniques to establish the links between measurements both at the network and device-levels to the QoE rates of Skype, YouTube and Facebook applications as shown in Figure 4.2. Since the application is still under development, the supported applications so far are Skype and YouTube 720P. The authors in [46], described ACQUA as a work in progress saying that ACQUA measurements are 70% accurate. Hence, this can limit the accuracy and practicability of of our work.

In total, 230 survey participants installed the ACQUA application on their Android smartphones and gave their feedback over a period ranging from mid-April to the end of September, 2019. As a requirement, participants should switch their data connection to 3G and make their data connection "ON". While watching YouTube online videos using ACQUA, users are allowed to send their MOS rates in real-time. Users are can use the tool at any time and place and as frequently as they like. The collected datasets are then sent back using an App-ID which is unique for each ACQUA application installed. Finally, we receive these collected datasets as E-mail attachments form the ACQUA admins.

4.3.3 *Survey Participants*

Here, survey participants included in the data collection using ACQUA are summarized below.

Table 4.2: Gender Distribution

Gender	**Frequency (Count)**	**Percent(%)**
Female	79	34.35
Male	151	65.65

Gender-wise, from a total of 230 survey participants, little more than one-third of them (34.35%) are female participants and the remaining slightly less than two-thirds of them (65.65%) are male as shown in Table 4.2. However,. This is in line with the study by [47], who surveyed gender distribution in Ethiopia. The study findings give suggestion on the probability of finding male to female ratios. According to [47], there were only 34.7% female mobile Internet users and women accounted to only 30% of professional jobs in scientific and technical sub-sectors in Ethiopia.

Table 4.3: Age Distribution

Age	**Frequency(Count)**	**Percent(%)**
Below 18	5	2.17
18-35	138	60
36-53	70	30.34
Above 54	17	7.39

As shown in Table 4.3, only five (2.17%) of the survey participants are under the age of 18. Large number of participants, 60%(138) are young people between the age of 18 - 35 while 30.34%(70) of them are between the age of 36 - 53. However, only 7.39%(17) of the crowd-sourcing participants are older people with an average age of 54 or above. This seems in line with the general belief that the youth has more access to Internet.

Table 4.4: Educational Background

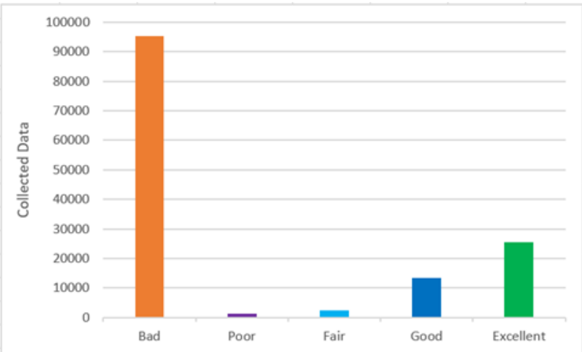
Educational Level	Frequency(Count)	Percent(%)
Diploma or Below	55	23.91
First Degree	123	53.47
Masters or Above	52	22.6

Table 4.4 shows that more than half of the participants or 53.47%(123) are first degree holders. 23.91% of the participants have college diploma or below; whereas, 22.6% have a masters or above.

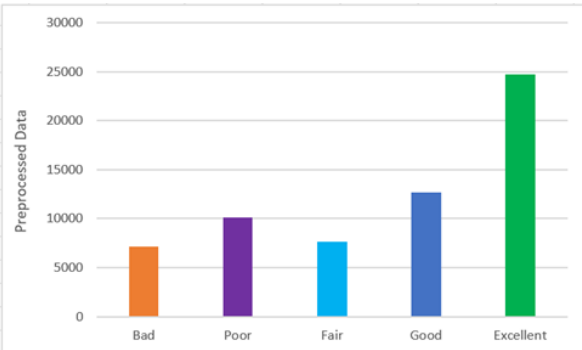
4.4 DATA PREPROCESSING

4.4.1 Collected Data Distribution

The total dataset collected from ACQUA is shown in Figure 4.3a. A whopping large number of data points (95,281 instances) of the collected datasets are labeled as MOS=Bad (the worst QoE possible). But MOS= Poor and Fair have small representatives, 1,306 and 2,240 instances respectively. The dataset exhibits an unequal distribution among its class labels.



(a) Collected MOS Distribution



(b) Final MOS Distribution

Figure 4.3: Collected vs Preprocessed Data Distributions

Possible reasons for obtaining such extremely large size of “Bad” MOS experiences collected from users could be due to:

1. The ACQUA crowd-sourcing tool takes a minute or two until it collects its environmental network conditions after it is started which varies depending on the type of smartphone device used. Survey participants are made aware of this by telling them to send their feedback after waiting for at least two minutes after they start their ACQUA tool. However, the nature of the collected datasets indicate that participants might have often forgotten the instruction and used to send feedbacks immediately. Datasets submitted within this time are always invalid due to the incorrect network measurement value recorded. For example RTT becomes infinite or RTT values become $1.5713X10^{308}$ nanoseconds.

2. We often remind survey participants via a telephone to use ACQUA and submit their feedbacks in their spare time. As we confirmed from some participants, they remember open the ACQUA tool and give their feedbacks during the morning, lunch-time, tea-time and/or in the evening when they are free from work or when they are back home. These periods are thought to be peak-hours or busy hours for mobile networks. Internet connection during this time becomes busy or QoE becomes poor. Therefore, the ACQUA based YouTube video streaming might actually "Bad" experience users.

4.4.2 Data Cleaning

Data points associated with measurement errors are not consistent with most instances are removed using an Oracle supported data cleaning. Because, these values may bias the outputs of the study. Such values come from experimental abnormalities or errors, and omitting them may improve algorithm performances. The data cleaning processes detected a whopping large, in total 91,775 such data points. These abnormal values are clearly identifiable instances by the human eyes. For example, survey feedbacks when there is no Internet, RTT is assigned by default "An infinite measurement value" or RTT value = $1.5713X10^{308}$ nanoseconds. These values are not consistent with the normal RTT measurement values which are in the order of some minutes.

The invalid data records are easily detectable by human eyes and refer to the total absence of an Internet connection. The five MOS classes have different distribution of these data points as these kinds of datasets are caused by measurement errors. From the collected 95,281 instances with "Bad" class, 91,775 instances are invalid and have been removed. So now, there are only 3,506 instances with "Bad" MOS class that are usable for an experimentation purpose. The total datasets are now reduced to only 45,976 instances.

4.4.3 *Outlier Detection and Removal*

Outliers are data points that differ greatly from the usual trend expressed by other values in the dataset. Before deciding whether to omit outlying values from a given dataset, we must identify the dataset's potential outliers. This is called outlier detection and it is difficult to detect outliers using human intelligence. In this work, the Inter-Quartile Range (IQR) algorithm is used to detect and remove outliers. As discussed in [48] in IQR, observations are first arranged in an ascending order starting from the smallest to the largest such as X_l, X_2, ..., X_n. The ordered data is broken into four quarters, the boundaries of each quarter defined by Q_1, Q_2, and Q_3, also called the 1^{st} quartile, 2^{nd} quartile and 3^{rd} quartile respectively.

The difference $|Q_3$ -$Q_1|$ is called what is called the inter-quartile range or IQR. The lower and upper thresholds for outliers are: $Q_1 - 3|Q_3 - Q_1|$ and $Q_3 + 3|Q_3 - Q_1|$ respectively· Observations falling beyond these limits are called major outliers and any observation, X_i, i = 1, 2, ..., n such that $Q_3 + 1.5|Q_3 - Q_1| <= X_i <= Q_3 + 3|Q_3 - Q_1|$ is called a possible outlier in the upper side. Similarly, $Q_1 - 3|Q_3 - Q_1| <= X_i <= Q_1 - 1.5|Q_3 - Q_1|$ is a possible outlier on the lower side. Out of the remaining total 45,976 data points, 2,436 data points have been removed automatically after being detected as outliers by the IQR algorithm.

4.4.4 *Class Imbalance Correction*

Imbalanced datasets consist of an unequal distribution of data samples. Data imbalance occurs in a multi-class problem where some datasets have small representatives in the dataset while other classes have larger samples. These with smaller representatives are called minority classes; whereas, the ones with larger representatives are called majority classes. Prediction model learned from an imbalanced dataset shows greater errors over the examples from the minority classes. This is a challenge especially to some ML algorithms as it becomes difficult to learn from the minority class data points.

There are two different sampling techniques to improve class imbalance problems[49]. (1) Under-sampling and (2) Oversampling techniques. Under-sampling methods work by reducing the number of instances of the majority class either randomly or by using some statistical knowledge to balance the class distribution. On the other hand, oversampling methods add new instances for the minority samples by random re-sampling the original minority class or by creating synthetic samples for the minority class. Although both approaches are used to improve classier performances over imbalanced data sets, oversampling is a lot more useful than undersampling.

Synthetic Minority Oversampling Technique (SMOTE) is a heuristic oversampling method that generates synthetic examples to over-sample the minority classes . Rather than replicating the minority observations, SMOTE works by creating synthetic observations based upon the existing minority observations. Its main idea is to form new minority class examples by interpolating between several minority class examples that lie together. By interpolating instead of replication, SMOTE avoids the over-itting problem and causes the decision boundaries for the minority class to spread further into the majority class space [50].

Table 4.5: Data Preparation

Collected Data	**Collection Period**	**Final Dataset**	**Training Set**	**Test Set**
137752 Instances	April 15 - Sep 24	62321 Instances	46740 Instances	15581 Instances

The number of SMOTE depends upon the amount of oversampling required to balance the data labels [50]. In this work, the new datasets are increased from 43,540 to 62,321 instances after SMOTE creating totally 18,781 new synthetic samples. Nevertheless, the data samples are not equally distributed among the five MOS classes. MOS classes 1, 2 and 3 are incremented by 100%, 1500% and 300% respectively, but MOS classes 4 and 5 remain the same after SMOTE is applied. Table 4.5 shows a summary of the total datasets collected, collection period, final datasets after preprocessing and how these datasets have been split into training and test sets.

4.5 *EXPERIMENTATION TECHNIQUES*

The three supervised learning algorithms of ANN, KNN and RF are selected based on their suitability for our multi-input multi-output problems. Their prediction performances are also among the best. Two ML experimentation techniques are used to build the our models: The K-fold CV and separate test.

A *K - Fold Cross Validation*

In K - fold CV, the dataset is divided into mutually exclusive and equal-sized K subsets. These subsets are trained k times on the union of K - 1 subsets and tested on the k^{th} subset. This is repeated iteratively changing the test subset from the 1^{st} , 2^{nd}, . . . to the k^{th} subset to get a distribution of the test error of the models. The average error rate of each subset is then the estimated error rate of the prediction model. K-fold CV is used to achieve an unbiased estimate of the model performances from the training and test datasets proportions. Ten-fold CV is the most commonly used and suitable technique for medium-sized datasets like our datasets and we also used the ten-fold.

B *Separate Test*

In user-supplied separate test, commonly known as the separate test, the user feeds the already split training and test datasets. The raining set is used to train a prediction model. To test it, the unseen the test sets are supplied by the user. The trained models are then tested using the unseen test sets.

4.5.1 *Performance Evaluation Metrics*

There are many prediction models performance evaluation metrics. The selected evaluation metrics are described as follows.

A *Confusion Matrix*

A confusion matrix is a table that is often used to describe the performance of a prediction model (classifier) for which the true values are known. All the performance metrics are derived from the confusion metrics but expressed in a different way. The table in Table 4.6 below shows the confusion matrix. The diagonal values from the top left to the bottom right represent the correctly predicted values (True Positive and True Negative); whereas, the diagonal values going from the top right corner to the bottom left values are the incorrectly predicted values (False Negative and False Positive).

Table 4.6: The confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

B *Accuracy*

Accuracy is one of the most commonly used performance metrics. Accuracy is the number of correctly predicted dataset instances/examples divided by the number of totally predicted instances as shown in Equation 4.5.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.5}$$

where *TP* is *True Positive*, *TN* is *True Negative*, *FP* is *False Positive* and *FN* is *False Negative* in the confusion matrix. The closer accuracy is to 1 or 100%, the better the model is. In this work, accuracy is a primary evaluation criterion.

C *Precision*

Precision is the number of true positive predictions divided by the total number of positive predictions as shown in Equation 4.6. Put another way, precision is the number of correctly predicted MOS examples for a given MOS class divided by the total number of MOS examples that are predicted as that MOS class.

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

where *TP* is *True Positive*, *FP* is *False Positive* in the confusion matrix.

D *Recall*

Recall is the number of true positive predictions divided by the number of actual positive class values in the training data as shown in Equation 4.7. In another way, recall is the number of correctly predicted MOS class examples divided by the total number of actual MOS class examples collected as that MOS class in the training set.

$$Recall = \frac{TP}{TP + FN} \tag{4.7}$$

where *TP* is *True Positive* and *FN* is *False Negative*.

E F-Measure

F-measure is also called the F-Score or the F1-Score and it conveys the balance between the precision and the recall. F-measure is the combination of both precision and recall into one and it is better than accuracy when correctness is very important.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.8}$$

F Root Mean Square Error

RMSE is a quadratic scoring rule which measures the average magnitude of the errors between the actual MOS class examples and the model predicted class examples. RMSE in another way means, the average of the squares of the difference between the forecast and corresponding observed MOSs, and the square root of the average is taken as expressed in Equation 4.9.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)^2} \tag{4.9}$$

where x_i and y_i represent the *Collected Subjective MOS* and the *Predicted MOS* respectively. N is the number of instances used to train or test the models.

G Receivers Operating Characteristics/ Area Under the Curve

ROC is a plot of the True Positive Rate against the False Positive Rate where the formulas for True Positive and False Positive Rates. ROC is a two-dimensional graphical illustration of the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate (1-Specificity). According to [51], it illustrates the behavior of a classifier without having to take class distribution into much consideration.

5

RESULTS ANALYSIS AND DISCUSSION

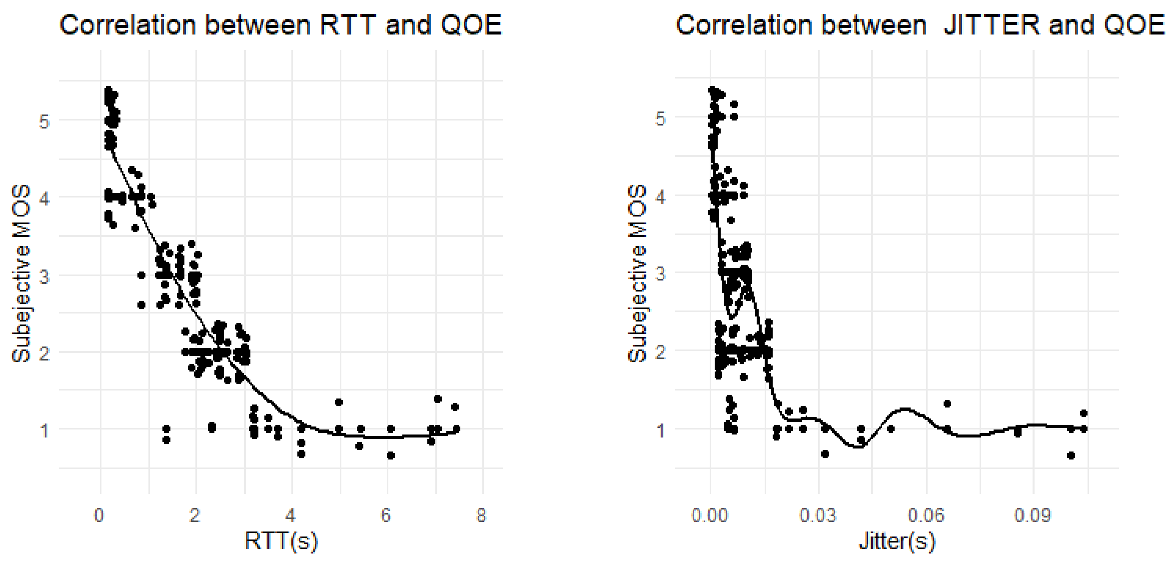
Experimentation results and performance evaluation comparisons of the resulting QoE estimation models are discussed here. The chapter begins with the correlational analysis between the QoS features measurements values and their corresponding MOS values. Developed models performance results are then discussed next.

5.1 CORRELATION BETWEEN QOS ATTRIBUTES AND QOE

The effects of selected QoS attributes on QoE as perceived subjectively by the user are examined. To understand the strength of the relationships between the independent variables (QoS features) and the dependent variable (MOS values) are further compared using Pearson's Correlation Coefficient (PCC) or R values.

When we see the correlation between RTT and end-user QoE, it is an exponentially degrading scatter plot as depicted in Figure 5.1a. When RTT measurements are concentrated around the X-axis (near to zero), the curve is observed to be at MOS = 5. However, when RTT values increase to one or two seconds, MOS rates go sharply down to 2. For RTT values between 3 seconds to 8 seconds, user QoE becomes the worst possible (or MOS = 1). R value for RTT is, R = -0.87, showing that QoE is strongly correlated with RTT. The negative sign (-) indicates that RTT has a degrading impact on user QoE.

Likewise, when there is no jitter (jitter values = 0), corresponding MOS values become maximum (MOS = 5). Nevertheless, when jitter raises to some fractions of seconds, the exponential curve drastically falls down to MOS = 3. At jitter values approximately from 0.005 to 0.02 seconds, MOS becomes 2. At around 0.02 seconds or 20 Millisecond (ms), MOS scores become the worst (MOS = 1) and remain there for all higher jitter measurements as shown in Figure 5.1b. Since jitter values vary greatly, the exponential curve looks like Figure 5.1b. For jitter, R = -0.524 indicates that jittter and MOS scores have an inverse correlation. On the other hand, when network jitter increases, the users QoE level degrades or vice versa.



(a) RTT vs QoE

(b) Jitter vs QoE

Figure 5.1: Correlation between RTT and Jitter against User QoE

Correlational relationship between LR and QoE is also expressed by the resulting R scores, R = -0.46. LR's resulting curve is an exponentially degrading curve similar to that of RTT and jitter. On the other hand, when the number of messages lost in the network increases, Internet user MOS becomes lower or vice versa. Unlike RTT, jitter and LR, throughput exhibited a positive, but the weakest correlation with R = +0.25. The positive(+) shows that the scatter plot is an increasing logarithmic curve indicating that when throughput increases, user QoE also increases or vice versa.

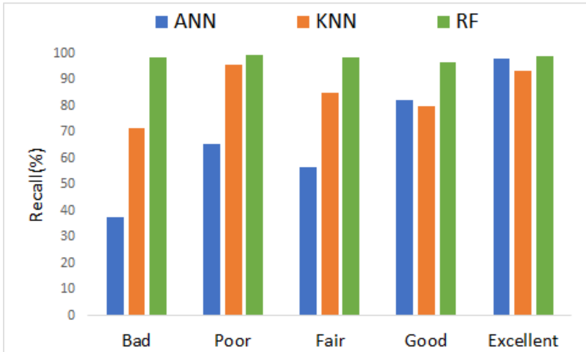
Similar findings (PLR (R = -0.91), PRR (r = -0.95) and VBR (r = +0.97)) were found in [4]. Here, the degree of correlation between the QoS features and QoE rates is weaker. This could be attributed to the subjective nature our data and measurement accuracy problems of our data collection tool.

5.2 MODELS PERFORMANCE ANALYSIS

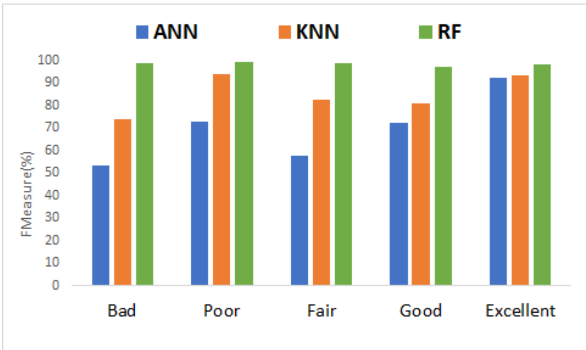
Here, we summarize the performance comparisons of our developed prediction models. Accuracy performances of the three ML algorithms using RStudio and WEKA for the ten-fold CV technique are first evaluated. Tools accuracy results show that both tools (RStudio and WEKA) have very close experimentation performances. In other words, All ML algorithms have no significant performance gap in both tools. This gives us more confidence to pursue our experimentation using the WEKA workbench to build our estimation models.

Accuracy is an important metric and we use the term overall accuracy, because accuracy values differ among the five MOS classes. The final accuracy values are the average of all individual MOS accuracy scores. Table 5.1 shows that RF outperforms both ANN and KNN with an overall accuracy of 98.39% in the ten-fold CV. ANN and KNN have overall accuracy performances of 77.58% and 87.48% respectively. Coming to RMSE as shown in Table 5.1, RF is the best with an RMSE value of 0.07. Nonetheless, ANN and KNN have RMSE scores of 0.26 and 0.22 respectively. So, taking accuracy and RMSE as performance metrics, RF is the best of all three whereas, ANN is the least performer.

(a) Precision



(b) Recall



(c) F-Measure

Figure 5.2: Models' Recall, Precision and F-Measure Performances

Figure 5.2 depicts precision, recall and F-measure respectively. Each MOS label is represented with different bar-plots indicating each algorithm performs differently for each of the Bad, Poor, Good Fair and Excellent QoE labels. The precision, recall and F-measure performances of RF is almost perfect for all MOS classes. ANN and KNN have comparable precision performances with KNN becoming slightly better than ANN in recall and F-measure. The excellent performance by RF for all five MOS classes agrees with the findings in [45] and [51]. The excellent performance by RF is because, it is less affected by class imbalances in comparison to other algorithms like ANN [51].

Table 5.1: Models Performance Summary

Validation Techniques	Algorithms	Time(S)		Performance Results					
		Build	Evaluate	Accuracy(%)	Precision(%)	Recall(%)	AUC(%)	F-Measure(%)	RMSE
Ten-fold CV	ANN	23.21	178	77.59	79.1	77.6	89.1	77.6	0.26
	KNN	0.01	90	87.49	87.4	87.5	92.0	87.4	0.22
	RF	17.79	177	98.39	98.4	98.4	99.7	98.4	0.07
Separate Test	ANN	8.95	0.04	85.30	84.5	85.5	94.1	85.2	0.22
	KNN	0.02	14.53	87.03	87.0	87.0	90.7	87.0	0.23
	RF	7.5	0.78	98.63	98.6	98.6	99.5	98.6	0.07

The summary of performances is depicted in Table 5.1. In ten-fold CV, the average precision, recall and F-measure for ANN is 79.1%, 77.6% and 77.6% respectively. KNN scores an average precision, recall and F-measure values of 87.4%, 87.5% and 87.4% respectively. RF outperforms both by achieving almost equally 98.4% for precision, recall and F-measures performances. So, in ten-fold, RF yields a very good performance and both KNN and ANN also achieve good performances with KNN significantly performing better than ANN.

In the separate test, significant models performance variations are observed. Looking at Table 5.1 once more, ANN's the overall accuracy is improved to 85.3% in ten-fold. However, this is not the case for KNN and RF that produce an overall accuracy of 87.03% and 98.63% respectively that show little improvements. Therefore, both ANN improved its performances in the separate test method.

In the separate test method, the average precision, recall and F-measure of ANN are 84.5%, 85.5% and 85.2% respectively exhibiting a significant improvement from the ten-fold technique. KNN scores the same (87.0%)in precision, recall and F-measure in the separate test to again become the second-best QoE estimation model. However, RF performs exceptionally well with the same performance of 98.6% in precision, recall and F-measure. The RMSE scores in the separate test technique remains almost similar to that of the ten-fold for all three models. Overall, RF with an accuracy of 98.6% and RMSE 0.07 is once the best performer in the separate test method.

The model building and evaluation times are important because, the models are going to be implemented in real-time. In the ten-fold CV method with building and evaluation time of 23.21 and 178 Seconds respectively, ANN is the slowest algorithm. In both methods, KNN has the least building time with less than 0.02 Seconds. However, in the separate test, ANN with 0.04 Seconds has the least evaluation time.

Observing Figure A.1 in Appendix A, the diagonal values from the bottom-left to the top-right corner of the graphs represent the trade-off between the Sensitivity (True Positive Rate) and 1-Specificity (False Positive Rate) for the produced models. This diagonal line has an Area Under the Curve (AUC) value of 0.5 and all AUCs should be above this threshold. For a well-performing classifier, the ROC curve needs to be drawn as far to the top left-hand corner as possible. As shown in Figure A.1, five ROC curves are drawn per each MOS class to get a better visualization of the performances of the algorithms. The average ROC performance comparisons of each algorithm for the 10-fold CV and separate test (supplied test) is included in Table 5.1.

Though class imbalances have partially been improved using the SMOTE algorithm, there is still dataset imbalance among the MOS classes. RF produces a perfect ROC curve for all five MOS with an AUC score of 99.7% in the ten-fold CV. In other words, RF has the best ROC stretching to the top left corner of the picture i.e. the upper 90^0 (0,1) covering large AUC. ANN and KNN achieve AUC of 89.1% and 92% respectively. Generally MOS classes with good sample representatives performed well than those that have fewer samples in ANN and KNN. This strengthens the findings in [45] and [51] that RF is less affected by class imbalance in comparison to similar ML algorithms.

In conclusion, RF is the best model that perfectly fits our QoE prediction solution based on the evaluation criteria used in this thesis. This can be because in addition to its robustness to class imbalance problems, RF is built out of many decision tree algorithms out of which the best model is selected using majority votes among the tree models. Similar outputs were found in [19] and [45], where RF outperforms ANN, KNN and M5P decision tree. Therefore, out of the three QoE estimation models proposed here, RF is the best model. For the experimentation techniques, ANN shows significant improvement in the separate test; whereas, both KNN and RF produce comparable results in the ten-fold and separate test methods.

5.3 *MODELS VALIDATION PERFORMANCES*

After developing our QoE estimation models, it is important to quantify how well they fit to future real observations. One of the simplest methods is to validate the models using test sets and measure the errors between the estimated and user collected MOS label counts. To validate performance accuracy of our multi-class MOS estimation models, we used the test sets. Here, all MOS labels are removed so that each model produces its own MOS labels for the unlabeled datasets based on the patterns that have been learned in the training stages.

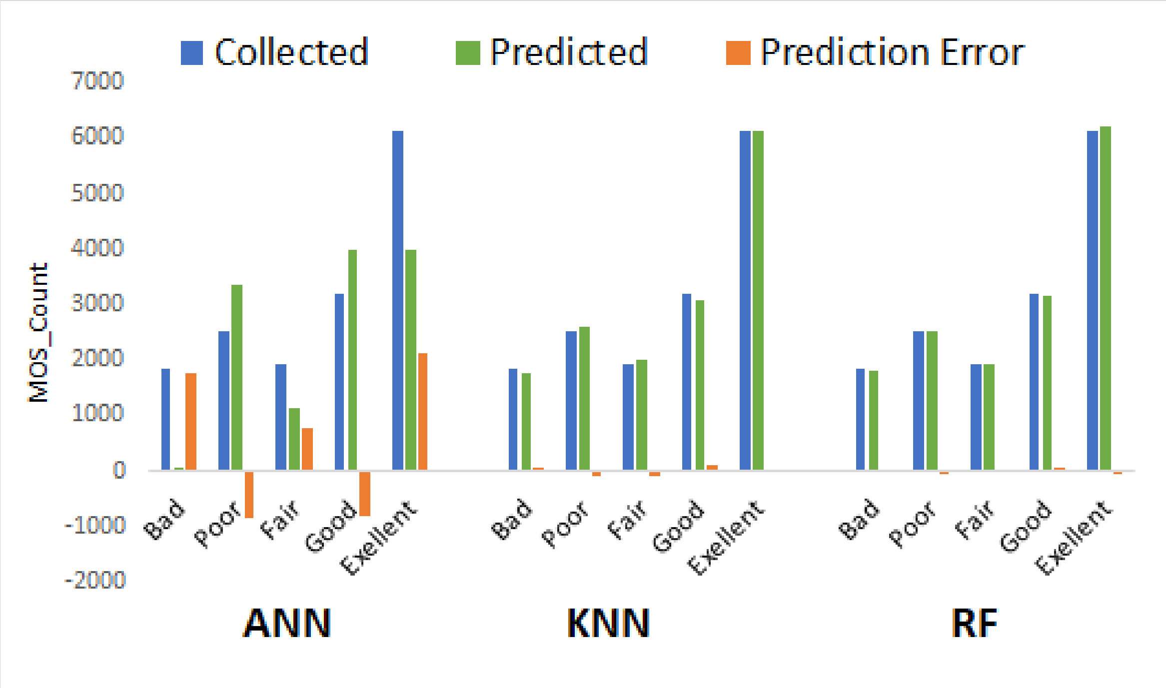


Figure 5.3: Validation Performances of ANN, KNN and RF Estimation Models

This provides an insight into the models' estimation accuracy when implemented in real telecommunications networks. The differences between the collected and predicted values are found by subtracting the estimated MOS counts from the collected MOS counts per each MOS label. As shown in Figure 5.3, if collected MOS counts are greater than models produced MOS counts for each QoE label, prediction errors become positive errors(red bargraphs) and lie above the X-axis. However, if the number of collected MOSs are smaller than models' produced MOS counts, prediction errors become negative and lie below the X-axis. Otherwise, if the models are accurate enough then their prediction errors become zero that is to mean, they have no or have very small validation errors.

Table 5.2: Validation Results of MOS Prediction Models

Validation Method	Test Set Size	MAE			R		
		ANN	KNN	RF	ANN	KNN	RF
Separate Test	15582 Instances	0.45	0.25	0.02	0.76	0.84	0.99

RF is the best model having very small prediction errors for all class labels as shown in Figure 5.3. KNN comes second with estimation errors observed slightly bigger than that of RF for all five MOS labels. However, ANN have larger prediction errors and it is the least accurate model. Mean Absolute Error (MAE) and R are also used to show the models' validation accuracy. ANN, KNN and RF have MAE of 0.45, 0.25 and 0.02 respectively. MAE is chosen because it gives a good insight into the MOS prediction accuracy. As shown in Table 5.2, R values for ANN, KNN and RF are 0.76, 0.84 and 0.99 respectively. This shows that RF has produced almost identical MOS labels to that of the collected MOS labels. So, RF is the most accurate and validated estimation model built in this thesis work.

6

CONCLUSION AND RECOMMENDATION

6.1 *CONCLUSION*

Collecting telecom users' QoE is one of the most important challenges for all TSPs. QoS focused quality management approaches have been used to overcome these challenges. However, this approach has been ineffective since QoE is the cumulative impact of many technical and perceptual factors. Therefore, QoE approaches are more preferable than QoS approaches in improving service quality for telecom services. Here, we propose ML-based QoE estimation solutions for UMTS networks in real-time.

First, non-linear relationships between the collected QoS features and QoE ratings are explored. Correlational results show that RTT, jitter and LR have a negative impact on user QoE. In other words, when measurements of these features increase, user QoE degrades or vice versa. The scatter plot between RTT, jitter and LR against user QoE also follows an exponentially degrading curve. Throughput against QoE, in turn, follows a logarithmically increasing curve, indicating a positive effect on user QoE. Meaning, when throughput increases, user QoE also improves, or vice versa. PCC or R results show that RTT has the highest influence on user QoE with R = -0.87. However, throughput has the least influencing QoS feature with R = +0.25. Similarly, jitter and LR have R values of -0.52 and -0.46 respectively.

ML models training and testing accuracy has been compared. RF produces an overall accuracy of 98.41%. KNN, with an accuracy of 87.49% is significantly better than ANN that has an accuracy of 77.59% as obtained from the ten-fold CV experimentation technique. RF is the best performer model as it is also observed in all performance metrics. In the separate test technique, the performance of RF is excellent with an overall accuracy of 98.63%. KNN scores an overall accuracy of 87.03%. The performance of ANN shows good improvement with an accuracy of 85.30% in the separate test, but ANN is still the least performer.

The proposed models are validated using test sets, but with MOS labels removed now to match the nature of real Internet traffics. Analysis results show that RF almost correctly estimates all

MOS labels having MAE and R values of 0.02 and 0.99 respectively. ANN and KNN produce MAE values of 0.25 and 0.45 respectively, and R values of 0.76 and 0.84 in that order. Generally, all models produce acceptable performances, but RF is the best of all three. The reason is, RF chooses the best model among multiple decision tree models using majority votes and it is less sensitive to data imbalance imbalance problem.

Our QoE estimation models can serve as better solutions in collecting QoE for video streaming services under varying network conditions in real-time. Since there is a paradigm shift from the traditional QoS-based to a more user-centric approach, our solutions have the potential to be good solutions if implemented in the telecom environment.

6.2 RECOMMENDATION

Our ML models could be of great importance to Ethio telecom in estimating user QoE for UMTS video streaming services. Our recommendations to future to Ethio telecom and future researchers are listed as follows.

- The correlational analysis results will help Ethio telecom or any other TSP in identifying which network factors are affecting Internet network performances.
- The proposed models if implemented will be more practical solutions to collect real time user satisfaction from Internet users.
- Our work is limited to YouTube-based video streaming services in UMTS, future works may include other Internet services and technologies like LTE.
- Results show, datasets with good sample representatives are more accurately predicted than the under-sampled datasets. So, more accurate models could be obtained by increasing training datases for our under-sampled datasets.
- Here, as most users are streaming service users, only downstream QoS measurements are used. Future works may include the upstream QoS measurements as additional features so that their solution will predict two-way Internet traffic.
- Since our crowdsourcing tool does not support location information, the spatio-temporal analysis of the collected data is not part of our work. Future studies may consider time and location data analysis.

BIBLIOGRAPHY

[1] Ericsson, "Mobility report: Mobile traffic estimates and forecast: More video on the horizon," Ericsson, Tech. Rep., 2018.

[2] N. K. Thanigaivelan, E. Nigussie, S. Virtanen, and J. Isoaho, "Conceptual security system design for mobile platforms based on human nervous system," in *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, Springer, 2019, pp. 437–446.

[3] M. Kovacs. (Jul. 2017). Global mobile device usagae expected to surpass five billion by 2022. English, itbusiness.ca, [Online]. Available: https://www.itbusiness.ca/news/global-mobile-device-usage-expected-to-surpass-five-billion-by-2022/92682.

[4] O. Issa, F. Speranza, T. H. Falk, *et al.*, "Quality-of-experience perception for video streaming services: Preliminary subjective and objective results," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, 2012, pp. 1–9.

[5] International Telecommunications Union, "New definitions for inclusion in recommendation itu-t p.10/g.100; vocabulary for performance and quality of service," ITU-T, Geneva, Tech. Rep., 2016.

[6] A. A. Laghari, K. U. R. Laghari, M. I. Channa, and T. H. Falk, "Qon: Quality of experience (qoe) framework for network services," in *Proceedings of the 4th international conference on software technology and engineering (ICSTE'12)*, 2012.

[7] P. Dhere, P. Chilveri, R. Vatti, V. Iyer, and K. Jagdale, "Wireless signal strength analysis in a home network," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, IEEE, 2018, pp. 1–5.

[8] I. Khan, N. Crespi, *et al.*, "Quantitative and qualitative assessment of qoe for multimedia services in wireless environment," in *Proceedings of the 4th Workshop on Mobile Video*, ACM, 2012, pp. 7–12.

[9] A. F. Mongi, "A conceptual framework for qoe measurement and management in networked systems," *International Journal of Computer Applications*, vol. 112, no. 8, 2015.

[10] C. W. Chen, P. Chatzimisios, T. Dagiuklas, and L. Atzori, *Multimedia quality of experience (QoE): current status and future requirements.* John Wiley & Sons, 2015.

[11] M. F. M. Hossain, M. Sarkar, and S. H. Ahmed, "Quality of experience for video streaming: A contemporary survey," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, IEEE, 2017, pp. 80–84.

[12] T. Begluk, J. B. Husić, and S. Baraković, "Machine learning-based qoe prediction for video streaming over lte network," in *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE, 2018, pp. 1–5.

[13] A. R. A. Shaban. (Nov. 2017). Ethiopia telecoms monopoly now africa's largest mobile operator. English, [Online]. Available: https://www.africanews.com/2017/11/16/ethiopia-telecoms-monopoly-now-africa-s-largest-mobile-operator/.

[14] E. telecom HomePage. (Jun. 2019). Mobile customers. English, Ethio telecom, [Online]. Available: https://www.ethiotelecom.et.

[15] A. W. Yusuf-Asaju, Z. B. Dahalin, and A. Ta'a, "Mobile network quality of experience using big data analytics approach," in *2017 8th International Conference on Information Technology (ICIT)*, IEEE, 2017, pp. 658–664.

[16] Ethio telecom, "Qos performance special reporting," Service Management Center Section, Tech. Rep., June, 2018.

[17] Ethio telecom and Addis Ababa University, "Customer satisfaction survey-national," Marketing Research and Intellignece Section, Tech. Rep., June, 2018.

[18] C. Meyer and A. Schwager, "Understanding customer experience," *Harvard business review*, vol. 85, no. 2, p. 116, 2007.

[19] T. Abar, A. B. Letaifa, and S. El Asmi, "Machine learning based qoe prediction in sdn networks," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, IEEE, 2017, pp. 1395–1400.

[20] C. Lv, R. Huang, W. Zhuang, X. Wei, and Q. Bao, "Qoe prediction on imbalanced iptv data based on multi-layer neural network," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, IEEE, 2017, pp. 818–823.

[21] V. A. Machado, C. N. Silva, R. S. Oliveira, A. M. Melo, M. Silva, C. R. Francês, J. C. Costa, N. L. Vijaykumar, and C. M. Hirata, "A new proposal to provide estimation of qos and qoe over wimax networks: An approach based on computational intelligence and discrete-event simulation," in *2011 IEEE Third Latin-American Conference on Communications*, IEEE, 2011, pp. 1–6.

[22] A. R. Bisrat, B. B. Haile, E. Mutafungwa, and J. Hämäläinen, "Quality evaluation for indoor mobile data customers in addis ababa business area using data from network management system, walk test, crowdsourcing and subjective survey," in *International Conference on Information and Communication Technology for Development for Africa*, Springer, 2019, pp. 164–175.

[23] N. Singh, "Theoretical and real world of 4g technologies (draft paper),"

[24] Y.-B. Lin, Y.-R. Haung, Y.-K. Chen, and I. Chlamtac, "Mobility management: From gprs to umts," *Wireless Communications and Mobile Computing*, vol. 1, no. 4, pp. 339–359, 2001.

[25] H. Holma and A. Toskala, *WCDMA for umts: hspa evolution and lte*. john Wiley & sons, 2007.

[26] C. Chioariu, "Qos in umts," in *Helsinki University of Technology Seminar on Internetworking*, 2004.

[27] I. Q. Karim and T. I. Yahiya, "A study of quality of service resource allocation schemes for mobile health," *UKH Journal of Science and Engineering*, vol. 1, no. 1, pp. 60–66, 2017.

[28] J. Shaikh, M. Fiedler, and D. Collange, "Quality of experience from user and network perspectives," *annals of telecommunications-annales des telecommunications*, vol. 65, no. 1-2, pp. 47–57, 2010.

[29] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[30] I. El Naqa and M. J. Murphy, "What is machine learning?" In *Machine Learning in Radiation Oncology*, Springer, 2015, pp. 3–11.

[31] G. Williams, *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media, 2011.

[32] K. K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons, 2011.

[33] M. Mohammed, M. B. Khan, and E. B. M. Bashier, *Machine learning: algorithms and applications*. Crc Press, 2016.

[34] Y.-S. Park and S Lek, "Artificial neural networks: Multilayer perceptron for ecological modeling," in *Developments in environmental modelling*, vol. 28, Elsevier, 2016, pp. 123–140.

[35] R. Zaheer and H. Shaziya, "Gpu-based empirical evaluation of activation functions in convolutional neural networks," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, IEEE, 2018, pp. 769–773.

[36] R. Tulaskar, K. S. Pooja Chaudhari, and R. P. Devashree Bhiugade, "Mall map application with indoor positioning and navigation (an android based mobile application)," IJSRCSEIT, vol. 2, 2017, pp. 2456–3307.

[37] P. Cunningham and S. J. Delany, "K-nearest neighbour classifiers," *Multiple Classifier Systems*, vol. 34, no. 8, pp. 1–17, 2007.

[38] A. K. Mishra, S. V. Ramteke, P. Sen, and A. K. Verma, "Random forest tree based approach for blast design in surface mine," *Geotechnical and Geological Engineering*, vol. 36, no. 3, pp. 1647–1664, 2018.

[39] H. Kahsu and T. Ephrem(PhD), "Sim-box fraud detection using data mining techniques: The case of ethio telecom," Dec. 2018, [Online]. Available: http://etd.aau.edu.et/handle/123456789/15238?show=full.

[40] A. A. Laghari, H. He, and M. I. Channa, "Measuring effect of packet reordering on quality of experience (qoe) in video streaming," *3D Research*, vol. 9, no. 3, p. 30, 2018.

[41] P. J. A. Gutierrez, *Packet scheduling and quality of service in HSDPA*. Aalborg Universitetsforlag, 2003.

[42] R Yount, "Populations and sampling: The rationale of sampling, steps in sampling, types of sampling inferential statistics: A look ahead, the case study approach," *Retrieved June*, vol. 12, p. 2014, 2006.

[43] D. R. Stockwell and A. T. Peterson, "Effects of sample size on accuracy of species distribution models," *Ecological modelling*, vol. 148, no. 1, pp. 1–13, 2002.

[44] O. Belmoukadam, T. Spetebroot, and C. Barakat, "Acqua: A user friendly platform for lightweight network monitoring and qoe forecasting," in *The 3rd International Workshop on Quality of Experience Management (QoE-Management 2019)*, 2019.

[45] P. Casas, A. D'Alconzo, F. Wamser, M. Seufert, B. Gardlo, A. Schwind, P. Tran-Gia, and R. Schatz, "Predicting qoe in cellular networks using machine learning and in-smartphone measurements," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6.

[46] K. Bouraqia, E. Sabir, and M. Sadik, "Youtube context-awareness to enhance quality of experience between yesterday, today and tomorrow: Survey," in *2017 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, 2017, pp. 1–7.

[47] H. Beyene, *Final report national assessment: Ethiopia gender equality and the knowledge society*, 2015.

[48] C. Wang, K. Viswanathan, L. Choudur, V. Talwar, W. Satterfield, and K. Schwan, "Statistical techniques for online anomaly detection in data centers," in *12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*, IEEE, 2011, pp. 385–392.

[49] S. Barua, M. M. Islam, and K. Murase, "A novel synthetic minority oversampling technique for imbalanced data set learning," in *International Conference on Neural Information Processing*, Springer, 2011, pp. 735–744.

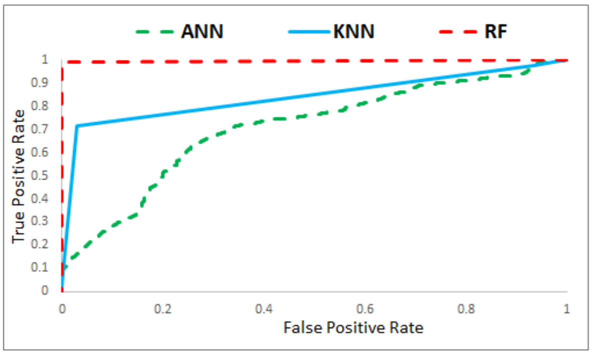
[50] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[51] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012.

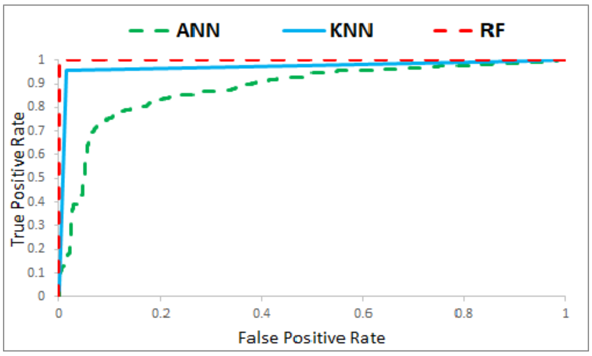
A

APPENDIX

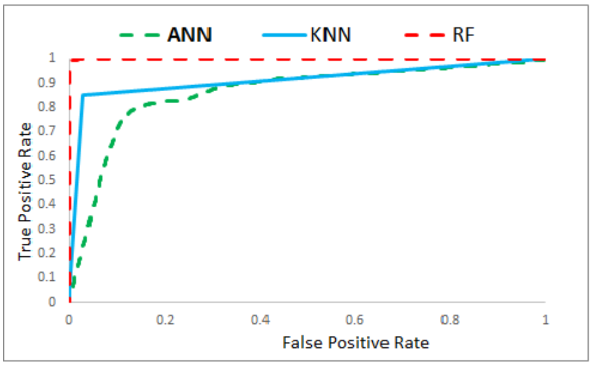
A.1 ROC CURVES



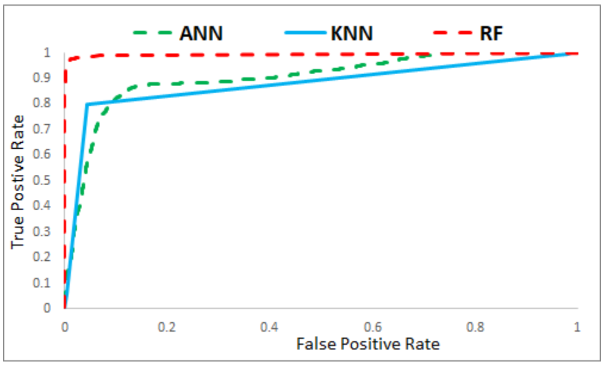
(a) MOS=1



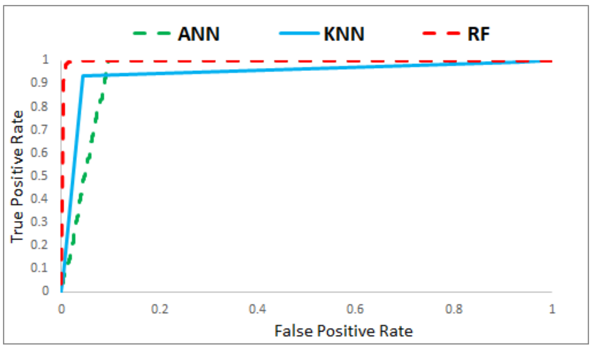
(b) MOS=2



(c) MOS=3



(d) MOS=4



(e) MOS=5

Figure A.1: Models ROC Curve Performances

A.2 SAMPLE DATASET

```
@attribute RTT numeric
@attribute JITTER numeric
@attribute THROUGHPUT numeric
@attribute LOSS_RATE numeric
@attribute MOS {1,2,3,4,5}
@data
0.616655,0.001965,0.591904,0.093431,5
2.331701,0.033861,0.687115,0.001433,1
0.477609,0.002331,1.133478,0.122829,3
```

```
2.376467,0.005007,1.008729,0.00301,3
2.419272,0.02746,0.607455,0,5
2.240937,0.00787,0.403884,0.018566,4
```

Figure A.2: Training Dataset Sample

A.3 SAMPLE SCRIPT IN RSTUDIO

```
> f2rfset<-read.csv(file.choose())
> head(f2rfset)
   RTT JITTER THROUGHPUT LOSS_RATE       MOS
1 0.55 0.0034      1.490      0.00      Good
2 0.35 0.0020      0.738      0.00 Excellent
3 0.15 0.0009      1.979      0.00 Excellent
4 0.82 0.0010      0.925      0.03      Good
5 0.48 0.0047      2.656      0.00       Bad
6 0.13 0.0003      4.782      0.00 Excellent
> x2rfsetnew<- frfset[,1:4]
> x2rfsetnew<- f2rfset[,1:4]
> y2rfsetnew<- f2rfset[,5]
> fr2fcontrolnew <- trainControl(method="repeatedcv", number=10, repeats=10)
> seed <- 7
> merric<- "Accuracy"
> set.seed(seed)
> mtry <- sqrt(ncol(x2rfsetnew))
> tunegrid <- expand.grid(.mtry=mtry)
> fr2fcontrolnew <- trainControl(method="repeatedcv", number=10, repeats=3)
> rf2_defaultnew<- train(MOS~., data=f2rfset, method="rf", metric=metric, tuneGrid=tunegrid, trControl=fr2fcontrolnew)
> print(rf2_defaultnew)
Random Forest

46740 samples
    4 predictor
    5 classes: 'Bad', 'Excellent', 'Fair', 'Good', 'Poor'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 10 times)
Summary of sample sizes: 42066, 42067, 42065, 42066, 42066, 42066, ...
Resampling results:

  Accuracy   Kappa
  0.9841107  0.9786699
```

Figure A.3: Sample RStudio Script for RF Experimentation

A.4 ACQUA APPLICATION USAGE INSTRUCTIONS

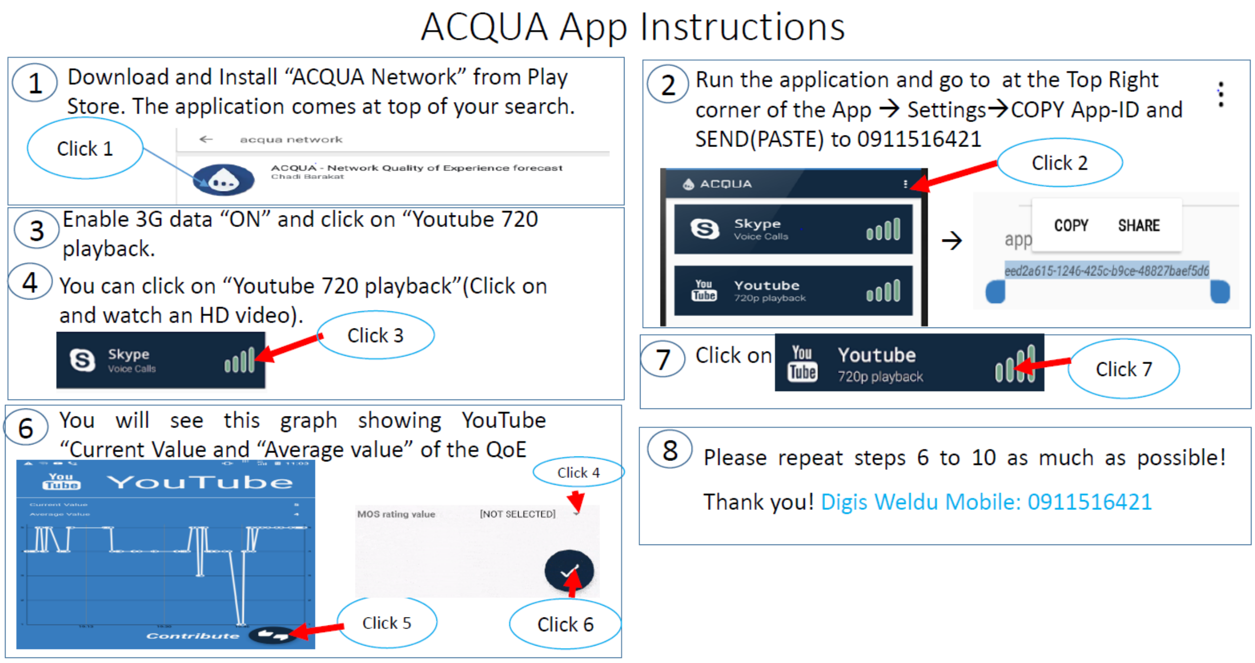


Figure A.4: ACQUA-based Crowd-sourcing Survey Steps