UMTS Network Coverage Hole Detection using Decision Tree Classifier Machine Learning Approach

Geleta Abdissa Wayessa

Advisor: Dr.-Ing. Dereje Hailemariam

Telecommunication Engineering Graduate Program
School of Electrical and Computer Engineering
Addis Ababa Institute of Technology
Addis Ababa University

A Thesis Submitted to School of Electrical and Computer Engineering, in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Telecommunications Engineering

February 2020

Abstract

Due to various innovative mobile services and applications, traffic is constantly increasing in size and complexity globally and as well as locally in Ethiopia. To fulfill these requirements in both quality and quantity, a wide range of radio frequency signal coverage areas are required. One means of satisfying this requirement is proper planning and devising proper network management during operational phase for network coverage hole detection for optimization of uncovered area. Measurement collection is a primary step towards analyzing and optimizing the performance of a telecommunication service. In this sense, this work aims to present a solution that contributes to reduce costs and time in network monitoring by exploiting user equipment Measurement Report (MR) data via the Minimization of Drive Tests (MDT) functionality.

An automatic coverage hole detection based on classification techniques, which is a Decision Tree (DT) classifier-based approach is used for rule induction to identify different scenarios of coverage holes and their respective areas for better service delivery purposes. The main idea is to jointly observe signal strength and signal quality for effective coverage-hole detection. It uses a new approach to classify four coverage scenarios such as "good coverage and good quality", "good coverage but poor quality", "poor coverage but good quality", and "poor coverage and poor quality" in Universal Mobile Telecommunications System (UMTS) network considering the last three coverage classes as coverage -hole with different severity levels.

The result showed that the applied model accuracy was 99.98%, and also the proposed approach could classify the target classes and allows the visualization of network performance in terms of signal strength and quality associated with a location. All four coverage scenarios were visibly observed and the results are almost uniform with validation results found from the driving test (with about 7dB and 1dB difference of RSCP and Ec/No respectively considering the cumulative distribution function value of 18%). 77% of coverage areas were classified as good coverage condition.

Keywords-UMTS, Coverage hole, MDT, MR, DT.

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

This is to certify that the thesis prepared by Geleta Abdissa Wayessa, entitled *UMTS Network Coverage Hole Detection using Decision Tree Classifier Machine Learning Approach* and submitted in partial fulfillment of the requirements for the degree of Master of Science in Telecommunication Engineering complies with the regulations of the University and meets the accepted standards concerning originality and quality.

Signed by the Examining Committee:

_____	_____
Chairperson	*Signature*
_____	_____
Examiner	*Signature*
_____	_____
Examiner	*Signature*
Dr.-Ing. Dereje Hailemariam	_____
Advisor	*Signature*

_____

Dean, School of Electrical and Computer Engineering

Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for a degree in this or any other University, and all sources of materials used for the thesis have been fully acknowledged.

_____

Signature

Geleta Abdissa Wayessa
February 2020

This thesis has been submitted for examination with my approval as a University advisor.

_____

Signature

Dr.-Ing. Dereje Hailemariam
February 2020

Acknowledgements

First and foremost, I am grateful to the almighty God, for giving me this opportunity and seeing me through all the challenges in my academic work.

My special gratitude goes to my advisor Dr.-Ing Dereje Hailemariam for his excellent guidance and valuable comments up to the submission of my thesis. I would also like to thank my evaluators Beneyam Haile (PhD) and Ephrem Teshale (PhD) for their valuable feedback during the thesis progress presentations and also thank Mr. Yonas Yehualaeshet for his dedicated support and guidance.

Next, special thanks go to my wife Gadise Regassa, my children Nani, Moti and Amen, my parents, brothers, and sisters. Without their help, encouragement and effort in all moments of my life nothing of this would be possible. I appreciate their support, patience, and encouragement they provided me throughout my studies.

I also wish to appreciate my colleagues and ethio telecom staffs for their support and encouragement throughout the course work. I also thank the member of Engineering department's staff for their support in providing data.

Last but not the least, a special acknowledgment extends to ethio telecom for giving me the opportunity to do my master study and be responsible for my full sponsorship.

Lists of Acronyms

2G	2nd Generations
3G	3rd Generations
4G	4th Generations
3GPP	3rd Generation Partnership Project
AI	Artificial Intelligence
CART	Classification and Regression Tree
CGI	Cell Global Identification
CHR	Call History Recording
CM	Configuration Management
CN	Core Network
CPICH	Common Pilot Channel
CS	Circuit Switched
DT	Decision Tree
Ec/No	Energy per chip to the total received power density
EDGE	Enhanced Data Rate for GSM Evolution
FN	False Negatives
FP	False Positives
GPRS	General Packet Radio Services
GSM	Global System for Mobile communications
GNSS	Global Navigation Satellite System
HLR	Home Location Register
HSPA	High-Speed Packet Access

IG	Information Gain
IMSI	International Mobile Subscriber Identity
I-RAT	Inter RAT
KPI	Key Performance Indicator
LTE	Long-Term Evolution
MCS	Mobile Crowdsourcing
MDT	Minimization of Drive Tests
ML	Machine Learning
MNO	Mobile Network Operator
MR	Measurement Report
MSC	Mobile Switching Center
NE	Network Element
NGMN	Next Generation Mobile Networks
OAM	Operation and Maintenance
PM	Performance Management
PS	Packet Switched
QoS	Quality of Service
RAN	Radio Access Networks
RAT	Radio Access Technology
REM	Radio Environment Maps
RF	Radio Frequency
RLF	Radio Link Failure
RNC	Radio Network Controller
RRC	Radio Resource Controller

ROC	Receiver Operating Characteristics
RSCP	Received Signal Code Power
RSSI	Received Signal Strength Indicator
SGSN	Serving GPRS Support Node
SON	Self-Organizing Network
TCE	Trace Collection Entity
TN	True Negatives
TP	True Positives
UE	User Equipment
UMTS	Universal Mobile Telecommunications System
USIM	UMTS Subscriber Identity Module
UTRAN	UMTS Terrestrial Radio Access Network
VLR	Visit Location Register
WCDMA	Wideband Code Division Multiple Access

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter provides background of this thesis and describes statement of the problem, objective, scope and limitation parts of the thesis. Moreover, it briefly describes reviewed literature that are related to the study, the methodology used and the contribution of the thesis. Finally, the thesis structure is outlined.

1.1 Background

Since the appearance of cellular networks, traffic is constantly increasing due to various innovative mobile services and applications [6]. The public interest has become higher as operators offer new services. Due to these and other reasons, the number of unique mobile users globally reaches about 5.1 billion in 2018 and will reach 5.8 billion by 2025 [7]. In order to fulfill subscribers' requirements in capacity, coverage and service quality a wide range of Radio Frequency (RF) signal coverage areas are required in terms of strength and quality. One means of satisfying this requirement is proper planning and deploying sufficient Network Element (NE)s and devising proper network management [8].

In Ethiopia, about two decades have passed since the first official Global System for Mobile communications (GSM) service launched for the first time in Addis Ababa in 1999 [6]. Later, for better mobile data services provisioning, ethio telecom, the sole service provider in Ethiopia, implemented different technologies like an enhancement to General Packet Radio Services (GPRS) and Enhanced Data Rate for GSM Evolution (EDGE) and also deployed 3rd Generations (3G) and 4th Generations (4G). Currently, ethio telecom is running multi Radio Access Technology (RAT) such as GSM, Universal

Mobile Telecommunications System (UMTS), and Long-Term Evolution (LTE) in parallel and now reached 42.9 million mobile subscribers in Ethiopia in the first quarter of 2019/2020 [9]. To provide good network coverage and quality of service, these mobile communication networks (GSM, UMTS, and LTE) require regular monitoring and optimization.

Ensuring RF signal coverage and quality within the network starts with a good network plan. A good network plan has to address the coverage, capacity and quality requirements of the area considered and also expected to provide a flexible network that allows network expansion without a major change of the existing sites. However, while deploying a RAT, coverage planning is a complex task for operators since they need to consider multiple and correlated network parameters as well as environmental conditions that are out of their control [10]. Hence, completely avoiding the existence of coverage holes in cellular networks is almost impossible. Not only due to planning case, but a coverage hole could also happen by different factors like obstruction, traffic overload, and network faults. Therefore, coverage hole detection and optimization processes are usually required during the operational phase [8]. A coverage hole has been defined by different scholars based on the consideration of the signal strength and signal quality [4, 11–14]. In this thesis, coverage hole is defined as, the area where the received signal strength or/and quality level of the serving and neighbor cell is below the levels required to maintain basic services.

Currently, ethio-telecom is managing different RATs in parallel, which contributes extra complexity of the network management. Hence, it is necessary to develop a new approach in which mobile system coverage detection and performance management become more effective and automated. In this context, the 3rd Generation Partnership Project (3GPP) introduced the Self-Organizing Network (SON) which is used for mobile network automation and minimization of human intervention in the cellular/wireless network management. This concept has been inspired by a set of requirements defined by the operators' alliance Next Generation Mobile Networks (NGMN) and has been introduced in 3GPP release 8 and expanding across subsequent releases [15].

As J. Moysen et al. mentioned in [16] the main objective of SON are classified into three main points:

1. to bring intelligence and autonomous adaptability into cellular networks;
2. to reduce Capital Expenditure (CAPEX) and Operational Expenditure (OPEX);

3. to enhance network performances in terms of network capacity, coverage, offered service/experience, etc.

3GPP also started on the standardization of cellular networks on the optimization of drive test since release 9 under the name of Minimization of Drive Tests (MDT) [4, 11]. Since release 10, a feature called MDT has been included in the standard, both for 3G (including also High-Speed Packet Access (HSPA)) and LTE [15]. The key idea of the MDT is that the network operator can request the User Equipment (UE)s to perform and report specific radio and Quality of Service (QoS) measurements associated with the UE location, which are the powerful input parameters for troubleshooting and optimization efficiency.

Both SON and MDT address the same objectives, which are to reduce operational efforts, increase network performance, quality and, at the same time, decrease the maintenance costs and time. Both techniques are very promising for network optimization and can be used independently one from the other. The main differences between SON and MDT are: SON is aiming at instant/automated reaction on short to middle-term network issues, while MDT is more about collecting measurements for further analysis and processing (either manually or automated) [15]. The use case applicability of both systems are: SON includes self-configuration, self-optimizations, and self-healing, while MDT mainly focuses on optimization.

Key features of MDT are that a mobile device reports its location information along with performance measurements using to Trace Collection Entity (TCE) via Radio Network Controller (RNC), thereby allowing to have a much more fine-grain view of a cell's performance [3]. It addresses the issue that often drive tests to have to be executed to monitor and assess mobile network performance in an efficient way. Hence, the motivation of this thesis is to detect coverage hole of UMTS network based on UEs Measurement Report (MR) data generated by MDT functionality.

1.2 Statement of the Problem

Providing a good quality of service is one of the major concerns of telecommunications operators. To provide sufficient coverage and QoS for subscribers in indoor and outdoor environments, mobile operators need to carry out various radio coverage measurements. Network coverage is an important and high priority target that has to be achieved by operators in the planning and optimization process to provide the required QoS. Hence,

to optimize the network with coverage hole problem, efficiently detecting the coverage hole area is the task to be performed by capturing required radio measurements.

Operators, including ethio telecom, are using the traditional method of radio measurement data collection for network coverage and quality evaluation which is called drive test. However, there are some challenges and limitations in traditional drive testing that could be improved. Firstly, it is a resource-consuming task requiring a lot of time, specialized equipment, and the involvement of highly qualified engineers. Secondly, it is difficult to capture the whole coverage data from every geographical location by using manual drive testing, since most of the UE generated traffic comes from indoor locations, while drive testing is often limited to roads. Thus, the drive test method cannot offer a complete and reliable picture of the network coverage situation at a reasonable cost and time.

So far, some researches have been conducted on network coverage hole detection by implementing a data-mining approach and other techniques. One is based on the analysis of the extended Radio Link Failure (RLF) triggering report (event-triggered report) without needing field measurements [17]. However, in this method, only the most severe cases, which occur after the link failures can be detected, but other situations like signal degradation before failure and other information in the periodic reports cannot be identified. Another study also conducted on coverage hole detection by using Inter Inter RAT (I-RAT) handover data. I-RAT handover is the handover process between different technologies (e.g., from 2nd Generations (2G) to 3G). In this study, only the heterogeneous deployment scenario has considered, which cannot be a solution for the case of homogeneous scenarios.

1.3 Objectives

1.3.1 General Objective

The main objective of this thesis is to detect coverage hole in UMTS network by using Decision Tree (DT) classifier supervised Machine Learning (ML) algorithm. Received Signal Code Power (RSCP) and Energy per chip to the total received power density (Ec/No) data collected from UEs via the Nastar tool serves an input for the classifier.

1.3.2 Specific Objectives

The specific objectives to be addressed in this thesis are:

- Review related literature on coverage hole, coverage-hole detection/classification methods.
- To identify and collect required key measuring parameters for coverage hole evaluation.
- Classify the collected parameters using the selected algorithms as per their pre-defined classes.
- Analyze the performance of the implemented classification algorithms.
- Locate the detected area of coverage-hole on a Google map (by considering the relative density of poor RF).
- Doing driving test and verify the detected coverage hole.
- Discuss the results and draw recommendations based on the findings.

1.4 Scope of the Research

This thesis addresses one of the use cases of MDT for UEs MR data collection which is used for coverage hole detection. The scope of the thesis is to investigate the coverage hole detection method based on MR data of the UMTS mobile network using a DT algorithm. The thesis is limited to coverage hole detection only rather than diagnosis the root cause of the problem. The captured data has considered only busy hours traffic time(condition) i.e. the data generated within certain periods of a day is considered. Moreover, some parts of Addis Ababa UMTS Node-B sites at the selected area are considered for analysis. However, the author is confident that the selected area will represent and can be replicated to the actual whole ethio telecom network with reasonable accuracy, and this factor does not significantly limit the applicability of the research.

1.5 Contributions

This thesis can contribute in reducing extra cost and time of the operator like ethio telecom by providing instant data collection and processing in a more flexible and cheaper way for network performance assessment. It can also contribute to network planners in providing sufficient information about the RF signal status of the specific geographic area of the active network for optimization triggering action. This is due to the capability of the model in considering the joint effect of RSCP and Ec/No at the same instance and capturing all possible locations covered by the network including indoor environments where drive testing cannot be addressed and the algorithm used in exploring accurate information.

1.6 Literature Review

A number of researches have been conducted in implementing ML algorithms and other techniques for network coverage hole detection. Recently, automatic coverage-hole detection method through mobile data analysis got attention in operators' real scenarios due to its efficiency. This data can be generated from UE or/and mobile operator NEs and captured properly for further analysis. Among these researches, related works which focus on coverage detection and optimization use cases are mentioned in this section.

O. F. Celebi et al. [18] and A. Gómez et al. [12] conducted the study on the use of big data for coverage hole detection. Both studies used I-RAT handover information to pinpoint the coverage-hole problem in the network considering the coexisted technologies. The study in [18] analyzed the Base Station Subsystem Application Part (BSSAP) messages from A-interface in a Hadoop platform to identify handovers from 3G to 2G. The simulation results show that the identified 3G coverage holes are consistent with the drive test results. The strength of this study is that the parameter used is easily handled and applicable but it considered only the scenario of different technologies which means it cannot be a solution for the case of the same technology deployed scenario. Whereas, [12] used mobile trace data with full information of geo-location and timestamp as an input. The objectives were to propose a method on how to detect coverage hole in LTE network, to implement root cause analysis and troubleshooting. They used an application of IF-THEN rules as a tool to analyze the data and finally detect the coverage hole of LTE network and as well as they could categorize the

severity level of the holes into three classes. The study used individual user-based trace data on both serving and targeted cells that make better detection accuracy however, it might be somewhat complex for huge data processing.

Another literature by Galindo et al. and V. Dalakas in [19, 20] conducted the study to detect the coverage hole by using the data collected mainly from the UEs remotely and used Bayesian interpolation to forecast the unknown location information from the collected sample data. The authors in [21] followed the cognitive-tool based approach that provides location awareness which is Radio Environment Maps (REM)s, whereas, the study by Lin et al. in [20] used graph-theory based mobile network insight analysis framework to detect the coverage hole from both network data and user behavior data. The detection accuracy of the study depends on the sample data size used for the interpolation case and the reference locations are limited to site-level like site ID, and location (latitude, longitude) for graph-theory based and so that it has an impact on accuracy.

Other studies were conducted on the detection of coverage hole and quality of service estimation by the name of MDT. These studies captured the data by tracking mobile devices and their real-time information about where, when, and what information about mobile users to analyze the coverage status of the network. Most of the approaches use supervised and unsupervised learning techniques to provide different solutions for this use case. These approaches are observed in [22–24], where the authors address the QoS estimation by selecting different Key Performance Indicator (KPI)s and correlating them with common nodes measurements, to establish whether a UE is satisfied with the received QoS or not. The studies are focused on QoS verification use case of MDT and used simulation environment. In another study [17], extended RLF reports are used to classify network problems into three categories that are downlink coverage, interference and handover problems using DT classifiers. Based on the results, the coverage, interference and handover problems can be differentiated by using the RLF reports containing RF measurements from both the serving and neighboring cells. The study tried to diagnosis the causes of the failures which minimize the operator challenge in solving the problem. However, only the event-triggered data is considered which may not provide full information on the network.

Additionally, in [25], the authors focus on multi-layer heterogeneous networks. They present an approach, based on regression models, which allows predicting QoS in heterogeneous networks for UEs, independently of the physical location of the UE. This work is extended in [26] by taking into account the most Principal Component

Analysis (PCA) on the input features and promoting solutions in which only a small number of input features capture most of the variance, the number of random variables has reduced. Based on previous results, in [27] the same authors defined a methodology to build a tool for smart and efficient network planning, based on QoS prediction derived by proper data analysis of UE measurements in the network.

Generally, the studies tried to follow efficient data collection methods; implemented ML algorithms for adaptive learning and parameters used are easily handled and applicable for coverage hole detection. However, the methods that they followed lack the generality, in detecting the coverage-hole for the heterogeneous network; event-triggered data was used, which cannot provide sufficient information about coverage hole; complex for huge data processing and the detection accuracy of the study depends on the sample data size used.

In this thesis, the measurements are captured using MDT framework from ethio telecom network by the support of the Nastar tool to investigate RF signal strength and quality. This approach is of special importance to network service providers in identifying or locating a particular area experiencing network problems due to coverage-hole prior to the link failure. In addition to these, the method explores four types of coverage scenarios by the support of ML that can provide extra information about the specific area for the experts and so that the diagnosis of the problem will be simplified.

1.7 Methodology

Since the goal of this thesis is to detect the areas having poor RF conditions (coverage-hole) in different scenarios such as: “Good Coverage (good RSCP power) and Poor Quality (poor Ec/No)”, “Poor Coverage (poor RSCP power) and Good Quality (good Ec/No)”, “Poor Coverage (poor RSCP power) and Poor Quality (Poor Ec/No)”, the main task is implementing classification techniques and identifying the location of the UEs sending Poor RF signals. Supervised ML techniques are used by considering the previously collected measurement report data from Nastar system as a training data and test data-set. DT algorithm is used for classification in this work. The tools like python, MS excel, and MapInfo are used.

Generally, the methodology used in this research include:

- An extensive literature review has been conducted related to the work to understand coverage hole and hole detection, ML algorithms, and network parameters used for coverage hole detection.
- The required data for the study has been captured using the Nastar tool and preprocessed using MS Excel and python platform.
- Target classes have been defined based on currently in-use thresholds in ethio telecom on the signal strength and quality and classification of the collected data have been done using the DT algorithm.
- Predicted classes have been displayed on Google map and the more densely visible locations considered as detected areas.
- Finally, the results have been discussed using Google map visualization, plots, graphs, and tables and published in the form of a final thesis paper.

1.8 Thesis Layout

The thesis has organized into six chapters, including this one. Chapter 2 reviews a state-of-the-art, related to this work. It constituted an overview and structures of UMTS radio communications, UMTS functional units, and interfaces. UMTS coverage hole and coverage hole detection techniques also included in this section. It also discusses a UMTS network performance data collection methods such as performance management, configuration management, drive tests, crowdsourcing, and network traces. Chapter 3 explores data mining techniques, specifically supervised learning. It also contains an overview, structure, attribute selection criteria, algorithm learning, and evaluation of the DT algorithm. Chapter 4 presents the experimentation that comprises area selection, threshold definition, coverage scenario definition, and system processes. Chapter 5 contains result discussion and validation. Finally, in Chapter 6 the work main conclusions and feature works are drawn.

Chapter 2

UMTS Overview

UMTS is a 3G mobile technology based on Wideband Code Division Multiple Access (WCDMA) radio technology with better spectrum efficiency and capacity than GSM. It supports data transfer rates through its enhancement HSPA. HSPA accounts for the majority of worldwide broadband networks today [28].

An overview of the UMTS system architecture and the functionality of its main nodes are presented in this chapter. The interfaces between the system nodes and radio interface channel definitions are also explained. Finally, about network coverage hole, coverage hole types, detection of coverage hole and performance data collection methods are discussed as well.

2.1 UMTS Structure

UMTS is a 3G cellular telecommunication system designed to cope with the growing demand for mobile and internet applications with the required quality of service parameters. The UMTS system consists of three interfacing domains; UE, UMTS Terrestrial Radio Access Network (UTRAN) and Core Network (CN) [29, 30]. The main function of the CN is to provide switching, routing, and transit for user traffic to connect them to the external network and it also contains the databases and network management functions. The UTRAN performs all the functions related to wireless communication and the UE interfaces with UTRAN through the air interface standard. The system architecture of UMTS system in Figure 2.1 below illustrates the network elements and the logical interfaces between them.

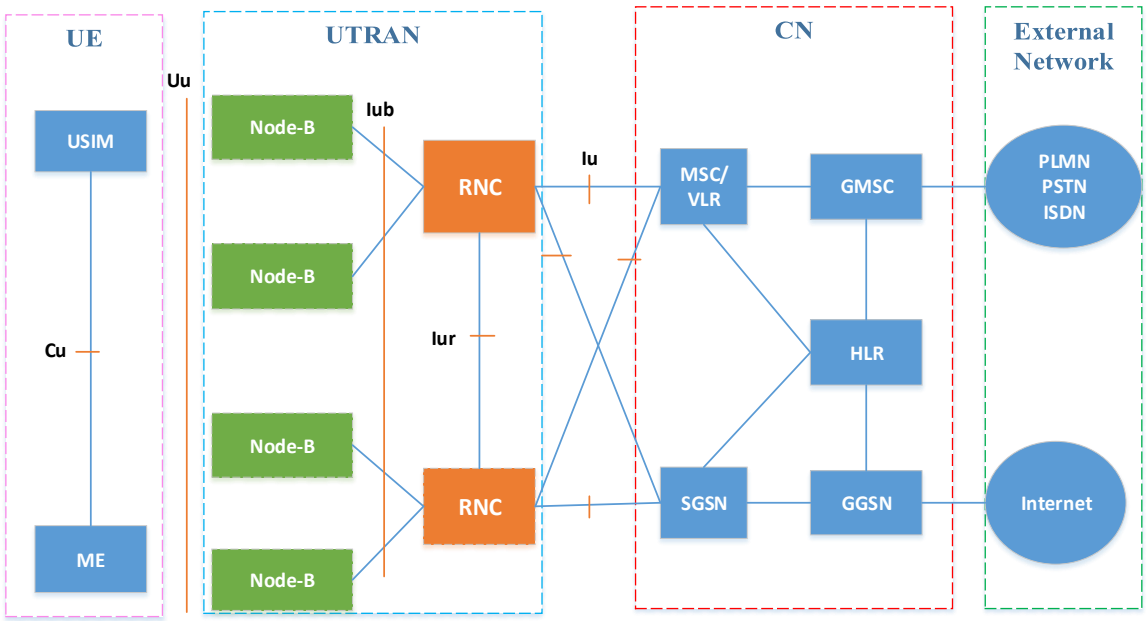


Fig. 2.1 UMTS system architecture [1]

2.2 UMTS Functional Units

2.2.1 User Equipment

UE is a wireless equipment/terminal used by the user to access UMTS services via the Uu interface. It consists of the Mobile Equipment (ME) and an intelligent card UMTS Subscriber Identity Module (USIM), which records the user ID, performs the authentication algorithm, and stores information such as authentication information and keys.

2.2.2 Node-B

Node-B is the name given to the 3G base stations and it is the logical node responsible for radio transmission/reception in one or more cells to/from UE. The main function of a Node-B is to establish the physical implementation of the Uu interface and the Iub interface. Other functions of the Node-B include spreading, scrambling, modulation, channel coding, power control, interleaving, synchronization, and measurement reporting [31]. It controls the data flow between the Iub and the Uu interfaces, terminates the physical layer, extracts the Media Access Control(MAC) protocol data units, and

transports them across the Iub interface to the RNC. It also participates in radio resource management.

2.2.3 Radio Network Controller

The RNC is the central unit in 3G Radio Access Networks (RAN). It is a governing element in the UMTS radio access network and is used for controlling the Node-Bs that are connected to it [1]. It is also responsible for controlling the use of all 3G radio resources by performing Radio Resource Management (RRM) procedures, handover decision and transmission scheduling [30, 32]. It also plays an important role in configuration management because the radio-related parameters for the whole Radio Network Subsystem (RNS) are stored in RNC. For performance management, the RNC updates performance counters, which are later used to calculate the KPIs for RAN. RNC is also responsible for fault management by keeping track of the alarms in any Node-B controlled by that particular RNC and also serves as the intermediate node which connects CN to RAN.

2.2.4 Home Location Register

Home Location Register (HLR) is a database located at the local system of the user, used to store the master copy of the subscriber service features [30]. Such features include information on the services allowed, roaming areas and information of value-added services. This database is created when a new subscriber registers to the system for network access and is maintained throughout the service period. To find a route to the UE for the incoming service, the HLR also stores the location information of the UE.

2.2.5 Mobile Switching Center/ Visit Location Register

Mobile Switching Center (MSC) / Visit Location Register (VLR) is a switching center and a database, providing Circuit Switched (CS) services for the UE at its current location. The MSC processes circuit-switched services, while the VLR stores a copy of the service feature description of the roaming subscribers, and more accurate information on the location of the UE in the service system [30]. The network part connected through the MSC/VLR is usually referred to as the CS domain.

2.2.6 Gateway Mobile-services Switching Center

Gateway Mobile-services Switching Center (GMSC) is the switching equipment at the connection between the UMTS network and the external circuit domain network. All incoming/outgoing CS connections go through the GMSC.

2.2.7 Serving GPRS Support Node

Serving GPRS Support Node (SGSN) function is similar to that of the MSC/VLR, except that it is used for Packet Switched (PS) services. The network part connected through the SGSN is referred to as the PS domain.

2.2.8 Gateway GPRS Support Node

Gateway GPRS Support Node (GGSN) function is similar to Gateway Mobile-service Switching Center (GMSC), but it is used to deliver PS services.

2.2.9 External Networks

External networks fall into two groups such as CS network and PS network. CS network provides circuit-switched connections, such as the existing telephone services like Integrated Service Digital Network (ISDN) and Public Switch Telephone Network (PSTN) whereas PS network provides packet-switched connections. Internet is an example of the PS network.

2.3 UMTS NE Interfaces

Interfaces are the logical connections between the UMTS NEs. All the interfaces are open, which allows an operator to build its UTRAN and CN by using equipment of different manufacturers, thus reducing the cost for network construction. The main open interfaces defined in the UMTS network are [30]:

- Uu interface: Serves as the air interface of the WCDMA system to connect a UE to a Node B.
- Iu interface: Connects UTRAN and CN.

- Iur interface: This interface is used to connect two RNCs. It allows soft handover between the RNC equipment of different manufacturers as an open interface.
- Iub interface: Connects Node-B and RNC.
- Cu Interface: The interface between USIM smartcard and mobile equipment.

2.4 UMTS Network Coverage Hole and Coverage Hole Detection Techniques

2.4.1 Introduction

In wireless mobile networks QoS changes dynamically due to large variety of factors. Because of that Mobile Network Operator (MNO)s monitor and optimize their network regularly in order to provide a good network coverage and quality of service. There could be different reasons that cause coverage holes such as [4, 33]:

- New building construction and other obstruction which shadows a certain area;
- Planning problems such as non-optimum cell design, wrong site implementation, wrong parameter planning, wrong or missing neighbor relations;
- Overloading of cells, and
- Network faults.

2.4.2 Network Coverage Hole

Coverage hole is the area where the received signal strength or/and quality level of the serving and neighbor cell is below the levels required to maintain basic services. The presence of coverage holes in mobile networks is a common problem for mobile operators. It is an aspect a user can easily observe and which mainly influences the user-experience. Completely avoiding the existence of coverage holes in cellular networks during the planning phase is almost impossible and therefore, coverage optimization processes are usually required during the operational phase [19]. Without coverage provisioning, it is difficult to talk about service, or quality of service provisioning. Therefore, cellular

coverage-hole detection and enhancement is one of the basic tasks that MNOs have to give attention.

In terrains with uneven morphology, the signal coverage and quality are affected by a lot of factors such as human structures, non-uniform human/vehicular traffic, hills, vegetation and the like. Other key limiting factors are distance from the cell, intercellular/intracellular interference and random background noise in the network environment [34]. Within this context, it highly needs a regular systematic assessment of deployed and operational mobile communication networks. This will provide up-to-date information for mobile operators to support the network engineering parameter tuning process and guarantee end-user satisfactions.

To detect and improve such problems and the others, radio measurements are needed. Some of the key indicators or metrics for performance evaluation at the system level and perceived at the UE are RSCP and Ec/No [34]. These measurements can be done with developed equipment directly at the NE (base station) or by drive tests.

2.4.3 Classification of Coverage Holes

As described in Section 2.4.1, coverage hole can happen due to different factors. However, all effects are mostly reflected in two aspects which are signal strength (RSCP) and quality (Ec/No). Despite receiving a high-power level, communication can be poor because of the interference effect which leads to communication degradation and so that the transmission rates could be reduced [35, 36]. Interference is typically measured by the Ec/No of the Common Pilot Channel (CPICH), that infers, how clear is the signal received. This means that the CPICH power level does not guarantee the coverage of the network unless the quality of the network is fulfilled.

In 3G networks using WCDMA, mobile terminals receive signals from multiple node-Bs. On the contrary, in a cellular system where all the air interface connections operate on the same carrier, the number of simultaneous users directly influences the receivers' noise floors. During times of peak use, distant users/customers may experience lower signal than normal as the interference increase. The increased interference causes a need for additional power in order to maintain the link quality, which in turn effects additional capacity and coverage degradation [37]. Therefore, it is possible that the mobile terminal cannot start logging network because several pilot signals are received with high reception, but none of them is sufficiently dominant so that the mobile can choose. Not only this, other reason can be overshooting cells, that means the presence

of unwanted signals in a region. Hence, coverage-hole can be reflected in different forms like poor signal strength, overshooting, and the area without dominant cell [38].

Poor signal strength: It refers to the coverage of some base stations where the pilot (or reference) signal power is in between the lowest network access threshold and the lowest value required for assigning full coverage.

Overshooting: Overshoot occurs when coverage of a cell reaches far beyond what is planned. It can occur as an "island" of coverage in the interior of another cell, which may not be a direct neighbor. Reasons for overshoot may be reflections in buildings or across open water, lakes etc. UEs in this area may suffer call drops or high interference. Possible actions to improve the situation include changing the coverage of certain cells.

Pilot Pollution: It means that too many pilots are received on a point, in which there has not dominant pilot. The possible reasons for pilot pollution may be the unreasonable site distribution, too high the location of the base station or too high antenna, unreasonable directional angle of the antenna, and the influence caused by the back radiation of the antenna, an unreasonable setting of pilot power and the influence of around environments. Where, the influence of around environment can be the obstacle of tall building or mountain, signal transmission extended along streets or water area, or signal reflection caused by the glass curtain wall on the tall building.

2.4.4 Network Coverage Hole Detection

Network coverage-hole detection is the process of identifying the area where the signal strength or/and quality level is below the required value for the given specified service standards by using different techniques and tools. It is a special case of signal degradation on the specific area due to different reasons like cell-overload, malfunctioning base station, blockage or planning problem. One approach for detecting a coverage-hole is to monitor the signal strength and quality of the network using different techniques. MNOs are following different techniques to manage their network coverage status. However, the techniques and tools operators use can affect the business due to inefficiency in time and cost.

As it was mentioned in Section 1.6, historical data analysis by using modern data mining techniques is recently in use to improve the coverage hole detection efficiency. Most of the approaches use supervised and unsupervised learning techniques to provide different solutions for this use case. The detection is based on employing the knowledge mining method to find hidden patterns from the UEs MR databases. Classification is

one of the data mining techniques used widely. There are many algorithms used for classification purposes such as DTs, neural networks, Bayesian networks, and many others. In this thesis, a supervised learning technique is used to mine the knowledge from the data using DT algorithm. It is carried out by conducting classification on the gathered MR reports which are a key performance indicators for revealing the coverage hole. The analyzed MR reports contained RSCP and Ec/No radio measurements from the serving cell.

The detection is based on the joint processing of the RSCP and Ec/No measurements. The DT-based approach is used to classify target classes (good coverage and good quality, good coverage but poor quality, poor coverage but good quality, poor coverage and poor quality). The reason why these joint processing needed is to visualize the coverage scenario of the geographic area easily and so that it will improve the quick diagnosis of the problem. For example, poor RSCP but good Ec/No scenario can be reflected due to less interference from other sites and the site itself. Generally, the relation between Ec/No, RSCP and Received Signal Strength Indicator (RSSI) [39, 40] can specify more about the coverage scenarios as shown in Equation (2.1) below.

$$E_c/N_o = \frac{RSCP}{RSSI}(dB) \tag{2.1}$$

From this equation, we can observe that RSSI and everything that affect it have a very big impact over E_c/N_o. In other words, we may have good RSCP, but if RSSI is bad (because of pilot pollution, overshooting, very high speed, external interference, huge traffic load), then Ec/No will be negatively affected. On the contrary, when we have relatively low RSCP but the RSSI is good enough, then Ec/No will be good.

2.5 UMTS Network Performance Data Collection

Because of the rapid increase in the use of wireless devices and continued expansion of cellular networks, effective control of the cellular cells' coverage has become more important to ensure QoS provisioning. The level of network coverage provided to various parts of a region under consideration has to be measured on a regular basis. This regular check is to determine any coverage holes produced due to construction, planning, network failure and other factors. Measurement collection is a primary step towards analyzing and optimizing the performance of a telecommunication service. So, it is important to utilize network-based or user-based measured information.

In order to know if the network performance is within recommended parameters, a set of methods are used to collect and analyze data from networks side and user side to understand whether the required QoS is delivered to the end-user or not. Some methods by which the network performance data can be gathered and analyzed or calculated are Performance Management (PM), Configuration Management (CM), drive tests, crowdsourcing, trace data or customer complaints [12, 41, 42]. The granularity in how the measurements are collected depends on the use case and efficiency.

2.5.1 Performance Management

PM is an information, consisting of counters reflecting the number of times that some event (e.g., soft handover, handover failure, call drop and others) has happened in a NE during a certain period [41]. It comprises evaluation and reporting the behavior and effectiveness of NEs by gathering statistical information, maintaining and examining historical logs, determining system performance and altering the system modes of operation [43]. It is highly important to the proper and efficient functioning of wireless telecommunication networks in the context of resource utilization, network planning, network optimization, problem diagnosis, and network availability monitoring [29]. Most of the time, KPIs are used to get an overview of the entire network performance. They result based on counters installed along network elements from statistical calculations, that can register, the number of voice calls performed as well as data calls, blocked calls, dropped calls, handover types or failed handovers. PM tools handling statistical counters show cells which are having problems, but the exact causes of the problems, or locations of the problems within the cell, are usually not provided [44].

2.5.2 Configuration Management

CM provides the operator with the ability to assure correct and effective operation of the 3G network as it evolves [8]. CM actions have the objective of controlling and monitoring the actual configuration of NEs and network resources.System modification service component and system monitoring service component are some parts of the CM service components. The first one is an action performed to introduce new or modified data into the system due to optimization or configuration, for example software upgrading. The second component, provides the operator with the ability to receive reports on the configuration of the entire network, or parts of it, from managed NEs. In terms of CM functions, they encompasses operator assistance in making the most

timely and accurate changes, ensure that CM actions will not result on secondary effects, traffic has to be protected from effects of CM actions and the mechanisms has to devised to overcome data inconsistencies [43].

2.5.3 Drive Tests

Drive tests are a method of measuring and assessing the coverage, capacity, and QoS of a mobile radio network by using moving vehicles. It is one of the methods with which user-side measurements can be collected [32]. The real performance of the cellular network is usually viewed from the perspective of mobile subscribers and this is the reason why operators use drive tests in assessing the coverage and the quality of their networks as the tests give the results from the field. It provides accurate real-world capture of the RF environment under a particular set of network and environmental conditions [32].

The most important reasons for drive tests performed in the network is for the optimization of capacity, coverage, mobility or QoS verification [45]. By measuring what a subscriber is expected to experience, in a particular location, MNOs can then make corrective planning for network performance improvement. It is however costly due to the recruiting of skilled engineers/surveyors and using vehicles. Moreover, it is constrained in both time and space and rarely covers in-building areas which greatly limits the validity of data as it restricts real-life scenarios.

2.5.4 Mobile Crowdsourcing

Mobile Crowdsourcing (MCS) refers to a group of people who voluntarily collects and shares data using widely available mobile devices. Mobile devices are equipped with abundant sensors (e.g., Global Positioning System (GPS), accelerometer, camera, etc) and powerful computing capabilities, which allow them to collect various types of data such as image/voice/video, location, and ambient information [46]. This trend enables individuals to sense, collect, process and distribute data around people at any time and place. Moreover, advances of communication technologies such as Wi-Fi, and Bluetooth, offer mobile devices direct connectivity to the Internet to exchange data at high speed at anytime and anywhere.

Having the capabilities on measurement collection at the user-side, operators are using this powerful tool in analyzing their network's performance. As an example,

radio coverage and quality can be monitored with such practice, by using subscribers' smartphones, which collect information on a signal level and quality that they are receiving and send such information to the operator. In this way, crowdsourcing can be used as a method of monitoring the network performance which can minimize drive tests and broaden the areas that are being monitored in real-time. In spite of the great benefits that MCS gains, it still faces several serious problems in terms of security, privacy, and trust. Not only these, but it may also increase costs as sometimes the subscribers are paid in order to provide this information. There are also other issues regarding this technology; among them, how should the network be built and structured to support MCS traffic or how to avoid the overwhelmed processing of subscribers' machines [47].

2.5.5 Traces

Traces are a means of network performance data acquisition resulting from communication between RNC and network elements attached to it including the user equipment. When connected to a Node-B from a certain RNC, a UE is constantly exchanging data within the network to inform about its communication conditions. All data from each UE are collected and logged at the RNC and represents a powerful feature to analyze and monitor the network performance. These data are called traces, and they are protocol events resulting from communication between RNC and network elements attached to it, UE (via Node B), Node-B, other RNCs and CN. All these events can be collected at the RNC through tools developed by vendors by activation of the trace functionality in an RNC [43].

Traces generate a huge amount of information, providing useful data on the quality of the communications. As an example, they contain measurements made by UEs relative to the quality of the signal that they are receiving with the storage of RSCP or Ec/No values [43].

Traces Recording Tools

Vendors have developed tools to process traces information to monitor their networks. As an example, there is the Ericsson's General Performance Event Handler (GPEH) system from Ericsson, and Call History Recording (CHR) for Huawei [48, 49]. Huawei CHR is a trace logging feature in GSM, UMTS and LTE that collects call and cell information, radio measurement and messages for all calls in the network and sends

to Nastar server. In this thesis Nastar tool is used for data collection and few details about the tool is provided in the following subsection.

Nastar: The Nastar performance analysis system is an intelligent and integrated tool developed by Huawei Technologies. It allows locating and analyzing wireless network quality problems and is applicable for GSM, GPRS, EDGE, UMTS, and LTE networks [2]. It supports the operations of multiple users; various wireless performance analysis and it is a basic support platform for further analyzing and locating wireless networks problems. The Nastar stands on a server-client architecture and includes a set of functions as service geographic observation, cell and terminal performance analysis as well as coverage, neighboring and pilot pollution analysis. Figure 2.2 shows the location of Nastar tool in the network.

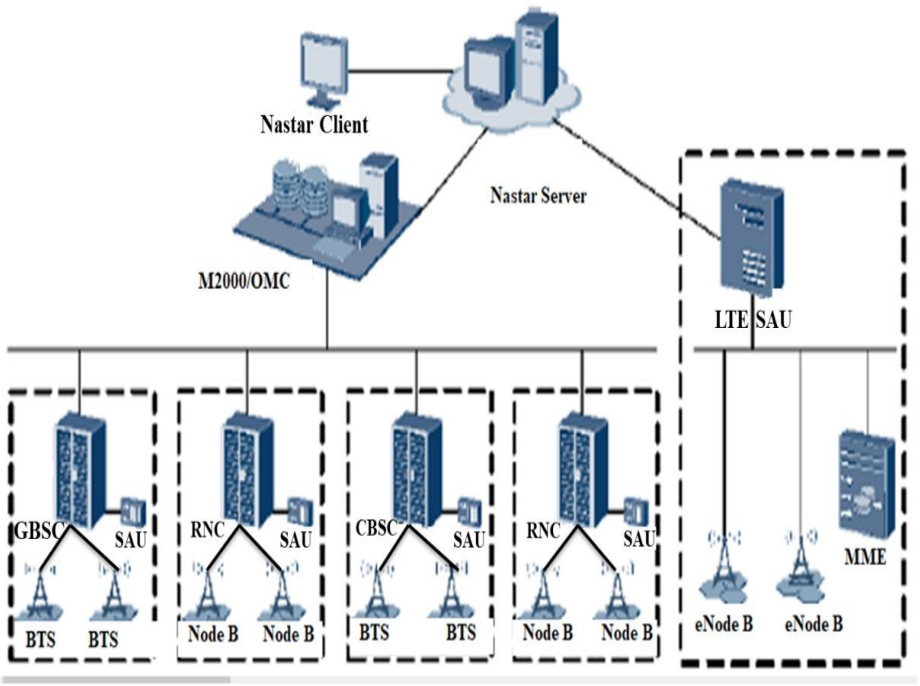


Fig. 2.2 Position of the Nastar in Network [2]

With the correct and precise analysis of traces information, many valuable possibilities are created, either in network performance management, processes automation (e.g., SON,MDT), QoS improvement or in costs reduction [50]. All these factors are made interesting for users and, even more important, for operators where the mechanisms contribute for easily manage the network and reducing unnecessary costs.

Trace Data Use Cases

Trace data are used for different use cases in telecommunication for network performance assessment. As per [51] trace data can be used in different use cases, such as to check radio coverage in a certain area, interoperability between UEs from different vendors, QoS profile for a subscriber after a subscriber complaint, to check the malfunctioning of mobile station or to test new features. This thesis focuses on one use case of trace data which is checking radio coverage status of UMTS network. To draw measurements, operators in the NGMN alliance proposed a standardized solution in 2011 called MDT (i.e., in 3GPP release-10 specification for LTE and UMTS networks)[52]. Key features of MDT are that a mobile device reports its location information along with performance measurements using Trace Collection Entity (TCE) via RNC, thereby allowing to have a much more fine-grain view of a cell's performance [3]. MDT addresses the issue that often drive-tests to have to be executed to monitor and assess mobile network performance in an efficient way. Some details are presented on MDT(architecture, reporting, location estimation, and measurement) in the following subsection.

2.5.6 Minimization of Drive Tests

MDT is a standardized mechanism introduced to provide operators with network performance optimization tools in a cost-efficient manner. It enables operators to collect UEs measurements together with location information, to optimize network management while reducing operational effects and maintenance costs [52]. It can provide the same type of information as those of the collected drive test by enabling normal UEs. The motives behind drive test evolution have been considered in 3GPP standardization and introduced this new feature (MDT) since 3GPP release 9 [13]. The main targets of MDT standardization were to optimize the drive test inefficiency. This approach allows operators to provide measurement data for radio network fault detection and optimization in all possible locations covered by the network. A great advantage is that UEs can retrieve and report parameters from indoor environments where drive testing cannot be addressed.

MDT Architecture

The aim of MDT functionality is to collect UE-specific radio measurements by using an architecture based on control plane signaling in UTRAN and E-UTRAN networks. The

benefit of using control plane architecture is that it allows RAN nodes (i.e., an eNB or RNC) to include additional data in UE measurements [3]. The measurements for MDT can be configured either by using management-based or signaling-based configuration procedures [4]. In signaling-based MDT, UE selection is performed in the Operation and Maintenance (OAM) based on a permanent UE identity, which uniquely identifies the UE in the network, such as International Mobile Subscriber Identity (IMSI) or International Mobile Subscriber Identity and Software Version (IMEI SV) [3]. In management-based MDT, trace functionality is used to configure a specific RAN node for collecting measurements for a certain area.

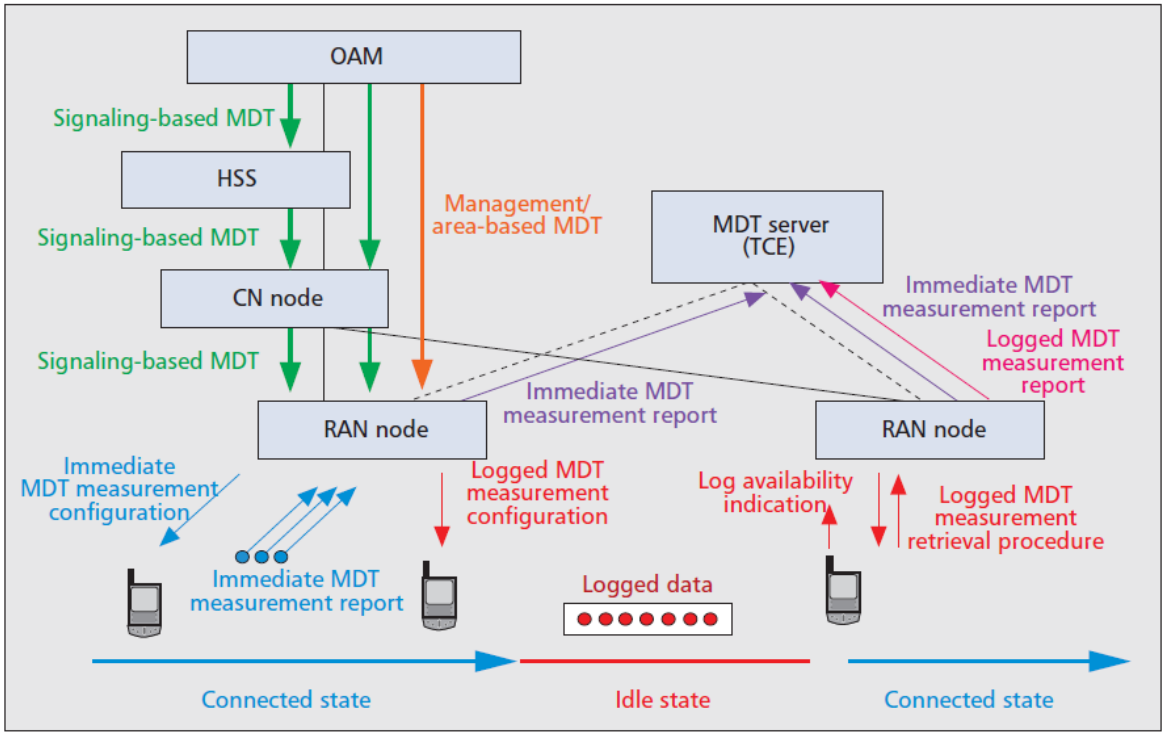


Fig. 2.3 MDT architecture [3]

Since this thesis is more about processing and analyzing the data collected from a set of UEs, the following sections focus on applications that employ a management-based MDT procedure. The MDT data collection is initiated and controlled by the OAM system and then the UE and RAN collects the data and sends it to the TCE, which stores the data and can be used for post-processing analysis [4]. An illustration of management-based MDT architecture is depicted in Figure 2.3 which is based on [3].

MDT Reporting

The overall MDT operation aims at delivering MDT reports to a data repository and the collected measurements are transferred to a specified file server which is called a TCE. There are two MDT operational modes of how the measurement collection can be done, namely immediate MDT and logged MDT [3, 4]. Immediate MDT uses the normal Radio Resource Controller (RRC) measurement configuration and reporting principles with the exception that the reported measurement data may include the UE location information at the time of measurement results are obtained. Logged MDT is a new mechanism for idle state UEs to store the radio measurement results to be reported later when the connection is set up next time [3]. Whenever RAN node receives the RRC MRs from UEs, it stores the measurements in a trace record together with its additional MDT information such as timestamp, trace parameters, and vendor-specific data and then forwarded to the TCE [3]. The procedure used for immediate MDT is shown in Figure 2.4.

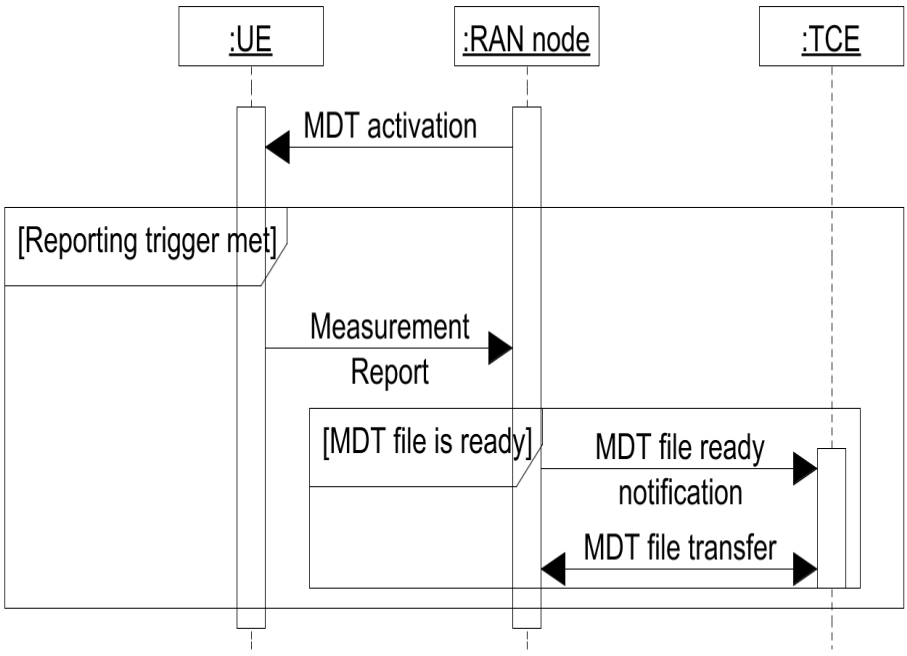


Fig. 2.4 Immediate MDT configuration and reporting [4]

The reporting mechanism of logged MDT and immediate MDT are not uniform. There is a little bit different among them. In logged MDT reports delivery, the RAN node or RNC is not aware of trace relevant configuration (MDT context is released after RRC connection release) before the MDT data is transferred to TCE. Hence, trace

relevant parameters (trace reference, trace recording session, TCE ID), memorized by the UE, are reported back to the network and attached by the UE to the MDT log. Trace reference and trace recording session reference is used to correlate the data at the TCE that belongs to the same trace (MDT) session [3, 4]. The procedure used for logged MDT is shown in Figure 2.5.

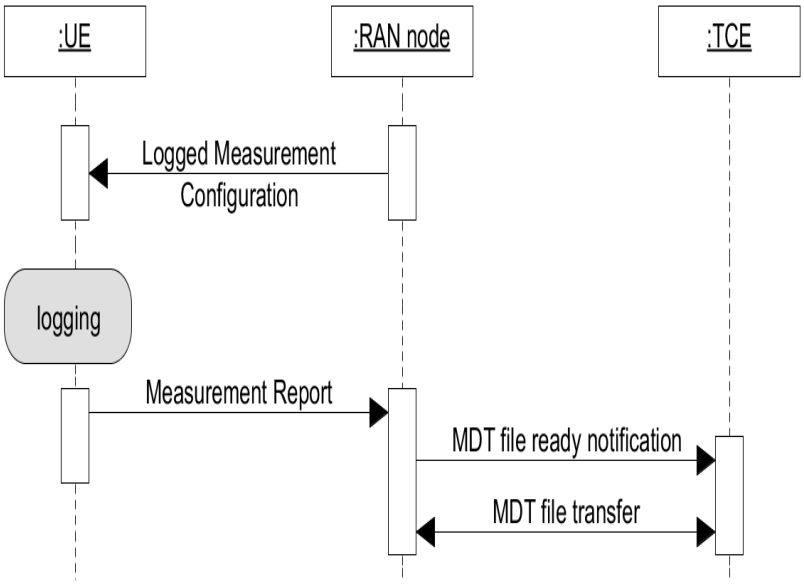


Fig. 2.5 Logged MDT configuration and reporting procedures [4]

Location Estimation in MDT

The UE location information can be included in the MDT reports in a best-effort manner. There are three ways how the UE location may be determined [52]: Cell ID/ serving Cell Global Identification (CGI); RF fingerprint (using neighbor cell measurements); Detailed location information/using stand-alone Global Navigation Satellite System (GNSS) positioning function. The most location information is the CGI. The cell ID or CGI will be available always known with immediate reporting or serving cell, and the serving cell ID will be stored with logged MDT. RF fingerprint is a profile of measured signal strength from neighboring cells. This information can be used by the net-work to calculate the approximate location of the UE by means of e.g., triangulation of the geographical location of the cell or the base station [3].The neighbor cell measurement results will be included in the report/log whenever available. As the UE may not be able to detect all the time neighbor cell signals (e.g. when close

to the serving cell base station), the positioning with RF fingerprint is not guaranteed in all locations [52].

In the best case, detailed location information is obtained from the GNSS if the satellite positioning has been activated by another function or application. If detailed location information is obtained from GNSS, then the MR shall consist of latitude and longitude, and a GNSS timestamp. With immediate MDT reporting, the UE does not send time stamp information as it does in the case of logged MDT. Instead, the RNC/eNB is responsible for adding the time stamp to the received MDT MRs when saving them to the trace records. However, if GNSS was used, the GNSS time information is included as a way to validate the detailed location information [53]. Even though active, the GNSS may not be able to provide position information continuously when poor signals received from the satellites, where indoors or are some locations in urban areas.

The MDT measurements and GNSS functions are normally independent functions, and therefore, also the timing when the measurement results and GNSS coordinates become available, can be random. The MDT function at the UE shall tag the measured result with the latest location information. A certain location sample shall be used only once in the GNSS report/log. The next accurate location information shall be included first when new coordinates are provided by the GNSS. The cell identification information consists of the serving cell CGI or Physical Cell Identification (PCI) of the detected neighboring cells. The measurements for both the serving and neighboring cells include the common pilot channel RSCP and Ec/No for a UTRAN system [53]. This thesis focused on the analysis of these measurement logs (RSCP and Ec/No) in order to detect where the coverage hole is for the optimization input purposes. Hence, each parameter is discussed in brief as follows.

MDT Measurements

The MDT procedure allows operators to collect radio measurements, such as received signal strength and quality with UE location information and a time stamp. In immediate MDT, the measurements can be conducted either periodically or network event-triggered based whereas, in logged MDT, the measurements are collected periodically [53].The MDT measurements consist of the location information with the longitude and latitude (if available); time stamp either from a UE or RNC/eNB depending on the MDT mode; cell identification data and the radio measurements for serving cell and detected intra frequency, inter-frequency and inter-RAT neighboring cells.

There are different mechanisms for estimating user location. The most location info is the serving CGI and in the best case, detailed location information is obtained from the GNSS. If detailed location information is obtained from GNSS, then the MR shall consist of latitude and longitude, and a GNSS time stamp [53]. With immediate MDT, the UE does not send time stamp information as it does in the case of logged MDT. Instead, the RNC/eNB is responsible for adding the time stamp to the received MDT MRs when saving them to the trace records. However, if GNSS was used, the GNSS time information is included as a way to validate the detailed location information [53]. The cell identification information consists of the serving cell CGI or Physical Cell Identifications (PCI) of the detected neighboring cells. The measurements for both the serving and neighboring cells include the common pilot channel RSCP and Ec/No for a UTRAN system [53]. This thesis focused on the analysis of these measurement logs (RSCP and Ec/No) in order to detect where the coverage hole is for the optimization input purposes. Hence, each parameter is discussed in brief as follows.

RSCP: is the signal code power measured by the receiver of a particular UE. It is used as an indication of received signal strength. It is measured on a CPICH and it can be obtained in both active and idle mode. RSCP measurement unit is dBm and has the range of -115 to -40 with a resolution of 1dB [40]. Handover process, cell selection and resection in the network rely heavily on the RSCP reported readings to the UE, which keeps measuring RSCP from the serving cell and the neighboring cells as well. RSCP provides information about the signal power but not the signal quality.

Ec/No: It is the ratio of the received energy per chip and the total received power spectral noise density of CPICH in the band. It is a radio quality measure for valuing the level of interference generated by the other cells [34]. Ec can be called RSCP value and No is the total receive power including thermal noise and interference. It is measured in dB as it's a relative value and has the range of -24dB to 0dB with a resolution of 1dB. The better this value the better can a signal of a cell be distinguished from the overall noise. The value is negative as the RSCP is smaller than the total received power. This value can be used to compare different cells on the same carrier and handover or cell reselection decisions can be taken.

RSSI: is the overall power (dBm), comprising the power of the serving cell, interference, and noise power received by the UE over the whole channel [34, 40]. RSSI helps in determining noise and interference information. UTRA carrier RSSI is given with a resolution of 1 dB with the range of [-94, ..., -32] dBm. Therefore, Ec/No measurement depend on both RSCP and RSSI [39] and it can be calculated using the Equation (2.1).

Chapter 3

Data Mining

Data mining is the science and technology of exploring data in order to discover previously unknown patterns and is a part of the overall process of knowledge discovery from databases (KDD) [54]. The two approaches which are ML and rules-based systems are widely used to make inferences from data. The two approaches have their strengths and weaknesses. Rules-based systems do still have their place in exploring data. They are a simple kind of Artificial Intelligence (Artificial Intelligence (AI)), which uses IF-THEN statements that guide a computer to reach a conclusion or recommendation with threshold values tailored to the evaluation scenario. Rules-based systems are typically built from the combined knowledge of human experts related with problem domain. These domain experts define all the steps to be taken to make a decision and how to handle any special cases. This full knowledge of the experts has incorporated into the system [55].

Writing and implementing rules in rules system is relatively easy. If we know about the situation of interest, we can create rules based on simple IF-THEN statements [55, 56]. However, rules-based systems are deterministic. Not having the right rule can result in false positives and false negatives, so the system of rules can become bulky over time as more and more exceptions and rule changes are added and can be difficult to grasp. Another challenge by rules-based systems is when the data and scenarios change faster than we can update the rules. They are always limited by the size of their underlying rule base (knowledge base) and are said to have rigid intelligence. For this reason and the other, rule-based systems can only implement narrow AI at best. The maintenance of these systems also too time-consuming and expensive. As such, rules-based systems are not very useful for solving problems in complex domains but

simple domains. They have been designed to perform a conservative detection and so that, they lack the ability to learn from experience. They cannot automatically update their knowledge base based on new information and they stick to the rules always [57].

ML is an alternative approach that can help to address some of the issues with rules-based methods. The methods typically only take the outcomes from the experts rather than attempt to fully emulate the decision process of an expert or best practice. For ML, exactly how the expert arrived at their decision is not important, only what their decision was is sufficient. Focusing on the outcomes rather than the entire decision-making process can make machine learning more flexible and less susceptible to some of the problems encountered with rules-based systems [55, 56].

ML is probabilistic and uses statistical models rather than deterministic rules unlike rules-based methods. In ML approaches the outputs (identified by historic outcomes data) can be described by the assumption of a combination of input variables and other parameters. The input variables can be numerous, and some models can use hundreds of inputs or features. The learning system is in principle unlimited in its ability to simulate intelligence and create its models. It is said to have adaptive intelligence,in which the existing knowledge can be changed or discarded, and new knowledge can be acquired. This quality and the other makes learning systems so different from rule-based testing. A machine learning model is trained on historic data outcomes already identified or labeled by human experts. As it is more amenable to continuous adaption and improvement through data preparation, algorithm selection, and algorithm parameter tuning it will be better in the long-run. Machine learning algorithms tend to be one step away from the human involvement in favor of optimization for computers [55].

Hence, most data mining techniques are based on inductive learning where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The primary assumption of the inductive approach is that the trained model is applicable to future, unseen examples. ML is an important field for data mining because of the algorithms that are used in data mining methods belong to algorithms that exist in the ML field [58].

3.1 Machine Learning Algorithms

ML is an application of AI that provides systems the ability to automatically learn, identify patterns and improve from experience with minimal human intervention

without being explicitly programmed [59]. It is applicable in different disciples due to its' capability to perform complex tasks (both tasks performed by human beings and beyond the capabilities of human beings), and its' adaptive nature to environmental change. Classification, clustering, association rule and numeric prediction are some examples of tasks performed using ML algorithms. Classification uses a set of labeled data for model learning and so that, new instances are classified into a set of predefined classes. In clustering the datasets are divided into a number of groups such that data points in the same groups are more similar to other data points in the same group whereas the data points are dissimilar to the data points in other groups. Association rule learning is a rule-based ML method that can be used for determining interesting relations between variables in large databases. Whereas numeric prediction is a machine learning method in which a constructed model or a predictor will predict a continuous valued function or ordered value [58].Mainly there are three primary ways in which machine can learn to do things such as supervised, unsupervised, semi-supervised [58, 60]. The focus of this thesis is on the use of a supervised learning for UEs MR data classification tasks for the purpose of cellular network coverage insights.

3.1.1 Supervised Learning

It is the ML task of learning or inducing a function that maps an input to an output based on labeled input-output pairs data. It is the type of learning that makes use of classes or categories already given to existing data, which are utilized for training purposes [61]. Some popular examples of supervised machine learning algorithms are: Linear regression for regression problems, DT for classification and regression problems, Support Vector Machine (SVM) for classification problems. . These algorithms are readily available and highly accessible through some programming tools like Python, among many others with varying levels of coding requirements [62]. This tool can have the capability to handle large amounts of data (too large data set) much more quickly, automate a process or run repeatable task analysis effectively. It can also reproduce previously conducted analyses on new datasets, create complex data visualizations while other tools like MS-Excel have serious limitations on these areas [62, 63].

In this work DT which applied for classification purpose will be focused. The reason why DT classifier focused is that, it is simple to interpret, do not need to be familiar with machine learning techniques to understand what a decision tree is doing as it can be visualized graphically, it requires little data preprocessing, it can have the capability to handle multi output , fast and applicable to any domain [64].

3.2 Decision Tree

DTs are a group of supervised learning methods in the concept of data mining and ML approach [58]. They are a flow-chart model in which each internal node represents a test on an attribute, each leaf node represents a response, and the branch represents the outcome of the test [65]. Similar to the expert system, the outcome of the DT algorithm can be regarded as a sequence of IF-THEN statements, with the difference that now these rules are being determined automatically. DT are important in data mining for various reasons but the most important reasons are that they provide accurate results and the tree concept is easily understandable compared to other classification methods [66].

Generally, DT has many appealing features than other classifier like: they can be visualized graphically and so that it's simple to understand and interpret; require very little data preparation whereas other techniques often require data normalization or standardization of features; the creation of dummy variables and removal of blank values; can handle both categorical and numerical data whereas other techniques are specialized for only one type of variable; can handle multi-output problems; use a white box model i.e. the explanation for the condition can be explained; can be directly converted to a set of simple if-then rules; robust and work well on noisy data; has well performance on large data sets [61, 64]. However, DTs are dependent on the coverage of the training data as with many classifiers. They are sensitive to the specific data on which they are trained. If the training data is changed the result of the decision tree can be different and as the result the predictions can be different. Moreover, they are also susceptible to over-fitting.

3.2.1 Decision Tree Structure

DTs consists of a root node, internal nodes and leaf (terminating nodes) that are connected by branches just like any other tree concepts [58, 67]. DT structure is shown in Figure 3.1 and the brief descriptions are presented below.

- Root node: It is a starting point of the tree where there are no incoming edges but zero or more outgoing edges. From the outgoing root node an internal node or leaf node is produced. It is usually an attribute of the DT model.

- Internal node: Appears after a root node or an internal node and is followed by either internal nodes or leaf nodes. It has only one incoming edge and at least two outgoing edges. Internal nodes are always an attribute of the DT model.
- Leaf node: These are the bottom most elements of the tree and normally represent classes of the DT model. Leaf nodes have one incoming edge and no outgoing edges that holds a class label.
- Depth: It is the maximal length of a path from the root node to a leaf node.

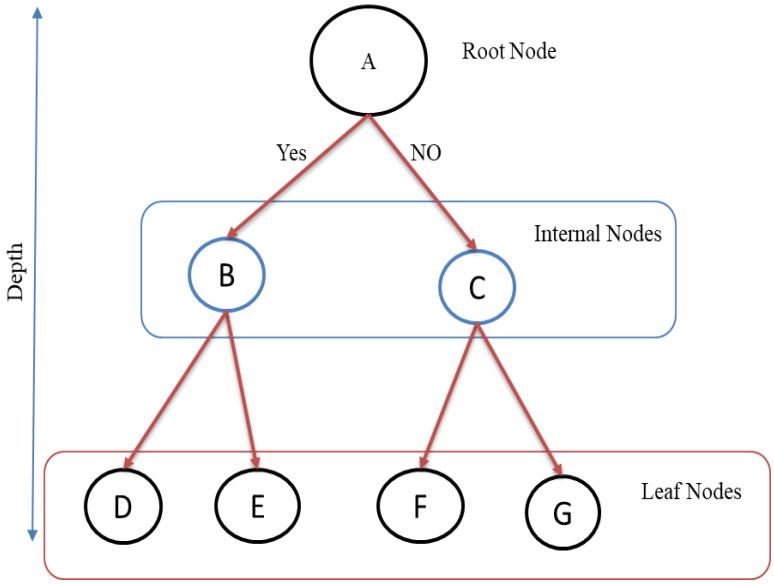


Fig. 3.1 Tree structure

DT are considered in two different categories as classification and regression. When a DT is used for classification tasks, it is called a classification tree and when it is used for regression tasks, it is referred to as a regression tree [54].

3.2.2 Decision Tree Types

Decision trees used in data mining are mainly of two types such as classification trees and regression tree. Classification trees are used when the predicted outcome is the class (discrete) to which the data belongs and are designed for data that have a finite number of class values. The attributes can take numerical or categorical values. Whereas regression trees used when the resulting leaf nodes of the tree are continuous [58].

3.2.3 Attribute Selection Criteria

Attribute selection is the idea of how to determine the best attribute that splits the data efficiently at each stage starting from the root. It is one of the fundamental properties of building a DT by selecting the attribute that is most useful for classifying the training data, which gives the maximum degree of discrimination. The selection of the attribute can affect the entire DT as it will have an impact on the efficiency and accuracy of the built tree [68]. The main idea is based on the purity of the dataset in most of the cases. This means the node that will be tested should be split into leaf or internal nodes that would be as pure as possible. The aim of purity is to partition the data instances in training data so that the partitioned group would either have all or most of the data instances in the same class category so that the entropy measure will be low [67].

To build a DT, identifying attributes for the root node at each level is required and so that in order to do that, attribute selection measures are used to select the attributes that partition the tuples into distinct classes. The popular measures/metrics used for attribute selection are Information Gain, Gain Ratio, Gini Index and Chi-Squared criterion [65, 67].In this work, information gain which is the function of entropy is used to measure how well the attribute splits the data.

Entropy

Entropy is the measure that tries to calculate the average amount of information contained in each message received [67]. In ML terms, entropy tries to find the most valuable attribute that would be beneficial for a model to be learned. It is a weighted sum of the logs of the probabilities of each possible outcome when we make a random selection from a set. It controls how a DT decides to split the data and draws its boundaries. The weights used in the sum are the probabilities of the outcomes themselves so that outcomes with high probabilities contribute more to the overall entropy of a set than outcomes with low probabilities. Mathematically, entropy can be defined as Equation (3.1).

$$Entropy = Info(D) = \sum_{i=1}^{m} (p_{(t=i)}) \times \log_2 (p_{(t=i)}) \tag{3.1}$$

Where, $p_(t = i)$ is the probability that the outcome of randomly selecting an element, t is the type i. At the beginning of the equation the minus sign is added to convert

the negative numbers resulted by the log function to positive ones. Equation (3.1) is a measure of the impurity or heterogeneity of a set. If samples are homogeneous all the elements are similar, then entropy is 0; else, if the samples are equally divided then entropy is maximum, which is 1 [68].

Information Gain

Given entropy as a measure of the impurity in a collection of training examples, it is possible to define a measure of the effectiveness of an attribute in classifying the training data. Information Gain (IG) is a measure of the expected reduction in the overall entropy of a set of instances that is achieved by testing or partitioning on a descriptive feature. It computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values by using the tree generation algorithms like ID3, C4.5, and C5.0 [67]. It measures how much information a feature gives us about the class. It generates a sequence of a test that divides the training data into pure sets with respect to the target values, then samples are labeled according to the test set sequence with consideration of corresponding target values.

To clarify the procedures, let node N represents the tuple of partition D. The attribute with the highest information gain is chosen as the splitting attribute for the node N. The expected information needed to classify a tuple in D is given by Equation (3.2).

$$Info(D) = -\sum_{i=1}^{m} p_i \times \log_2 p_i \tag{3.2}$$

Where, p_i is the probability that an arbitrary tuple in D belongs to class C_i and is estimated by $\frac{|C_{i,D}|}{|D|}$. Info(D) is the average amount of information needed to identify the class label of a tuple in D. Info(D) is also known as the entropy of D.

Equation (3.3) defines how we compute the entropy remaining after we partition the dataset using a particular descriptive feature. The expected information required to classify a tuple from D, based on the partitioning by attribute A is calculated by Equation (3.3).

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} \times Info(D_j) \tag{3.3}$$

Where, j indicates all the possible values that attribute A can take, D is the whole collection sample, D_j is the subset of the whole collection sample D for which attribute A has value j, $\frac{|D_j|}{|D|}$ weight of the j^{th} partition.

Information gain is defined as the difference between the original information requirement (i.e. based on the classes) and the new requirement (i.e. obtained after partitioning on A).Hence, by using Equation (3.2) and (3.3), we can now formally define information gain made from splitting the dataset using the feature A as Equation (3.4).

$$Gain(A) = Info(D) - Info_A(D) \tag{3.4}$$

Where, the first term in the equation is the entire entropy before partitioning the dataset and the second term of the equation is the entropy after splitting the instances using attribute A. This means that the information gain $Gain(A)$ is the expected reduction of entropy after knowing the value of attribute A [61].

Gini Index

Gini Index is a measure of inequality in the sample. It has a value between 0 and 1. Gini index of value 0 means samples are perfectly homogeneous (same class) and all elements are similar, whereas the Gini index of value 1 means maximal inequality among elements. Gini Index is an attribute selection measure used by the Classification and Regression Tree (CART) DT algorithm [64]. As in Equation (3.5), it measures the impurity of D, a data partition or set of training tuples [65]:

$$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2 \tag{3.5}$$

Where p_i is the probability that a tuple in D belongs to Class C_i and is estimated by $\frac{|D_j|}{|D|}$. D is a dataset, i is the set of levels in the domain of the target feature. The sum is computed over m classes. The attribute that reduces the impurity to the maximum level (or has the minimum Gini index) is selected as the splitting attribute.

Decision Tree Algorithm Learning

Algorithm learning is the systematic approach for learning a classification model given a training set. To be able to conclude new predictions from the existing datasets, DT

Table 3.1 Confusion matrix

		True Class	
		Positive	**Negative**
Predicted Class	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative(FN)	True Negative (TN)

induction is used. Induction is the process of using a learning algorithm to build a classification model from the training data [69]. It can be considered as the algorithm for a DT that use the basic tree concepts like creating a node, branching and combines these concepts with methods like attribute selection and pruning to build a tree model. Efficient algorithms have been developed to induce a reasonably accurate, optimal, decision tree in a reasonable amount of time [69]. The most prominent family of DT algorithms widely used are ID3 (Iterative Dichotomiser 3), C4.5, CART and CHAID (CHi- squared Automatic Interaction Detector) [64, 69].

3.2.4 Evaluation of Decision Trees

After doing data preprocessing and implementing a model and getting some output in forms of a class, the next step is to find out how effective is the model based on some metric using test datasets. Different performance metrics are used to evaluate different ML algorithms. This is achieved by using measures and metrics that will estimate the overall performance of the inducer's model for future use [70]. Well-known evaluation metrics to measure the classifiers performances are confusion matrix, accuracy, precision and recall-measure and Receiver Operating Characteristics (ROC) curve [71].

Confusion Matrix: The confusion matrix contains the numbers about actual and predicted class of the model used. During testing the raw data produced by classification scheme are the number of the correct and incorrect classifications from each class. The basic performance of a classifier can be indicated or evaluated by comparing these predicted labels against the true labels of instances as shown in Table 3.1.

The diagonal cells of the confusion matrices show the outcome of a true-positive test which indicates the likelihood that a sample is correctly labeled according to the class it belongs to. A false-positive test indicates the likelihood that samples are labeled incorrectly and have been assigned to the wrong classes. To evaluate the performance of the classification model, some terms which are derived from two-class labeled (positive and negative) data are important [70, 71].

True Positives (TP): These refer to the positive instances that were correctly labeled as positives by the classifier.

True Negatives (TN): These refer to the negative instances that were correctly labeled as negatives by the classifier.

False Positives (FP): These are the negative instances that were incorrectly labeled as positive by the classifier.

False Negatives (FN): These are the positive instances that were mislabeled as negative by the classifier.

A confusion matrix can provide the required information to determine how a classification model performs correctly, however summarizing this information into a single number makes it more appropriate to compare the relative performance of different models. This can be done using an evaluation metric such as accuracy, precision, recall, f-measure which are computed in the following way.

Accuracy: It measures the rate of all correctly classified instances by the total number of instances and is given in Equation (3.6).

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{3.6}$$

Precision (Positive predictive value): It represents the ratio of the number of correctly classified positives to the number of all the correctly classified positive and incorrectly classified positive instances. It is also called positive predictive, which can be calculated by the Equation (3.7). It represents the ratio of the number of correctly classified positives to the number of all the correctly classified positive and incorrectly classified positive instances.

$$Precision = TP/(TP + FP) \tag{3.7}$$

Recall (Positive sensitivity value): It represents the ratio of the number of correctly classified positives to the number of all the positive instances. It is also called positive sensitivity value, which can be calculated by Equation(3.8).

$$Recall = TP/(TP + FN) \tag{3.8}$$

F-measure: It is a model metric that can be used when we want to seek a balance between precision and recall (see Equation (3.9)).

$$\text{F-Measure} = 2/(1/Precision + 1/Recall) \tag{3.9}$$

Receiver Operating Characteristics (ROC): is a measure of classifier model which illustrate the tradeoff between true positive to false positive rates in graphical form. It is a two-dimensional graph in which the True positive Rate (TPR) represents the y-axis and False Positive Rate (FPR) is the x-axis.

Hence, in the next section the mentioned evaluation metrics especially confusion matrix and accuracy will be implemented for the evaluation of classifier (model) used for target classes prediction.

Chapter 4

Experimentation

As has been indicated in the objective part of Chapter 1, data mining technique is used to adapt a model that detects coverage hole in UMTS mobile network using UEs MR data as an input to the model. The detection has been performed by implementing the selected algorithm, which is DT, for knowledge mining. Such knowledge allows for better operation of the whole network by improving monitoring efficiency, specifically for early detection of hidden risks related to the joint effect of signal strength and quality by classifying into four coverage classes. To implement this task, specifically in this work, defining study-area,thresholds, coverage scenario and system process is required as shown in the following sections.

4.1 Area Selection

The test was carried out on 32 different mobile sites comprising 267 cells that are distributed over an area of $3000m \times 3000m$ with geographic coordinate range of (38.7076451E - 38.7348105E , 9.0404363N - 9.067913N) longitude and latitude respectively. The selected UMTS base stations are illustrated in Figure 4.1. It generally displays the distribution of all Addis Ababa UMTS Node-Bs and the selected region with the number of Node-Bs as well as their distribution on Google map. Based on the selected area, information like site ID, cell ID, RNC ID and location information (latitude and longitude) are captured. UEs MR parameters such as RSCP and Ec/No used in this work also captured by the support of these geographic information.

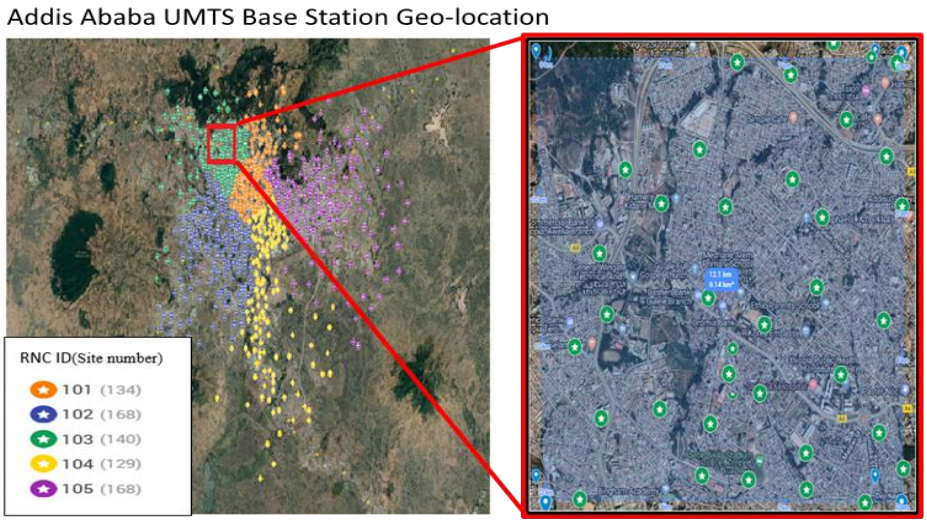


Fig. 4.1 Geo-location of deployed Addis Ababa UMTS base stations

Table 4.1 Geographical coordinates of available sites and the selected UMTS Node-Bs

	UMTS Node-B sites in Addis Ababa	**Selected Node-Bs**
Longitude	38.645469 – 38.940409	38.7076451 - 38.7348105
Latitude	8.818630 – 9.0943901	9.0404363 - 9.067913
Total Number of RNCs	5 (RNC101- RNC105)	1 (RNC 103)
Total Number of sites	744	32
Total Number of cells	7159	267

4.2 Threshold Definition

Mobile services are considered to be available when radio signal level values are above the minimum thresholds that allow their use. However, the thresholds may be varied as per the requirement of mobile operators, vendors, service requirements or technology [72–77]. For example in [74, 78], WCDMA coverage areas are classified considering the thresholds below -95dBm and -15dB as poor for signal strength and signal quality respectively. Also in [75, 77], different thresholds are defined for 19 European countries to qualify if there is outdoor coverage or not (covered / not covered). As it is observed from Table 4.2 there are five classes of coverage/quality levels (Poor, Fair, Good, Very good, and Excellent) [5]. In this thesis, the thresholds are considered based on the current use case of ethio telecom for coverage assessment which represents -95dBm for RSCP and -13dB for Ec/No [5, 33].

Table 4.2 Classification of signal coverage and quality based on RSCP and Ec/No level [5]

Coverage Quality Levels of UMTS	**RSCP(dBm)**	**Ec/No(dB)**
Poor	$-115 < RSCP < -95$	$-24 < Ec/No < -13$
Fair	$-95 \leq RSCP < -85$	$-13 \leq Ec/No < -10$
Good	$-85 \leq RSCP < -75$	$-10 \leq Ec/No < -8$
Very good	$-75 \leq RSCP < -65$	$-8 \leq Ec/No \leq -5$
Excellent	$65 \leq RSCP < Max$	$-5 < Ec/No \leq Max$

4.3 Coverage Scenario (Target Class) Definition

This section addresses the definition of the scenarios for network coverage status classification. As per [79], downlink CPICH coverage has to be verified by considering not only if the RSCP of the pilot channel is sufficient, but also by estimating the level of interference generated by the other cells. Such interference is typically quantified by the Ec/No of the CPICH. Where, Ec is an expression of power in CPICH and No is the cumulative sum of own cell interference, surrounding cell interference and noise density [39]. Ec/No value effectively estimated how much of the received signal can be used at a given location or how clean is the signal received. Different works of literature categorized network coverage status (grades) in different classes by considering RSCP and Ec/No separately, meaning that they did not use the joint effect of these two parameters at the same instant. To know the coverage problem of the area, they separately analyzed the parameters [72]. Hence, in this thesis, the defined scenarios are considered the joint effect of the parameters by benchmarking thresholds that ethio telecom is using for coverage/quality grading purposes.

In addition to this, what needs to be noted here is that even though there are more than two network coverage classifications like Very Good, Fair, Poor, this thesis considered only the threshold that separates the network coverage into two classes such as “good” and “poor” scenarios. By considering these two classes and the joint effect of two parameters (RSCP and Ec/No) the overall MR data were classified into four coverage scenarios as shown in Table 4.3. The brief description of coverage scenarios are explained below.

Coverage scenario 1 (Class-1): This scenario illustrates the area where both the RF signal strength (RSCP) and signal quality (Ec/No) are below the threshold as depicted in Table 4.3. This means that the RF signal strength (RSCP) and signal

Table 4.3 Summary of coverage scenarios

Weighted Target Classes (Coverage Scenarios)	Target Classes	State
4	Good Coverage and Good Quality	RSCP $\geq$ -95 (dBm) Ec/No $\geq$ -13 (dB)
3	Good Coverage and Poor Quality	RSCP $\geq$ -95 (dBm) Ec/No $< -13(dB)$
2	Poor Coverage and Good Quality	RSCP $<$ -95 (dBm) Ec/No $\geq$ -13 (dB)
1	Poor Coverage and Poor Quality	RSCP $<$ -95 (dBm) Ec/No $<$ -13 (dB)

quality (Ec/No) are below -95dBm and -13dB respectively. It implies that the areas have a critical coverage hole problem both in signal strength and quality.

Coverage scenario 2 (Class-2): This scenario illustrates the area where RF signal strengths (RSCP) are below the threshold but the signal qualities (Ec/No) are above the threshold sated. This depicts that the areas have coverage hole problems due to poor signal strength.

Coverage scenario 3 (Class-3): This scenario illustrates the area where the RF signal strength (RSCP) is greater or equal to -95dBm and signals quality (Ec/No) is below -13dB which indicates the coverage hole due to signal quality.

Coverage scenario 4 (Class-4): This scenario illustrates the area where both RF signal strength and signal quality are above or equal to the thresholds. This means the areas are in good condition both in signal strength and quality. The summary of all scenario is illustrated in Table 4.3.

4.4 System Process

Having the thresholds and coverage scenarios illustrated the main processing steps of the framework are shown in Figure 4.2. As can be seen, the analysis starts with the data collection and preprocessing of MR data and then split the data as a training and testing set. Applying the data on the model, model learning, testing, and evaluations are parts of the model framework. In the next process, the model classifies the required target classes which are coverage scenarios. Finally, those coverage scenarios are

visualized on a Google map which represents the coverage classes in location. The brief description is explained below.

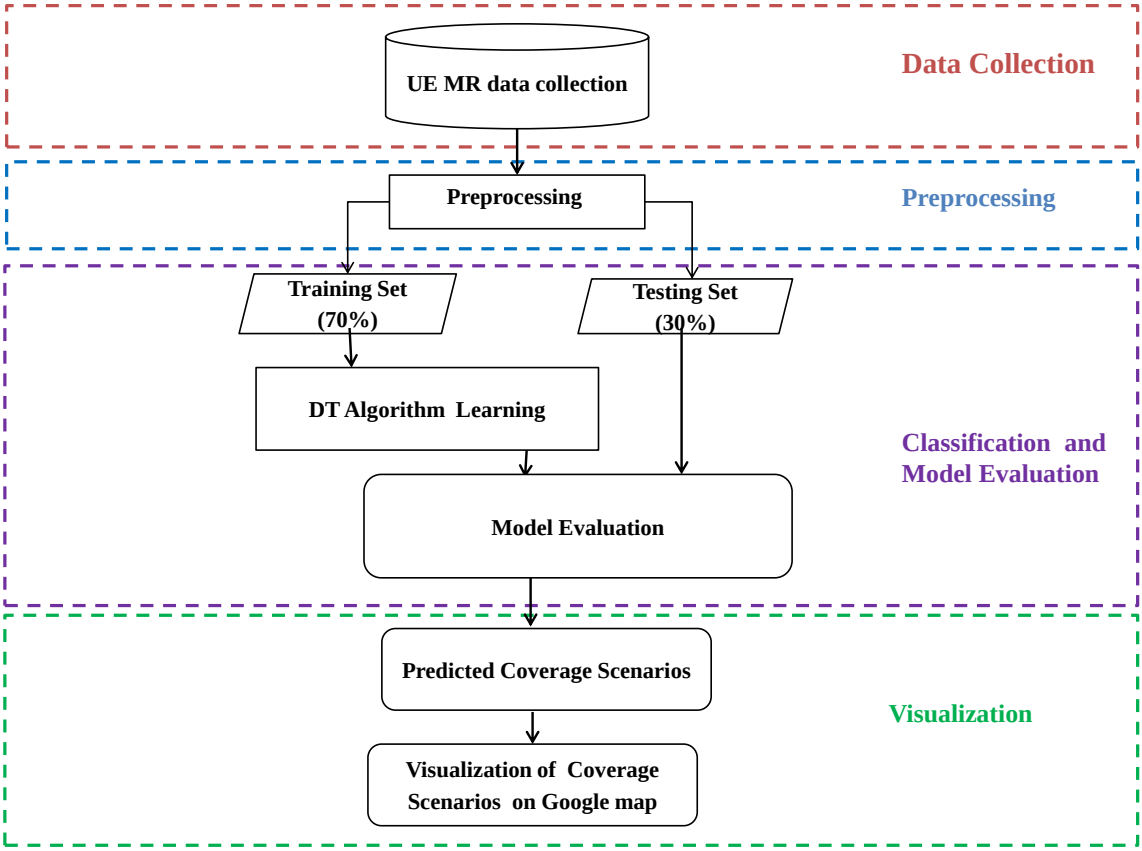


Fig. 4.2 System process

1. **Data collection:** In order to create an initial database, the UE MR data are collected from RNC interface using Nastar tool. RSCP and Ec/No parameters, including serving Cell ID and event locations (latitude and longitude) are captured for the analysis. Three months of data with averaged daily busy hours traffic was considered.

2. **Preprocessing:** In order to obtain a good performance during the evaluation, the input variables of the different measurements must be preprocessed. This is the step of the system process (model) where the data cleaning, dimensionality reduction, and missing values have been filtered. After this stage datasets are separated into two parts as a training set and test set. Table 4.4 shows samples of input dataset.

Table 4.4 Input dataset sample

Index	Date	Long.	Lat.	Cell	Ec/No	RSCP
0	1/6/19	xxx	xxx	xxx	-11.99	-88.33
1	1/6/19	xxx	xxx	xxx	-10.11	-91.14
2	1/6/19	xxx	xxx	xxx	-8.92	-84.32
...	...	...	...	...	...	...
350834	31/7/19	xxx	xxx	xxx	-16.83	-103.67
350839	31/8/19	xxx	xxx	xxx	-17	-92

3. **Classification and Model Evaluation:** After the preprocessing step, the DT algorithm was applied to the data for classification. The dimensions of the dataset used for classification is 2×350,839. These datasets are separated into two pieces such as the training set, which takes exactly 70% of the randomly selected data points, and testing set, which contains the remaining 30% of data samples. Here model evaluation also performed by applying test data set to the trained model and finally tuning parameters (maximum depth, maximum leaf node setting) are also performed for DT optimization. To evaluate the performance of the classifier we compare the original labels, described data, with the prediction results of DT classification. Table 4.5 shows the data sample after prediction. Figure 4.3 shows the tree generated after learning the model.

Table 4.5 Data sample after prediction

Index	Date	Long.	Lat.	Cell	Ec/No	RSCP	Predicted Class
0	1/6/19	xxx	xxx	xxx	-11.99	-88.33	4
1	1/6/19	xxx	xxx	xxx	-14.11	-96.1	1
2	1/6/19	xxx	xxx	xxx	-8.92	-84.32	4
...	...	...	...	...	...	...	...
105226	31/7/19	xxx	xxx	xxx	-15.3	-86.6	3
105238	31/8/19	xxx	xxx	xxx	-17	-106	1

We learn from Figure 4.3 that, 70% (245,587) of the samples are used for training the model out of 350,839 instances. RSCP is selected as the best splitter and used as a root node attribute to split the data sets. The 'value' row in each node tells us how many of the observations that were sorted into that node fall into each of four categories. Finally, four-leaf nodes (classes) were formed with observed samples value.

4. **Visualization:** After model evaluation, the target coverage scenarios displayed on Google map for visualization.

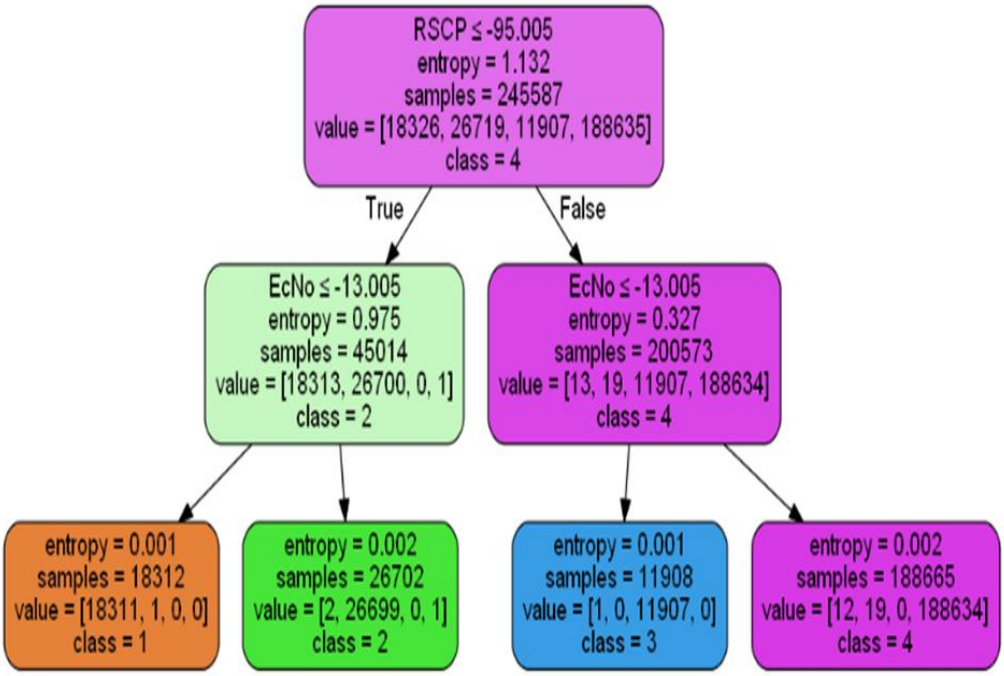


Fig. 4.3 Decision tree chart

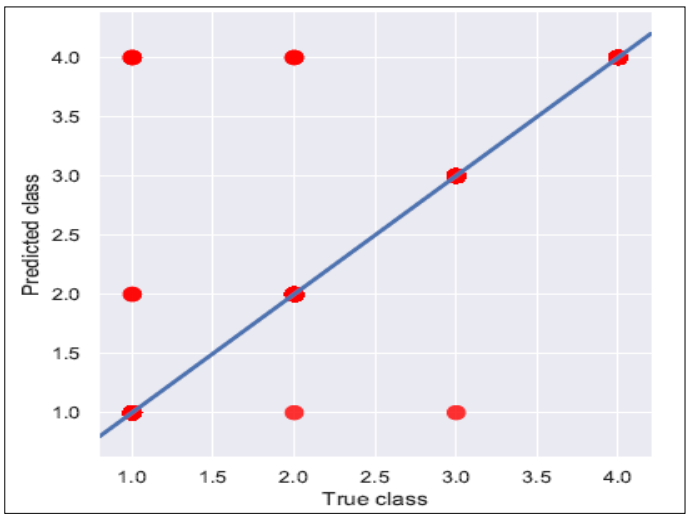


Fig. 4.4 Plot form of classifier confusion matrix

Table 4.6 Generated test result in confusion matrix form

		Predicted Class			
		Class-1	Class-2	Class-3	Class-4
True Class	Class-1	7880	2	0	4
	Class-2	1	11443	0	3
	Class-3	1	0	5158	0
	Class-4	0	0	0	80760

Figure 4.4 reports the predicted class against true class results for the classification of the four coverage scenarios from MRs in terms of the confusion matrix in plot form. We learned from the plot that there are points out of the diagonal line which represent the incorrectly classified instances. The points on the diagonal line show the datasets correctly classified as per their respective classes. Then, performance measure which is accuracy is calculated as per the corresponding formula presented in Equation (3.6). From the confusion matrix in Table 4.6, we can see that out of 105,252 test instances, the algorithm misclassified only 11. This resulted in 99.98% accuracy which is good. As it is observed from the table, there are four classes of coverage scenarios, such as class-1, class-2, class-3, and class-4. The values of TP1, TP2, TP3, and TP4 are 7880, 11443, 5158 and 80760, respectively, which represent the diagonal in the table.

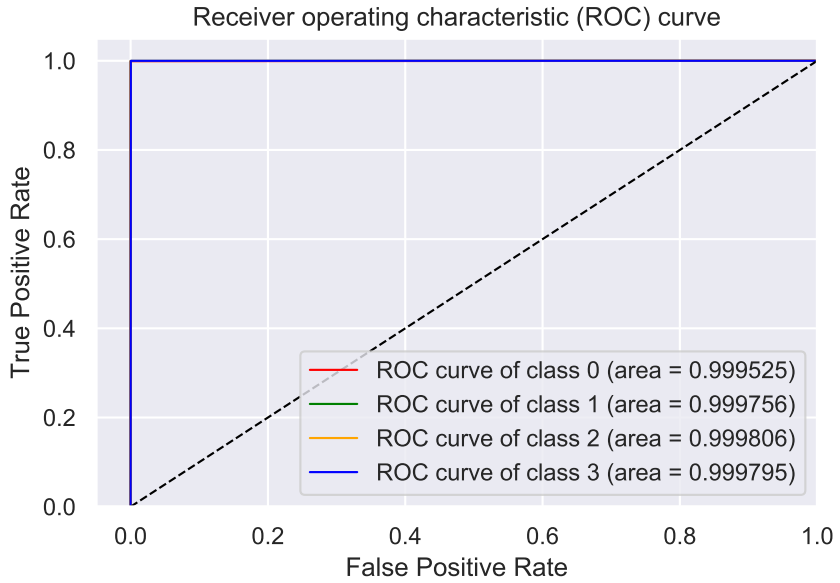


Fig. 4.5 ROC curve for DT classifier-based coverage hole detection

Figure 4.5 shows the graphical measures (ROC) curve for four classes (coverage scenario). In the ROC curve, there are four important points to be noted. The point at the lower-left corner (0,0) represents a classifier where there is no positive classification, while all negative samples are correctly classified and hence True positive Rate (TPR) = 0 and False Positive Rate (FPR) = 0. The point in the top right corner (1,1), represents a classifier where all positive samples are correctly classified, while the negative samples are misclassified. The point in the lower right corner (1,0) represents a classifier where all positive and negative samples are misclassified. The point in the upper left corner (0,1) represents a classifier where all positive and negative samples are correctly classified; thus, this point represents the perfect classification. Accordingly, Figure 4.5 shows perfect classification performance. Almost all curves(classes) (Red, green, orange, blue) rises vertically from (0,0) to (0,1) and then horizontally to point (1,1). These curves reflect that the classifier perfectly ranked the positive samples relative to the negative samples. The Area Under Curve (AUC) for all classes is approaches to 1 which shows a good performance of the classifier.

Chapter 5

Result and Discussion

The data used to achieve these results has collected from the RNC of Addis Ababa UMTS network using the Nastar tool. Periodic MR instances are used in this process, as the relevant information using MDT functionality. Cell IDs of serving cells, event location, and network performance indicator parameters such as RSCP and Ec/No values reported, are contained in those reports. These data collected from June 1, 2019, to August 31, 2019, which is three months' duration. For validation purpose, driving test task has performed and collected the required data. The results are indicated in terms of qualitative and quantitative aspects as shown in the following sections.

5.1 Qualitative Result

Figure 5.1 depicts the 3D scatter plot, reporting the target classes distribution. The classification was based on RSCP and Ec/No to exploit the joint effect of the parameters on coverage hole detection rather than observing separately. As it is observed from the figure the attributes could classify the target classes properly and so that someone can diagnosis the root cause easily and trigger optimization by setting priorities as per the requirements.

Figure 5.2 also reports what the target classes look like in a contour plot. From the contour map, the red color shows a section where a high concentration of poor signal strength and qualities are available. Here some variation of signal distribution is observed, which reflects small area has covered with this signal type and so that there is a fast transition to other classes. The green color shows the section where good signal strength and qualities are concentrated. This section looks smooth in most

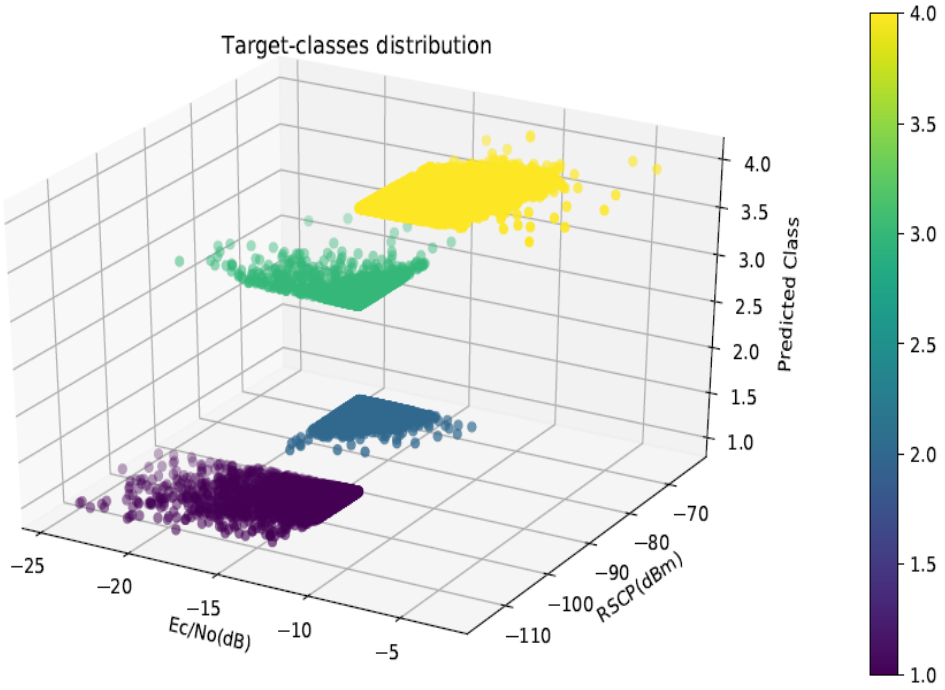


Fig. 5.1 3D scatter plot of target classes on testing data set

of the areas as the coverage of this class covers large areas with a uniform level of signal distribution. However, there are also small sections of other coverage classes distributed in this region. The violet color is the area where poor coverage and good quality signal levels are concentrated. Whereas the blue one shows where good signal strength but poor-quality levels are concentrated. From the plot, the other information to be observed is that there is the overlapping of classes 1 and 2 in the same area. The area with black color is where there is no signal measurement taken.

Figure 5.3 illustrates the overall qualitative distribution of CPICH coverage scenarios against the test area on the Google map for the entire test duration. The color of pixels corresponds to the coverage scenarios (classes) of the area. It can be learned from the figure that all target classes such as Class-1, Class-2, Class-3, and Class-4, which are represented by orange, violet, blue and green colors respectively, are reflected densely in some respective areas. The orange pixels in the figure labeled as A1, A2, A3, and A4 are the areas where a mobile terminal cannot detect CPICH signal at the required RSCP and Ec/No levels. Whereas the violet pixels labeled as B1 and B2 are the place

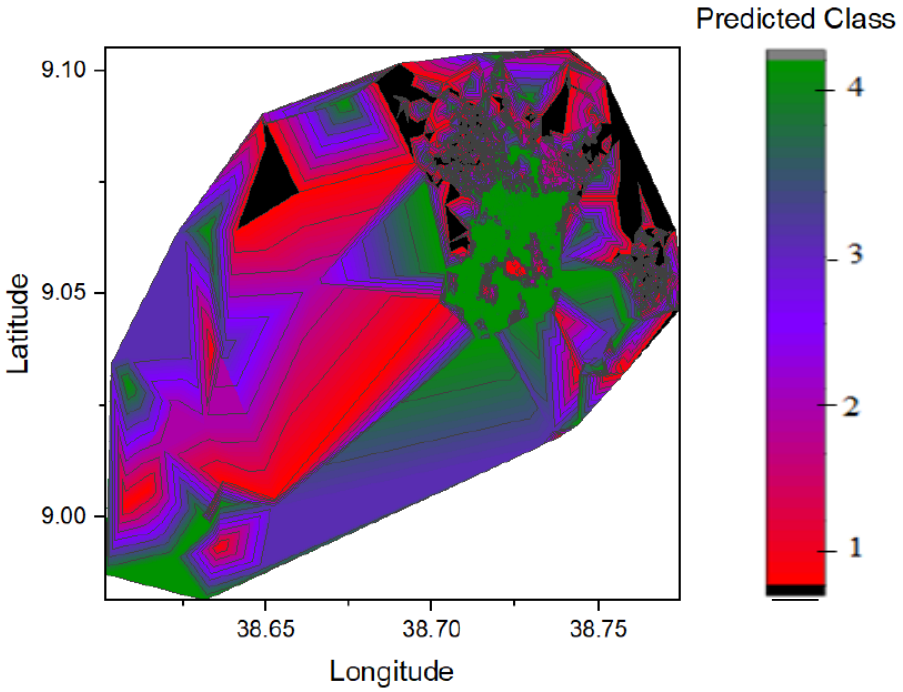


Fig. 5.2 Contour plot of target classes on testing set.

where a mobile terminal cannot detect required RSCP but Ec/No which tells us the areas have coverage problems related to poor signal strength problems.

The green pixels in Figure 5.3 shows the areas where mobile terminals detect CPICH signal at the required RSCP and Ec/No levels, which implies that the areas have no coverage-hole problem. Whereas, the blue pixels labeled as C1, C2, and C3 shows the areas where mobile terminals detect required RSCP but not Ec/No. This indicated that the areas have no signal strength problem but served by the poor signal quality, which infers the existence of coverage-hole problem at the area due to bad signal quality (interference).

The following discussions in Figure 5.4 to Figure 5.6 more explains the detail of each coverage scenarios and the relations to each other. Figure 5.4 depicts the distribution of coverage scenarios 3 and 4 which represented with blue and green pixels respectively. In this figure the majority of the areas covered by good signal strength and quality (coverage scenario 4). Parts of the area covered with good signal strength but with poor quality (coverage scenario 3). From the quality point of view, if an area has good signal coverage but has a poor Ec/No value, the provided service will not be at the desired level, it will be poor service. In the other way, there are uncovered areas with both scenarios, specifically those areas labeled as A1, A2, A3, and A4. What we

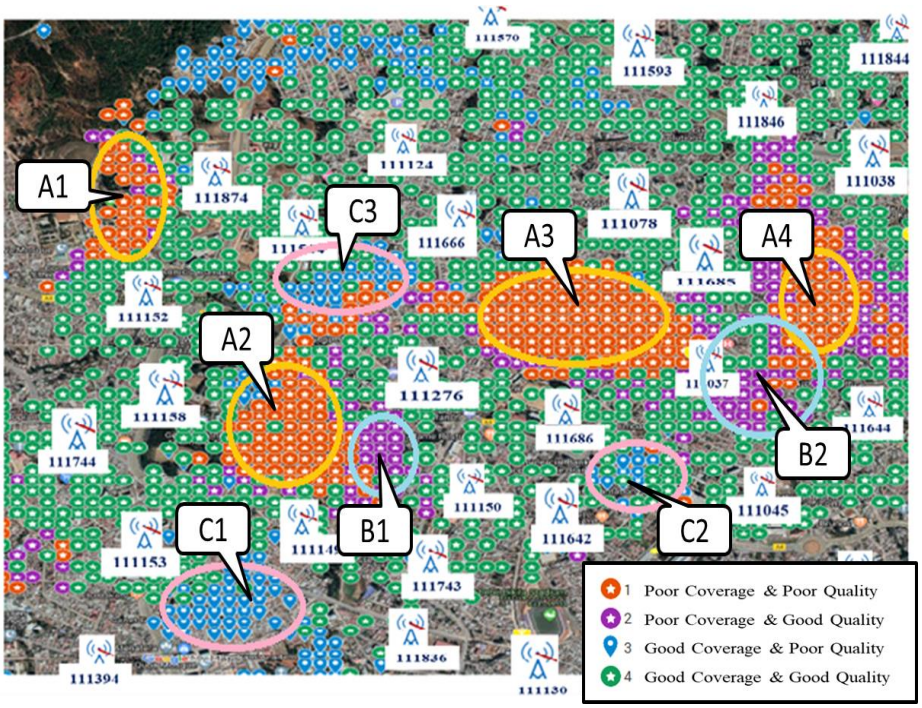


Fig. 5.3 Spatial distribution of coverage scenario

learned from this is that, throughout the test duration (for about 90 days) the areas did not get CPICH signals of scenario 3 and 4.

Figure 5.5, shows the spatial distribution of coverage scenario (2,3,and 4). we learn from this figure that the areas (A1, A2, A3, and A4) not covered so far in Figure 5.4 by coverage scenarios 3 and 4 have covered with coverage scenario 2. The coverage status of other areas looks almost uniform, with the result found in Figure 5.4. This means that, the uncovered problem(area) seen in Figure 5.4 has covered/served with poor coverage and good quality (scenario 2).

In Figure 5.6, the areas (A1, A2, A3, and A4) not covered with coverage scenarios 3 and 4 in Figure 5.4 are here covered with poor coverage and poor quality, which is coverage scenario 1.

Generally, what we have learned in Figure 5.3 to Figure 5.6, is that the certainty of the existence of the coverage scenarios on the respective area was high, which could not be the random existence of the coverage problem, due to any random environmental factors. All coverage-hole classes could be identified spatially. A high correlation of coverage scenarios 1 and 2 have observed against the area.

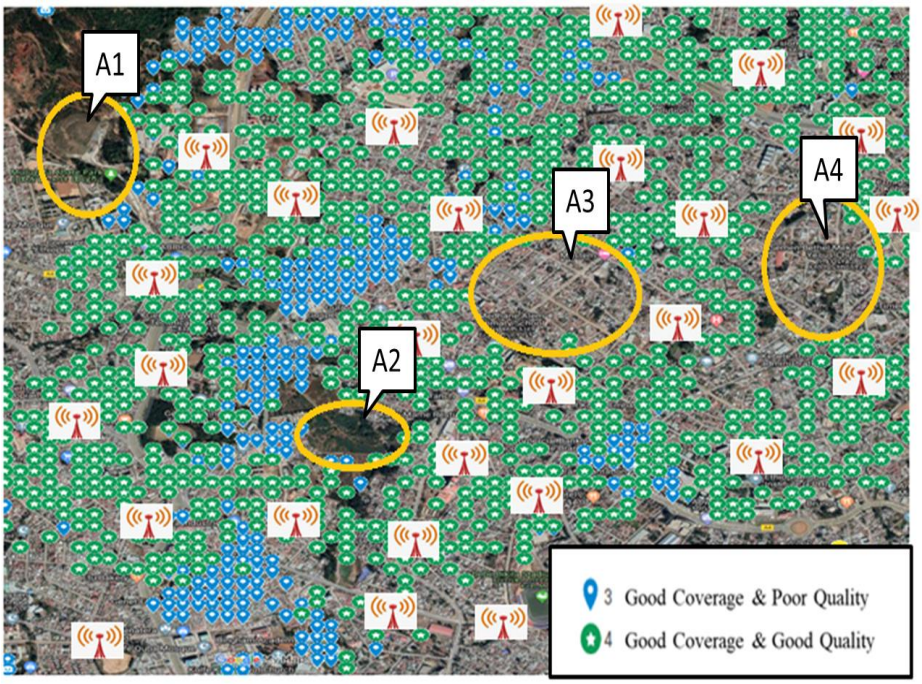


Fig. 5.4 Spatial distribution of coverage scenario 3 and 4

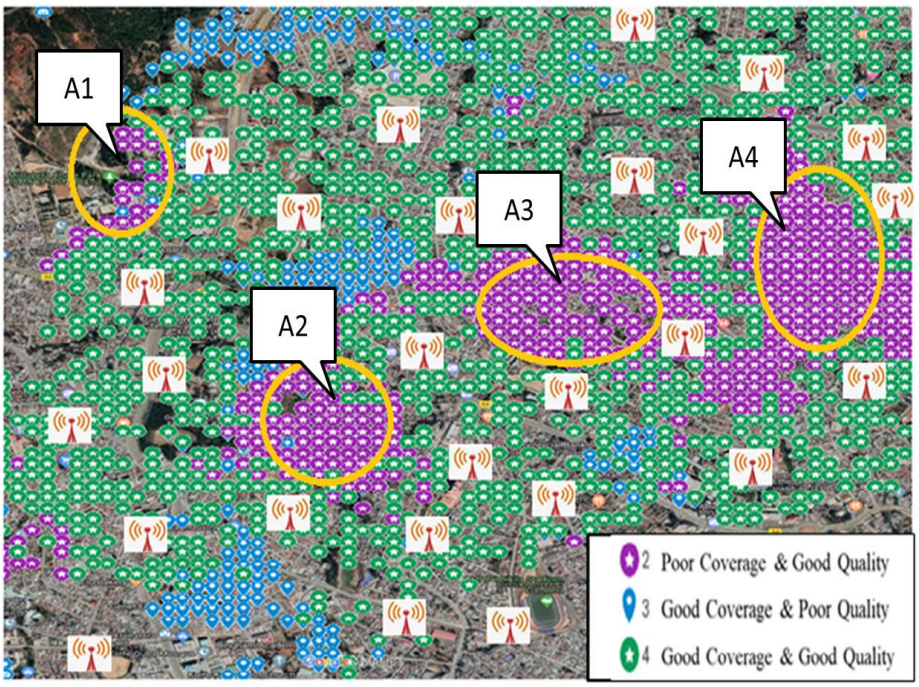


Fig. 5.5 Spatial distribution of coverage scenario 2, 3 and 4

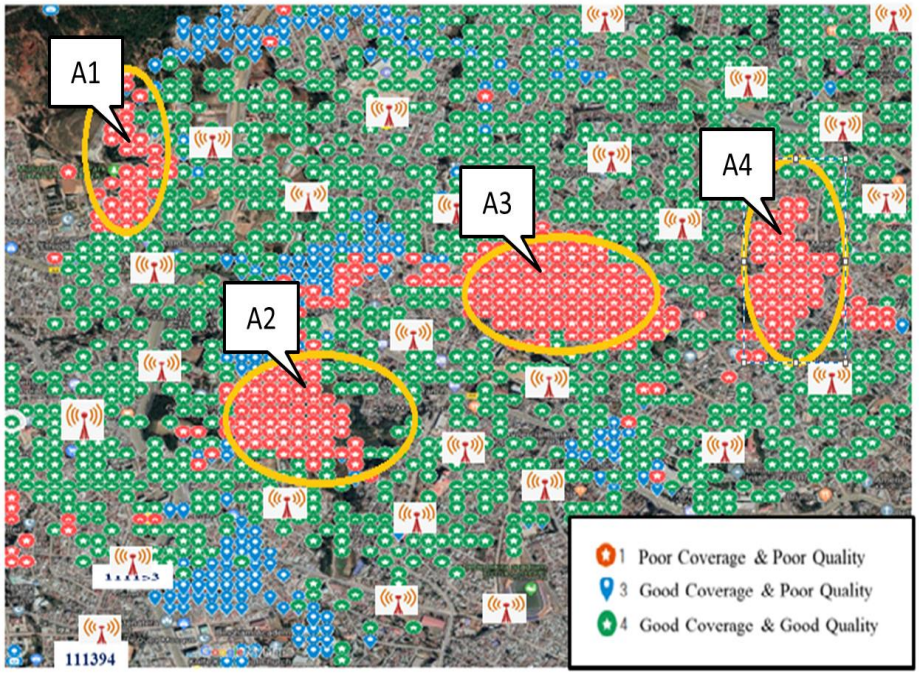
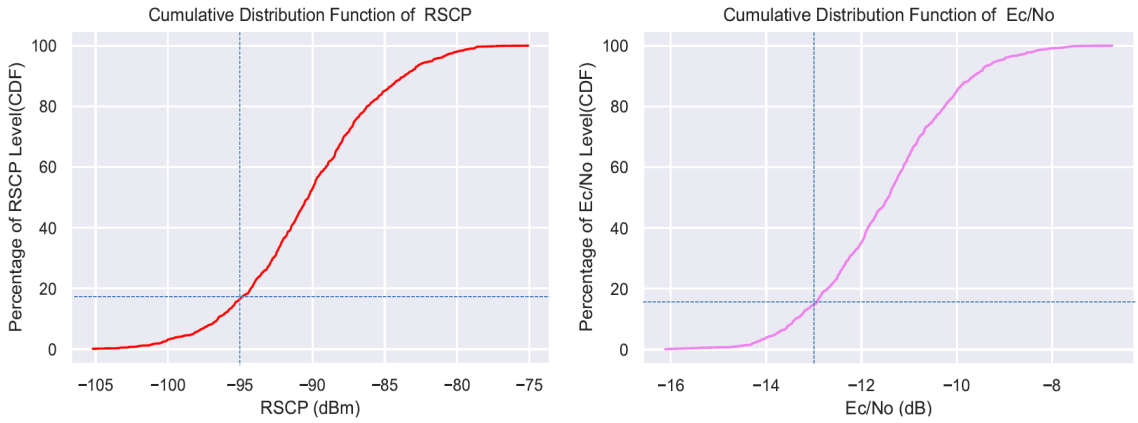


Fig. 5.6 Spatial distribution of coverage scenario 1, 3 and 4

5.2 Quantitative Result

Figure 5.7 illustrates the quantitative analysis of RSCP and Ec/No. Cumulative Distribution Function (CDF)s of MDT measurements in terms of RSCP and Ec/No, are present in Figure 5.7a and 5.7b respectively and the respective average, maximum, minimum, standard deviation, variance and medians of RSCP and Ec/No are also presented in Table 5.1.

Figure 5.7a and Figure 5.7b illustrate the signal strength and quality respectively. We can learn from the figures that about 18% of the total RSCP and Ec/No is below the threshold value. This means that 18% of the RSCP and Ec/No values are below -95dBm and -13dB respectively. When we see the average value, both parameters (RSCP and Ec/No) are in the range of the operator's thresholds value, which is -90.62dBm and -11.51dB respectively. However, there are variances of 25.27 for RSCP and 2.26 for Ec/No which need to be minimized for reliable QoS. Moreover, a significant variation is also observed from the minimum and maximum value of the parameters observed. The problem to be noted here is that the distribution of the CPICH signals coverage is not uniform and so that there are areas served with poor signal strength and quality



(a) RSCP CDF (b) EC/No CDF

Fig. 5.7 CDF of RSCP and Ec/No

throughout the test period that 18% of poor signal condition does not represent exactly the entire test area.

Table 5.1 Quantitative values (statistic) of RSCP and Ec/No

Quantitative Values (Statistic)	**RSCP**		**Ec/No**	
	MDT	**Drive test**	**MDT**	**Drive test**
Mean	-90.62	-73.88	-11.51	-9.67
Maximum	-64.00	-34.85	-3.00	-2.4
Minimum	-115.00	-118.96	-24.50	-25.9
Median	-90.58	-74.0	-11.55	-9.37
Standard deviation	5.03	12.2	1.50	3.73
Variance	25.27	149	2.26	13.95

From Figure 5.7a it can also be concluded that the area coverage probability is below the target for all morphologies (Dense Urban, Urban, and Suburban) against ethio telecom coverage probability requirement which was 95% for all -75dBm, -81dBm and, -85dBm for Dense Urban, Urban, and Suburban respectively [80] for outdoor RSCP requirements (dBm) assumption during UMTS coverage planning. However, as we have learned from Figure 5.3 the distribution of the RSCP was not uniform. Hence, the operator needs to be aware as there are some areas where signal strength almost totally out of the target range while a few areas fulfill the requirements.

In addition, when we see the coverage status of target classes, the majorities are covered by good RF signal conditions in terms of strength and quality as depicted in Figure 5.8. In this figure, the good coverage and quality scenario shares 77% of the entire sample

of the evaluation while the poor coverage and poor quality shared only 7% of the whole region. The other scenarios (good coverage but poor quality) and (poor coverage but good quality) shared 5% and 11% respectively. Even though there is a high share of good coverage and quality, there is an area critically served with poor signal strength and quality throughout the test period.

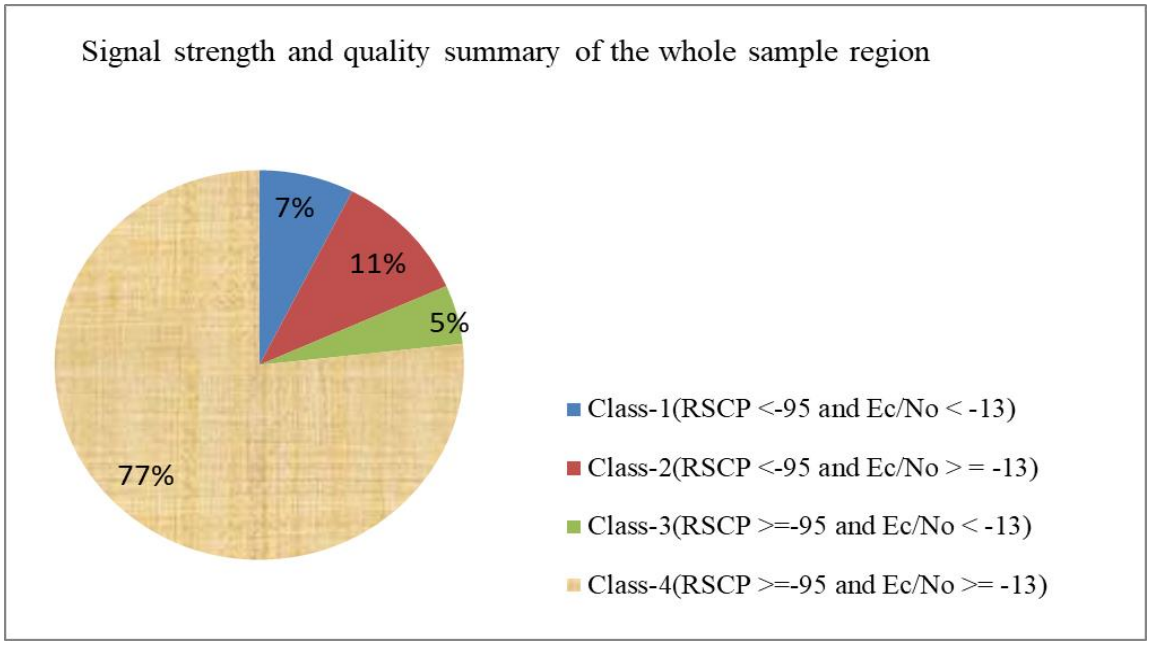


Fig. 5.8 Coverage scenarios share (%) of testing dataset.

5.3 Result Validation

In order to validate the detected coverage hole based on MDT data, drive tests were carried out in defined areas and used as a comparison. RSCP and Ec/No are the parameters considered in both cases since they can have the capability to show the full picture of the network status in terms of coverage and quality.

As shown in Figure 5.9, the result found or the detected coverage hole by using the MDT data is almost the same with the result found from drive testing. It can be learned from the figure, that there is a slight variation in signal strength. The signal strength of the drive test result looks better than the detected result, especially at the yellow circled area. This result is expected in the real scenario since the driving test has performed in the outdoor environment, where the degradation of signal strength due to the building is less.

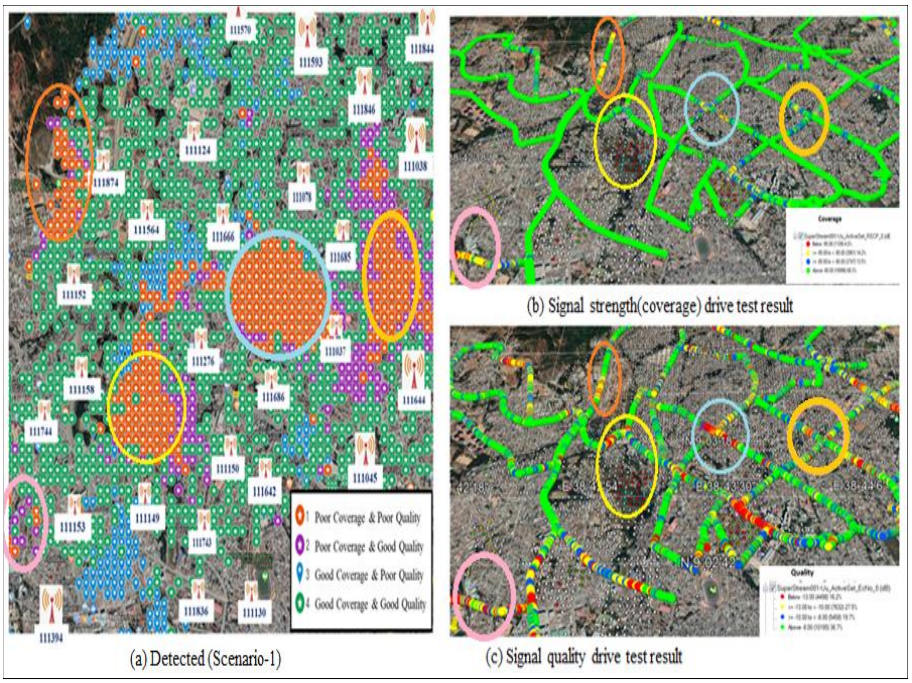


Fig. 5.9 Detected coverage scenario-1 comparison with derive test result.

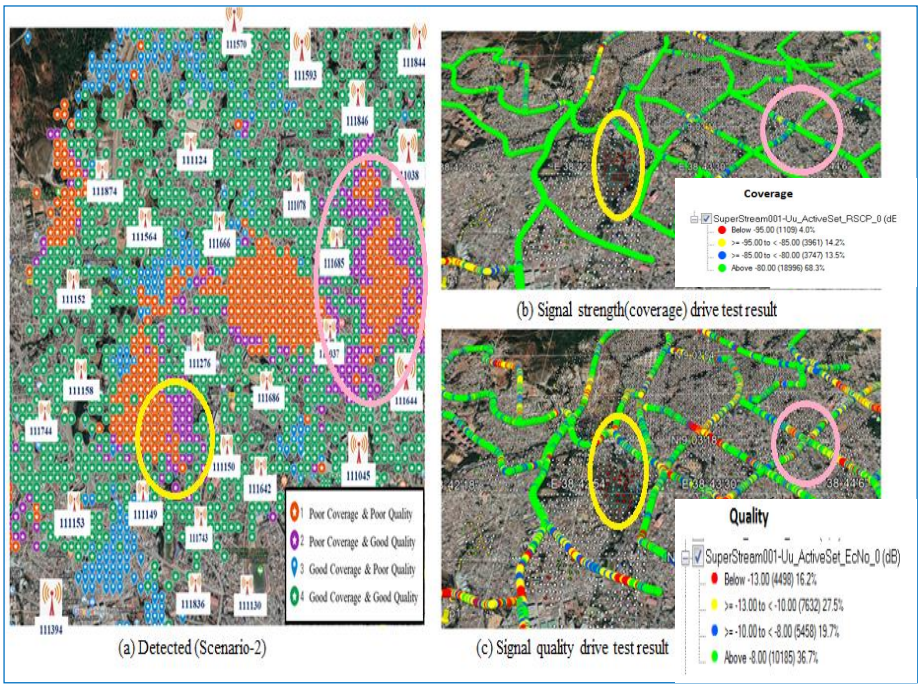


Fig. 5.10 Detected coverage scenario-2 comparison with derive test result.

In Figure 5.10, coverage scenario-2 has presented where the mobile terminals in the areas cannot detect required RSCP but Ec/No. The validation result showed that the

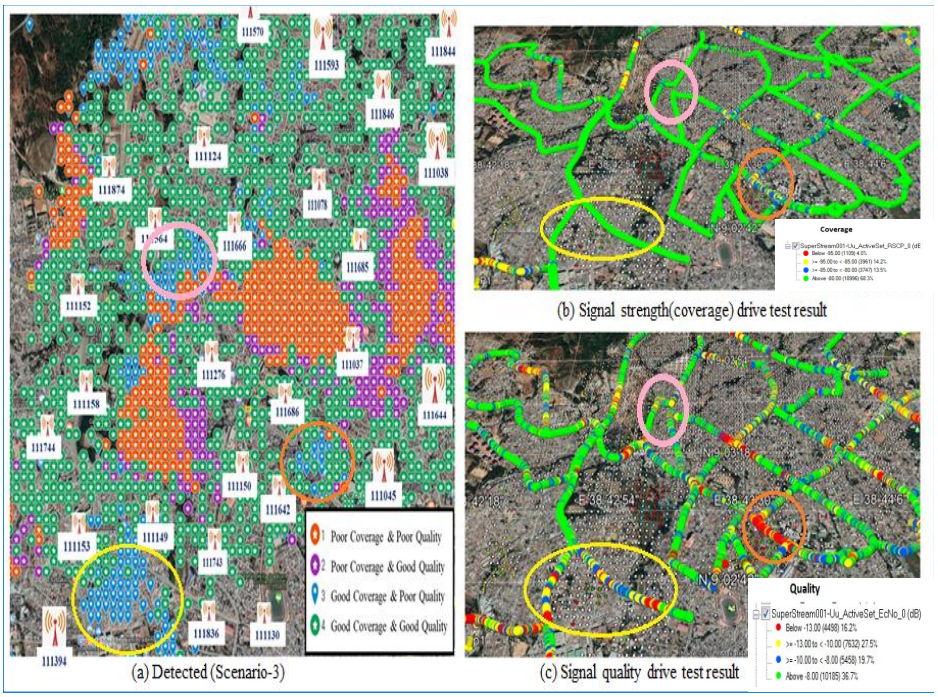


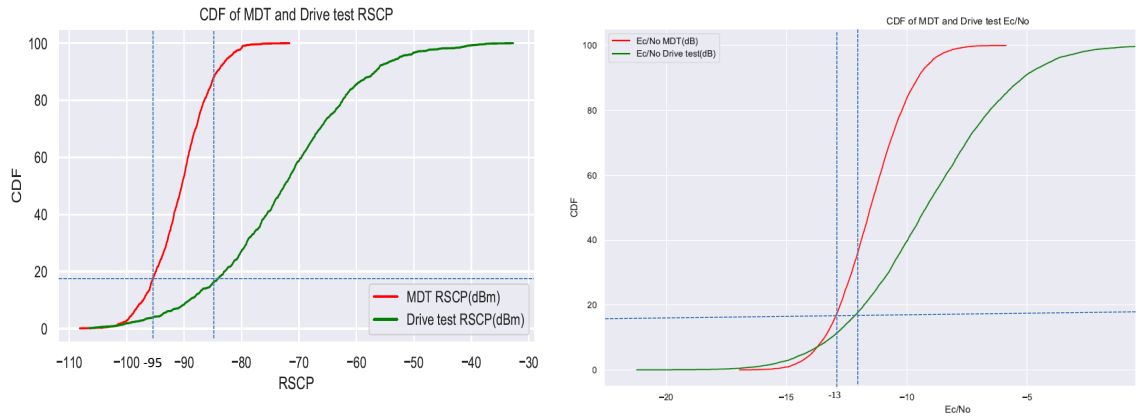
Fig. 5.11 Detected coverage scenario-3 comparison with derive test result

MDT based detected coverage hole also reflected in the drive test result with a slight variation in signal strengths. In addition to this, it can be learned that another coverage scenario which is scenario-1 also observed. Figure 5.11 shows coverage scenario-3 where mobile terminals detect required RSCP but not Ec/No. From the figure it can be observed that, the detected coverage-holes by using the UE MRs data are confirmed, and the same with the result found from drive testing.

To compare the results quantitatively, from Table 5.1 the difference between traces and drive tests reported are of -90.6 dBm and -73.88 dBm for average and -90.58dBm and -73dBm for median respectively. Also, in terms of Ec/No -11.5dB and -9.7dB for average and -11.55dB and -9.37dB for median respectively.

Figure 5.12a and Figure 5.12b illustrates the quantitative comparison of CDFs of MDT and drive tests in terms of RSCP and Ec/No. The CDFs curves have similar trends and they seem shifted from one another. As per Figure 5.12a, higher RSCP values are obtained in drive test case. Considering the CDF value of 18%, in MDT case, it corresponds approximately -95dBm in RSCP value, which indicates that 82% of the possibility for RSCP values are obtained above -95dBm; however, in drive test case, 82% of RSCP values are obtained approximately higher than -88dBm. That is 7dB gain in drive test case. Also, from Figure 5.12b higher Ec/No values are obtained in drive test.

Considering the CDF value of 18%, in MDT case, it corresponds approximately -13dB in Ec/No value, which indicates that 82% of Ec/No values are obtained above -13dB. In drive test case, 82% of the possibility for Ec/No values are obtained approximately higher than -12dB. That is 1dB gain in drive test case.



(a) RSCP CDF of MDT and drive test (b) EC/No CDF of MDT and drive test

Fig. 5.12 CDF of MDT and drive test

Generally, we learned from the validation result that the detected coverage hole and the drive test results are almost the same. A slight variation of the detected signal strength (RSCP) and quality (Ec/No) using MDT data collected and the driving-test based data was observed. This could be happened by different factors. The measurement data taken from the Nastar tool (MDT) has generated from both indoor and outdoor environments. Hence, the reported result might be affected as the data is the averaged result of both factors. Whereas the drive tests result is performed only outdoor. From this and other factors, we expect that the data collected using a Nastar tool is more degraded than that of the driving test due to building penetration effect. So, what we observed from the result has assured the reality, that the signal strength collected by driving test is better than that of the Nastar data that could be because of the mentioned reasons. Another issue to be noted is the drive testing should consider the time that the MDT data used and should be repeatable and if possible, on average to be accurate.

Chapter 6

Conclusion and Feature Work

6.1 Conclusion

In this thesis, the concept of a network coverage hole detection and the challenges of traditional data collection methods for coverage-hole detection were introduced. In addition to this, the knowledge mining technique, specifically the implementation of the DT classifier in processing and analyzing user equipment MR (MDT) data, which is used to optimize the inefficiency of the driving test were introduced. This MDT data were used to construct a model that classifies different scenarios of coverage problems such as "poor coverage and poor quality", "poor coverage but good quality", and "good coverage but poor quality" problems.

The main purpose of implementing MDT data-based coverage scenarios classification was to reduce extra cost and time of the operator by providing instant data collection, reducing human intervention and processing more flexibly and cheaply for network performance assessment. It has the capability to capture the whole coverage data from every geographical location including the UE generated traffic comes from indoor locations. It also provides the basic information help for acquiring the root cause of the problems and so that the optimization team could trigger the optimization process on time.

Drive test data was used for the validation of the result. The results indicated that most of the coverage scenarios were properly classified and so that the proposed method could increase the optimization efficiency in time, cost and accuracy. The proposed method addresses the area where the driving test cannot cover and so that it provides the overall picture of the network coverage status. The methodology could able to

provide the information that drive-test data can do and so that the operators need not go for a drive-test to check their network coverage status. They can exploit the methodology for efficient monitoring of their network through adaptive intelligence of ML capability. The method also easily adaptable for multifeatured scenarios and real-time based coverage hole detection. Additionally, it is also expected that the applied approach can be applicable to other technologies like GSM and LTE for coverage hole detection and other use cases.

6.2 Future Work

The aim of this thesis was only to detect UMTS network coverage-hole using the DT algorithm. In the future, this work can be extended to the diagnosis of the root cause of the detected coverage-hole and also predicting the coverage hole to happen could also another area of focus. Additionally, in this work, the detection considered only one threshold for each parameter (RSCP threshold and Ec/No threshold), so that extra thresholds and features can be considered for farther grading of the network coverage status. Moreover, the traffic dynamics impact on coverage-hole could be investigated in the future by considering none busy hour traffic in addition to busy hour traffic.

References

[1] H. Holma, M. Kristensson, J. Salonen, and A. Toskala, *WCDMA for UMTS: HSPA Evolution and LTE: Fourth Edition*, 2008.

[2] L. HUAWEI TECHNOLOGIES CO., "Network Optimization User Guide," 2013.

[3] W. A. Hapsari, A. Umesh, M. Iwamura, M. Tomala, B. Gyula, and B. Sébire, "Minimization of drive tests solution in 3GPP," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 28–36, 2012.

[4] A. Brandl, "Minimization of Drive Tests (MDT) in Mobile Communication Networks," no. March, pp. 43–49, 2014.

[5] A. Jiso, "Quality of Service Evaluation of Voice over UMTS Network : The case of Addis Ababa Quality of Service Evaluation of Voice over UMTS Network :," pp. 1–98, 2017.

[6] B. B. Haile, D. A. Bulti, and B. M. Zerihun, "On the Relevance of Capacity Enhancing 5G Technologies for Ethiopia," no. June, 2017.

[7] GSMA, "The mobile economy 2019," GSMA, 2019. [Online]. Available: https://www.gsma.com/r/mobileeconomy/

[8] J. Laiho, A. Wacker, and T. Novosad, *Radio network planning and optimisation for UMTS*. Wiley Online Library, 2002, vol. 2.

[9] A. Ahmed, "Ethio telecom 2012 efy (2019/20) quarter i business performance," Ethio Telecom, Oct 2019. [Online]. Available: https://www.ethiotelecom.et

[10] A. Galindo-Serrano, B. Sayrac, S. Ben Jemaa, J. Riihijarvi, and P. Mahonen, "Automated coverage hole detection for cellular networks using radio environment maps," *International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt2013)*, pp. 35–40, 2013.

[11] J. Johansson, W. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3GPP release 11," *IEEE Communications Magazine*, vol. 50, no. 11, pp. 36–43, 2012.

[12] A. Gómez-Andrades, R. Barco, and I. Serrano, "A method of assessment of LTE coverage holes," *Eurasip Journal on Wireless Communications and Networking*, vol. 2016, no. 1, 2016. [Online]. Available: http://dx.doi.org/10.1186/s13638-016-0733-y

[13] T. Specification, "TS 137 320 - V10.1.0 - Universal Mobile Telecommunications System (UMTS); LTE; Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall ," vol. 0, pp. 0–18, 2011.

[14] M. Sousa, A. Martins, and P. Vieira, "Self-Diagnosing Low Coverage and High Interference in 3G/4G Radio Access Networks based on Automatic RF Measurement Extraction," vol. 6, no. Icete, pp. 31–39, 2016.

[15] John Wiley & Sons, *LTE SELF-ORGANISING NETWORKS (SON) LTE SELF-ORGANISING NETWORKS (SON)*, C. S. Seppo Ha¨ma¨la¨inen, Henning Sanneck, Ed. John Wiley & Sons, Ltd Registered, 2012.

[16] J. Moysen and L. Giupponi, "From 4G to 5G : Self-organized Network Management meets Machine Learning," pp. 1–23, 2017.

[17] J. Puttonen, J. Turkka, O. Alanen, and J. Kurjenniemi, "Coverage optimization for minimization of drive tests in LTE with extended RLF reporting," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2010, pp. 1764–1768.

[18] F. Kurt, "On Use of Big Data for Enhancing Network Coverage Analysis," 2013.

[19] A. Galindo-Serrano, B. Sayrac, S. Ben Jemaa, J. Riihijärvi, and P. Mähönen, "Automated coverage hole detection for cellular networks using radio environment maps," in *2013 11th International Symposium and Workshops on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, WiOpt 2013*, 2013, pp. 35–40.

[20] M. Lin, Q. Ye, and Y. Ye, "Graph theory based mobile network insight analysis framework," in *2016 IEEE 7th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2016*, no. October 2016, 2016.

[21] A. Galindo-Serrano, B. Sayrac, S. Ben Jemaa, J. Riihijarvi, and P. Mahonen, "Harvesting MDT data: Radio environment maps for coverage analysis in cellular networks," in *Proceedings of the 2013 8th International Conference on Cognitive Radio Oriented Wireless Networks and Communications, CROWNCOM 2013*, 2013, pp. 37–42.

[22] V. Dalakas, "Automate Minimization of Drive Tests for QoE Provisioning: The Case of Coverage Mapping," *International Journal of Computer Applications*, vol. 126, no. 8, pp. 1–6, 2015.

[23] F. Chernogorov and J. Puttonen, "User satisfaction classification for Minimization of Drive Tests QoS verification," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, pp. 2165–2169, 2013.

[24] F. Chernogorov and T. Nihtilä, "QoS verification for minimization of drive tests in LTE networks," in *IEEE Vehicular Technology Conference*, 2012.

[25] J. Moysen, L. Giupponi, N. Baldo, and J. Mangues-bafalluy, "Predicting QoS in LTE HetNets based on location-independent UE measurements," pp. 124–128, 2015.

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

[27] J. Moysen, L. Giupponi, and J. Mangues-bafalluy, "A Machine Learning enabled network Planning tool," 2016.

[28] O. Osterbo and O. Grondalen, "Benefits of self-organizing networks (SON) for mobile operators," *Journal of Computer Networks and Communications*, vol. 2012, 2012.

[29] T. Specification, "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Performance Management (PM); Performance measurements; Definitions and template (3GPP TS 32.404 version 10.1.0 Release 10)," vol. 10.1.0, pp. 1–30, Oct. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/132400_132499/132404/10.01.00_60/ts_132404v100100p.pdf

[30] M. CHUAH and Q. ZHANG, *Design and Performance of 3G Wireless Networks and Wireless LANs*, 1965, vol. 111, no. 479.

[31] A. A. K. Muhammad Abdur Rahman Haider, Abu Bakar Bhatti, "Radio Resource Management In 3G UMTS Networks," Ph.D. dissertation, 2007.

[32] M. Ajay R, *Fundamentals of Network Planning and Optimisation 2G-3G-4G Evolution to 5G*. India: John Wiley & Sons Ltd, 2018.

[33] A. Jamel, "WCDMA RF Optimization Process GSM-to-UMTS Training Series V1.0," 2009. [Online]. Available: https://www.academia.edu/6103251/WCDMA_RF_Optimization_Process_GSM-to-UMTS_Training_Series_V1.0

[34] J. Isabona and V. M. Srivastava, "Coverage and Link Quality Trends in Suburban Mobile Broadband HSPA Network Environments," *Wireless Personal Communications*, vol. 95, no. 4, pp. 3955–3968, 2017.

[35] I. Zamudio-Castro, S. Vidal-Beltrán, J. Ponce-Rojas, and J. Sosa-Pedroza, "Experimental Analysis of a Node B Coverage Based on the CPICH and Ec/Io Values," *Wireless Engineering and Technology*, vol. 02, no. 01, pp. 23–29, 2011.

[36] M. A. E. Gutiérrez, S. V. Beltrán, and S. J. P. Rojas, "Impact on Quality of Service (QoS) of Third-Generation Networks (WCDMA) with Pilot Signal Pollution," *Procedia Technology*, vol. 7, pp. 46–53, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.protcy.2013.04.006

[37] J. Laiho, *Radio Network Planning and Optimisation for WCDMA*, 2002.

[38] M. Sauter, "Universal Mobile Telecommunications System (UMTS)," *Communication Systems for the Mobile Information Society*, vol. 0, pp. 121–215, 2006.

[39] T. Numatti and S. Saiyod, "Optimization Channel Control Power in Live UMTS Network," pp. 59–64, 2015.

[40] 3GPP, "Universal Mobile Telecommunications System (UMTS); Physical layer – Measurements (FDD) (3G TS 25.215 version 3.1.1 Release 1999)," vol. 0, pp. 0–23, 2002.

[41] V. Buenestado, M. Toril, S. Luna-Ramírez, J. M. Ruiz-Avilés, and A. Mendo, "Self-tuning of remote electrical tilts based on call traces for coverage and capacity optimization in LTE," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4315–4326, 2017.

[42] M. Molinari, M. R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," *C2B(I)D 2015 - Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data, Part of SIGCOMM 2015*, pp. 33–38, 2015.

[43] R. Alexandre and G. Borralho, "Developing a Geo-positioning C # Framework for Radio Network Optimization based on 3G Network Recordings," no. May, pp. 1–10, 2017.

[44] T. P. Description, "TEMS ™ Discovery Network 11 . 0 Technical Product Description," pp. 1–88, 2014.

[45] Telecom Technology Service, "Minimization of Drive Tests (MDT) https://www.ttswireless.com/minimization-of-drive-tests/ Accessed on 7/November/2019."

[46] J. Phuttharak and S. W. Loke, "A Review of Mobile Crowdsourcing Architectures and Challenges: Toward Crowd-Empowered Internet-of-Things," *IEEE Access*, vol. 7, pp. 304–324, 2019.

[47] W. Feng, Z. Yan, H. Zhang, K. Zeng, Y. Xiao, and Y. T. Hou, "A survey on security, privacy, and trust in mobile crowdsourcing," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2971–2992, 2018.

[48] H. T. Co, "(Huawei Industrial Base) Nastar V600R011 Client,Network Optimization User Guide," 2013.

[49] P. Gustås, P. Magnusson, J. Oom, and N. Storm, "Real-time performance monitoring and optimization of cellular systems," *Ericsson Review (English Edition)*, vol. 79, no. 1, pp. 4–13, 2002.

[50] M. Support, "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Network architecture (Release 13)," no. Release 11, 2015. [Online]. Available: http://www.3gpp.org

[51] T. Specification, "TS 132 421 - V6.6.0 - Universal Mobile Telecommunications System (UMTS); Telecommunication management; Subscriber and equipment trace; Trace concepts and requirements (3GPP TS 32.421 version 6.6.0 Release 6)," vol. 0, pp. 0–23, 2005.

[52] S. Hämäläinen, H. Sanneck, and C. Sartori, *LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency*, 2012.

[53] M. Andersson, "Universal Mobile Telecommunications System (UMTS)," vol. 0, p. 27, 2014.

[54] O. M. Lior Rokach, *Data Mining with Decision Trees: Theory and Applications*, H. Bunke and P. S. P. Wang, Eds. World Scientific Publishing Co. Pte. Ltd.

[55] Deparkes, "Machine learning vs rules systems," Nov 2017. [Online]. Available: https://deparkes.co.uk/2017/11/24/machine-learning-vs-rules-systems/

[56] "Ai approaches compared: Rule-based testing vs. learning." [Online]. Available: https://www.tricentis.com/artificial-intelligence-software-testing/ai-approaches-rule-based-testing-vs-learning/#

[57] E. T. Ogidan, K. Dimililer, and Y. K. Ever, "Machine learning for expert systems in data analysis," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Oct 2018, pp. 1–5.

[58] I. Sciences, "Impact of Evaluation Methods on Decision Tree Accuracy Batuhan Baykara," no. April, 2015.

[59] H. Al-Sahaf, Y. Bi, Q. Chen, A. Lensen, Y. Mei, Y. Sun, B. Tran, B. Xue, and M. Zhang, "A survey on evolutionary machine learning," *Journal of the Royal Society of New Zealand*, vol. 49, no. 2, pp. 205–228, 2019.

[60] J. Brownlee, "Master Machine Learning Algorithms Discover How They Work and Implement Them From Scratch," *Machine Learning Mastery With Python*, vol. 1, no. 1, p. 11, 2016. [Online]. Available: http://machinelearningmastery.com

[61] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," vol. 7, no. 2, pp. 5–18, 2017.

[62] D. Kopf, "If you want to upgrade your data analysis skills, which programming language should you learn?" Sep 2017. [Online]. Available: https://qz.com/1063071/the-great-r-versus-python-for-data-science-debate/

[63] P. Ferrari, "When excel doesn't cut it: Using r and python for advanced data tasks," Oct 2018. [Online]. Available: https://www.simplilearn.com/using-r-and-python-for-advanced-data-tasks-article

[64] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining," *International Journal of Computer Applications*, vol. 163, no. 8, pp. 15–19, 2017.

[65] R. A. Devi and K. Nirmala, "Construction of Decision Tree : Attribute Selection Measures," *International Journal of Advancements in Research & Technology*, vol. 2, no. 4, pp. 343–347, 2013. [Online]. Available: http://www.ijoart.org/docs/Construction-of-Decision-Tree--Attribute-Selection-Measures.pdf

[66] N. E. I. Karabadji, H. Seridi, I. Khelf, N. Azizi, and R. Boulkroune, "Improved decision tree construction based on attribute selection and data sampling for fault diagnosis in rotating machines," *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 71–83, 2014.

[67] M. Windows, M. Corporation, K. Hori, and A. Sakajiri, *Fundamentals of Machine Learning for Predictive Data Analysis: Algorithms, Worked Examples, and Case Studies*. MIT, 2015.

[68] T. G. Dietterich, *Machine learning in ecosystem informatics and sustainability*. McGraw-Hill Science/Engineering/Math, 2009.

[69] B. Concepts, "Classification: Basic Concepts and Techniques."

[70] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018. [Online]. Available: https://doi.org/10.1016/j.aci.2018.08.003

[71] G. Al-Naymat, M. Al-Kasassbeh, N. Abu-Samhadanh, and S. Sakr, "Classification of VoIP and non-VoIP traffic using machine learning approaches," *Journal of Theoretical and Applied Information Technology*, vol. 92, no. 2, pp. 403–414, 2016.

[72] M. Services and L. Coverage, "Assessment of Quality of Service," no. June, 2017.

[73] T. V. Wsd, "ECC Report 236," no. May, 2015.

[74] A. Sam *et al.*, "Analysis of quality of service for wcdma network in mwanza, tanzania," 2015.

[75] C. Science and C. Science, "Analysis of Quality of Service for WCDMA Network in Mwanza , Tanzania," vol. 5, no. 9, pp. 18–27, 2015.

[76] Q. I. Group, Engineering Services, "WCDMA Network Planning and Optimization," 2006.

[77] BEREC, "Common Position on information to consumers on mobile coverage," no. December, 2018.

[78] "Self-Diagnosing and Optimization of Low Coverage and High Interference in 3G / 4G Radio Access Networks."

[79] S. Pengpeng, "WCDMA Radio Network Planning and Optimization," QUALCOMM Incorporated, San Diego,USA, Tech. Rep., 2004.

[80] Ethio Telecom, "Low Level Design Documentation for UTRAN (UMTS) Mobile Network in Addis AbabaV14," pp. 1–89, 2014.