ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Cell Outage Detection Through Density-based Local Outlier Data Mining Approach: In case of Ethio telecom UMTS Network

By

Solomon Bekele

Advisor

Dr.-Ing. Dereje Hailemariam

A Thesis Submitted to the School of Electrical and Computer Engineering of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Telecommunication Network Engineering

November 2018

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Cell Outage Detection Through Density-based Local Outlier Data Mining Approach: In case of Ethio telecom UMTS Network

By

Solomon Bekele

Approval by Board of Examiners

_____	_____
Chair Person	Signature
Dr. -Ing. Dereje Hailemariam_____	_____
Advisor	Signature
_____	_____
Examiner	Signature
_____	_____
Examiner	Signature

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.

Solomon Bekele _____

Name Signature

Place: Addis Ababa

Date of Submission: _____

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

Dr. -Ing. Dereje Hailemariam _____

Advisor's Name Signature

Abstract

Mobile traffic growth increases exponentially over the years. To gratify the growing traffic, which requires capacity and coverage, densification of a network is a key solution.

As mobile network becomes larger and larger, it is difficult to manage the network manually rather it requires automated network management. Self-healing is one of self-organizing network (SON's) functionalities that implements automatic fault management in radio access network (RAN). In practice, mobile cell outage is the major problem in the radio access network and leads to the lack of network service. The automated and timely detection of a malfunctioning cell is one of the crucial challenges for network operators.

In this thesis, data mining model has been introduced to detect cell outage automatically. Density-based Local Outlier Factor (LOF) detection algorithm, which is a decisive part of the model, has been adopted and implemented using incoming handover statistical data to detect cell outage and sleeping cells in self-organizing manner. For this purpose, statistical handover data has been collected from real UMTS network and then preprocessed using filtering, aggregation, normalization and then profiling. Moreover, an improved version of LOF algorithm, fast anomaly detection with duplication (FADD), has also been implemented to improve the detection capability.

Receiver Operating Characteristic (ROC) curve is used to show the degree of the performance of the algorithms. The study shows that the two versions of LOF cell outage detectors have detected most cells in outage and locate their positions. But, FADD has detected 89% compared to 75% of the original LOF.

Keywords—*Self-healing, Self-organizing network, Local outlier factor, Receiver operating, characteristic, Sleeping cell, Cell out detection, UMTS, data mining, FADD.*

Acknowledgment

First and foremost, I would like to thank God Almighty for making it possible for me to come this far in my studies.

Secondly, my deepest gratitude goes to my advisor Dr.-Ing Dereje Hailemariam for his invaluable guidance and encouragement in carrying out my thesis work, especially during the first few months. Moreover, I am really grateful for his constant supervision and constructive comments up to the submission of my thesis

I wish to express my sincere thanks to Ethio telecom Engineering and Network Operation Center (NOC) department's staff, especially Ato Ashagrie Getnet, Ato Fisum Mergia, Ato Atnafu Dereje and Ato Frew Katito for their technical guidance and generous support.

My deepest thank goes to my family for their emotional support and encouragement. I also want to thank all the people who have directly or indirectly helped me throughout the course of this thesis work.

Finally, a special acknowledgment extends to Ethio telecom for giving me the opportunity to do my master study and be responsible for my full sponsorship.

Table of Content

List of Figures

List of Tables

List of Acronyms

2G	2nd Generations
3GPP	3rd Generation Partnership Project
AuC	Authentication Centre
AUC	Area under ROC curve **BS**Base station
BTS	Base transceiver station
CAPEX	Capital Expenditure
CBLOF	Cluster Based Local Outlier Factor
CM	Configuration management
CN	Core network
COC	Cell Outage Compensation
COD	Cell Outage Detection
COF	Connectivity based outlier factor
COM	Cell outage Management
CRISP-DM	Cross Industry Standard Process for Data Mining
CS	Circuit switching
DAP	Dynamic Affinity Propagation
DM	Domain Managers CSCircuit switching
DRNC	Drift RNC
EDGE	Enhanced Data for GSM Evolution
EMS	Element Management System
EIR	Equipment Identity Register:
FADD	Fast anomaly detection with duplication
FDD	Frequency Division Duplex
FM	**F**ault management
FMS	Fault management system
FN	False negative
FP	False positive
FPR	False positive rate
GERAN	GSM/EDGE radio access technologies
GGSN	Gateway GPRS Support Node
GMSC	Gateway MSC
GP	Granularity period

GPRS	General Packet Radio Service
GSM	Global Systems for Mobile Communications
H-CRAN	Heterogeneous cloud radio access network
HLR	Home Location Register
HO	Handover
HSDPA	High Speed Downlink Packet Access
HSPA+	High Speed Packet Access evolution plus
IP	Internet Protocol
IMEI	International Mobile Equipment Identity
IMSI	International Mobile Subscriber Identity
INFLO	Influenced outlier-ness
KDD	Knowledge discovery from databases
KNN	K-nearest neighbor
KPI	Key Performance Indicator
LOF	Local outlier Factor
LOFAD	LOF based anomaly detector
LoOP	Local outlier probability
LSH	Local sensitive hashing
LTE	Long Term Evolution
ME	Mobile equipment
MSC	Mobile Services Switching Center
MDT	Minimization of Drive Tests
NGMN	Next Generation Mobile Network
NMS	Network Management System
NOC	Network operation center
OAM	Operation, Administration and Maintenance
OCSVMD	One-class support vector machine-based detector
OPEX	Operational Expenditure
OSS	Operation support system
PAD	Probabilistic anomaly detection
PCA	Principal Component analysis
PM	Performance Management
PS	Packet switching
QoE	Quality of Experience
QoS	Quality of Service

RAN	Radio Access Network
RAT	Radio Access Technology
RNC	Radio network controller
RNS	Radio network subsystems
ROC	Receiver operating characteristic
RRM	Radio Resource Management
RSRP	Reference signal received power
RSRQ	Reference signal received quality
SEMMA	Sample, Explore, Modify, Model, and Assess
SGSN	Serving GPRS support node
SLA	Service level agreement
SM	Security management
SOM	Self-organizing maps
SON	Self-Organizing Networks
SRNC	Serving RNC
SVM	Support vector machine
TDD	Time division duplex
TN	True negative
TP	True positive
TPR	True positive rate
TT	Trouble tickets
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
UNMS	Unified Network Management System
USIM	UMTS Subscriber Identity Modules
UTRAN	UMTS Terrestrial Radio access network
VLR	Visitor Location Register
WCDMA	Wideband code division multiple access

Chapter I

1. Introduction

This chapter provides background of this thesis and describes the motivation, objectives, scope, and contribution of the research. Moreover, it discusses the methodology used and briefly describes reviewed literatures which are related to the study. Finally, the thesis structure is outlined.

1.1. Background

Mobile communication is a rapidly growing technology, which provides ease of access to different services backed by different technologies and seamless connectivity anywhere and anytime [1]. Due to different services offered by telecom operators, the public interest has become higher to get the new services. The number of mobile subscription globally reaches about 5.0 billion in 2017 and will reach 5.9 billion by 2025 [1]. This indicates that communication has become an integral part of the daily life. Hence, for every operator it is a challenge to provide adequate coverage and capacity to its customers that effectively increases an operator's subscriber base and generates more revenues.

To provide adequate capacity and coverage, operators need to increase their mobile network capacity by increasing the number of base stations and deploy the latest technology to accommodate a large volume of traffic, while keeping capital expenditures (CAPEX) and operational expenditures (OPEX) at minimum. Subsequently as the number of base stations becomes larger and new technologies are introduced, network operation and maintenance (O&M) activities become difficult and its OPEX increases significantly. Hence, it is necessary to develop new approach in which mobile system fault and performance management becomes more effective and automated [2]. Self-organizing network (SON) has been

identified for this purpose by 3rd Generation Partnership Project (3GPP) based on the business requirement provided by Next Generation Mobile Networks (NGMN) operators Alliance [3, 4]. Later, the 3GPP also included SON as an important element of the new standard of mobile communications. SON solution is categorized by its functionality as self-configuration, self-optimizations and self-healing in which use cases are developed and organized [4]. R. Barco et al. in [4] described that, SON mainly targets to reduce operational expenses, improve operational efficiency, and enhance and maintain a gratifying user experience, by means of automating tasks that are currently manually performed by highly experienced staff. SON also contributes to CAPEX reduction by a more efficient use of network elements and resources. The SON concept was first introduced by 3GPP as a fundamental element for Long Term Evolution (LTE) deployment in Release 8 [3]. Later, 3GPP further developed Releases 9 and 10[5].

Self-configuration functions focus on defining the configuration parameters of network elements automatically in the planning or deployment stage. Self-optimization functions adjust network parameters, which improve network performance, based on the situation during operation. Self-healing, in simple terms, is an automated fault management which automatically detects any fault occurring in the network, diagnoses those fault to avoid any service breakdown as well as maintaining Service Level Agreements (SLAs), and reduces operational costs of the network [1]. Figure 1.1 shows the self-healing functions.

One of the major mobile network faults is mobile cell outage. Traditional troubleshooting is a manual process which is carried out through alarms and user complaints. Such practice costs much time and effort for cell outage detection (COD) [6]. In order to implement the detection functionality accurately and timely, a number of COD frameworks have been adopted by different researches based on detection indicators such as key performance, cell-level statistical data and location information which can be collected from base stations and user equipment (UE). COD mechanisms generally involved data collection, preprocessing

and analyzing it to extract relevant information so that it can judge if a cell outage occurs. For such process data mining approaches are used to preprocessing and extract knowledge from collected data that comprise of huge amount of information.

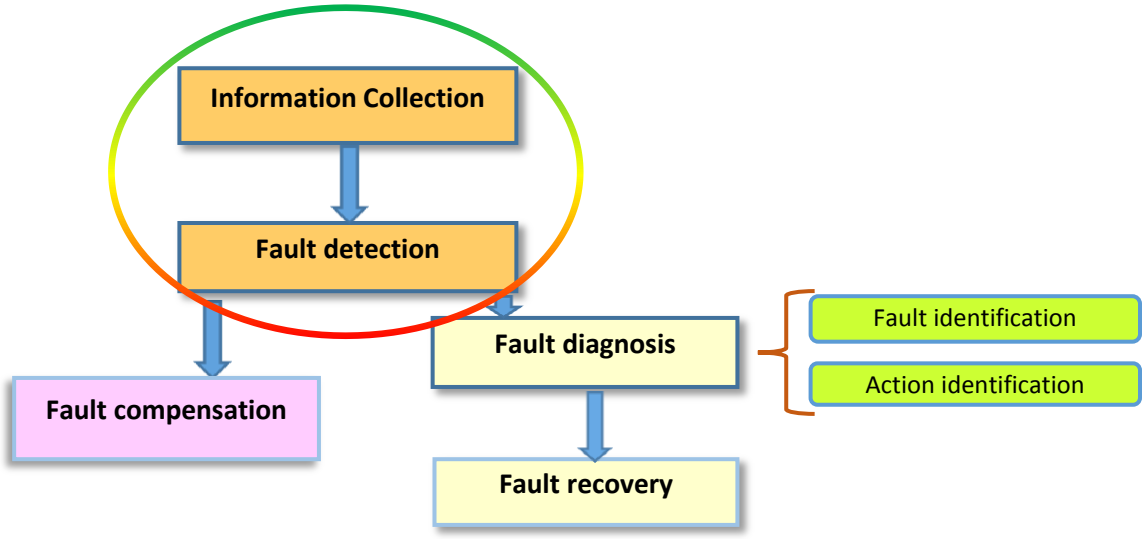

Figure 1. 1: Self-healing functions [4].

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data [7]. It is also popularly referred to as knowledge discovery from database (KDD) which is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams [8]. The abandoned growth of databases in recent years, for example global backbone telecommunication networks carry tens of petabytes (2^{50} bytes) of data traffic every day [7] , brings data mining to the forefront of new business technologies.

In many researches data mining techniques such as classification, clustering and outlier detection methods were used for network fault detection. One of the outlier detection methods in data mining is LOF that calculate the degree of being an outlier for each instance based on the local density around it. Detail description of LOF is presented in section 5.3,

since it is related to the scope and concern of this thesis. But, other data mining methods are also briefly described in section 4.2.

The motivation of this thesis work is to adapt mobile cell outage detection model using LOF data mining technique based on incoming handover statistical data of Universal Mobile Telecommunication System (UMTS). The thesis work also focuses on how handover (HO) data is collected and preprocessed to detect and localize outage cells. Moreover, it evaluates the performance of the detection system.

1.2. Statement of the Problem

Ethio telecom's mobile network consists of Global systems for Mobile Communications GSM, UMTS throughout the country and LTE in central part of Addis Ababa. These technologies have been deployed by different vendors. As per Ethio telecom website recent Press Release, Ethio telecom is now the largest mobile operator in Africa in terms of subscriptions, with 57.34 million mobile subscribers in 2017 [9].

Ethio telecom mobile network is generally characterized by multi radio access technology (multi-RAT), multi-vendor, and multi-layer. The deployment of various technologies and increase number of mobile base stations in the network usually creates complexity and introduces challenges in the planning, optimization and O&M activities of the network. The traditional network management activities result in longer execution time, more OPEX, and prone to human error [4, 5].

One of the categories of network management that detects, isolates, and correct malfunctions is fault management (FM) system. FM is used for the detection of faulty network element and generation, collection and presentation of alarms. In Ethio telecom, network FM system process involves teams from Network Operation Center (NOC), and O&M departments to classify and recover the fault in order to ensure reliable services. Root-

cause analysis (diagnosis) of faults is done manually that makes the management very labor-intensive and expensive as well as requires adequate expertise.

As per six month collected data and its subsequent report from Ethio telecom monitoring section, 57% out of a total number of telecom network faults for which Trouble tickets (TTs) have been generated is related to mobile network; especially, mobile radio access network (RAN) [10]. It is a significant percentage that affects the network performance, availability and customer quality of experience (QoE). Moreover, based on the above report, average resolution time for mobile related faults is about 31 hours per single fault/TT (fault clearing may range from a minimum of few minutes to a maximum of a number of days or weeks depend on the severity of a fault and its recovery effort), which has negative impact on quality of service (QoS) and company revenue generation.

One of the possible mobile radio access network problems for an operator is a cell outage. Several researches about COD which are based on the analysis of the performance of the problematic cell or its neighboring cells have been presented. However, with these methods only the most severe cases can be detected but many other outage situations may not be identified. For instance, sleeping cell, which is a specific type of cell outage that can occur in the network, is not identified by the fault management system. This is because sleeping cell is invisible for network operators via traditional alarms [11]. This means that FM system does not generate alarms based on network traffic. Sleeping cell problem can only be identified from customers complains or network performance evaluation.

This thesis focuses on automatic detection of cell outage that is the major type of fault in the radio access part of the UMTS mobile network. The detection is based on HO statistics data to reduce resolution time in fault handling process.

1.3. Objectives

1.3.1. General Objective

The main objective of this thesis is to investigate a COD model using density-based LOF data mining technique based on HO statistical data collected from Ethio telecom UMTS network and evaluate the performance of the detection scheme.

1.3.2. Specific Objectives

The specific objectives to be accomplished in this thesis are:

- Review related literatures and understand COD methods.
- Study the various Local Outlier data mining techniques and identify the available HO statistical data to come up with a COD model for the thesis work.
- Collect incoming HO statistical data and preprocess it for analysis by transforming the data into a format suitable for the proposed data mining detection algorithm.
- Develop a Matlab program for COD algorithm and test it by altering the control parameters to get best performance.
- Evaluate the performance of the detection algorithms using Receiver Operating Characteristic (ROC) curve.
- Discuss the results and draw recommendations based on the findings.

1.4. Scope and Limitation of the Thesis

This thesis addresses one of the use cases of self-healing which is cell outage detection. The scope of the thesis is to investigate cell outage detection model based on incoming HO statistical data of Ethio telecom UMTS mobile network using outlier anomaly detection algorithm. Even though cell outage detection concept is the same for all mobile

technologies, the thesis work is restricted to only UMTS network. Moreover, central part of Addis Ababa UMTS Node-B sites is selected for analysis. Due to time limitation and complex matters, the thesis is limited to cell outage detection only rather than diagnosis the root cause of the problem. Hence, further research can be conducted to investigate the cause of the cell outage.

1.5. Literature Review

SON is often used to categorize a mobile network for which the activities of configuring, operating, fault handling and optimizing are largely automated. It is a collection of functions for automatic configuration, optimization, and healing of mobile networks. Self-healing is one of the functionalities of SON that enables the mobile network to automatically perform the task of troubleshooting which includes detection, diagnosis, and correction of faults [12]. Cell outage Management (COM) is the main task of self-healing and it contains COD and cell outage compensation (COC).

There are a lot of researches conducted on the detection and diagnosis of faults in mobile networks, especially on radio access part of the network. Many detection methods use input data that collected from minimize drive testing (MDT), UE or network management system (NMS). This section presents a brief survey of basic literature on COD. But more of other relevant literatures are reviewed in Chapter three.

In some study Key Performance Indicators (KPIs) are used for Cell outage detection. Ahmed Zoha et al. in [5] proposed COD framework that adopts a model driven approach that makes use of mobile terminal-assisted data collection solution based on MDT functionality. They first collected UE reported MDT measurements and extracted minimal KPI representation by projecting them to a low-dimensional embedding space. They used two kind of anomaly detection methods namely local outlier factor-based detector (LOFD) and one-class support vector machine-based detector (OCSVMD) together with the embedded

measurements. The two learning algorithms were compared and evaluated. Moreover, the geo-location associated with each measurement of COD framework was used to localize the position of the faulty cell. Full dynamic LTE simulation tool was utilized to simulate the LTE network consists of 27 e-NodeBs and to test the detection performance of the OCSVMD and LOFD. The result shows that OCSVMD better to identify abnormal measurement than LOFD.

In the study presented by Szilágyi and Nováczki [13], integrated detection and diagnosis framework is presented that can perform fault classification based on statistical analysis and find the most probable root cause of problems. For detection, monitored radio measurements and other KPIs were compared to their usual behavior captured by profiles automatically without threshold and manual setting. But, diagnosis is depending on previous fault cases. The abnormality level was used to calculate the likelihood of a failure case. The target with largest likelihood value is considered to be the diagnosed failure.

Other detection approach is proposed by S. Rezaei & H. Radmanesh [14]. In this paper, an automatic unified detection and diagnosis framework has been presented using unsupervised clustering of both traffic and signaling KPIs for diagnosis. Moreover, experts reasoning also incorporated in the design to enable automatic decision support in an operating mobile communications environment for fault diagnosis. They considered the real data of a live GSM network for performance evaluation.

In another study in [15] by I. De-La-Bandera et al., HO statistics has been used for cell outage detection. The proposed Cell Outage Model includes different cases of cell outage. First, cell outage that does not affect the eNB in which eNB generates KPIs from the cell in outage. Second is site outage which affects eNB and there is no KPIs available from the Operation support system (OSS). The third case is when Cell is not in outage, but there is a failure in eNB-OSS connection. LTE simulator test results indicates that algorithm enables to detect a cell outage when KPIs from the cell are either available or not. The drawback of the

algorithm is that cell in outage with very low traffic cannot be detected.

There are also researches based on data mining and machine learning techniques. Automated network troubleshooting using data mining is presented by E. Rozaki [16]. A monitoring scheme is proposed for mobile networks based on the use of rules and decision tree data mining classifiers. The goal of the study is to improve anomaly detection and fault localization based on a top-down (Bayesian networks) model. The data mining techniques was used to train a system to learn network fault rules.

T. Zhang et al. in [17] proposed COD architecture based on the handover statistical data to detect small cell outage in heterogeneous network (HetNets). They use data mining methods and preprocessed sequential HO data spatially and temporally. Detection algorithms performance is evaluated using their own designed simulator with some reasonable assumptions. The results of simulation show that their system is more effective to detect small cell outage in comparison to the model using MDT measurements.

In a study by Y. Ma, et al. in [18], an unsupervised data mining algorithm called Dynamic Affinity Propagation (DAP) clustering which uses reference signal received power (RSRP) and reference signal received quality (RSRQ) as input data from UEs, eNodeBs and OAM to detect cell outages was introduced. The LTE-Advanced simulation environment is used to test the proposed algorithm and cell outages are successfully detected.

1.6. Methodology

In order to have a better understanding of this thesis work, different related literatures, journals, and books on COD and UMTS network have been reviewed. Relevant materials on detection algorithms of data mining techniques are also referred. Moreover, Ethio telecom process manuals and network design documents are also consulted to understand the UMTS network management.

This thesis was performed by using the following methods:

- The thesis work started with a literature review to understand more the purpose of the research and then familiarize with the available related research works.
- Necessary data for the study has been identified.
- COD model has been proposed.
- In ordered to accomplish the research process, COD algorithm Matlab program has been developed.
- A real Ethio telecom UMTS network scenario has been considered and the required data has also been collected in collaboration with domain experts.
- Data has been preprocessed using MS-Excel and detection system analyzed its output to detect outage cells.
- Finally, the results have been discussed and published in the form of a final thesis paper.

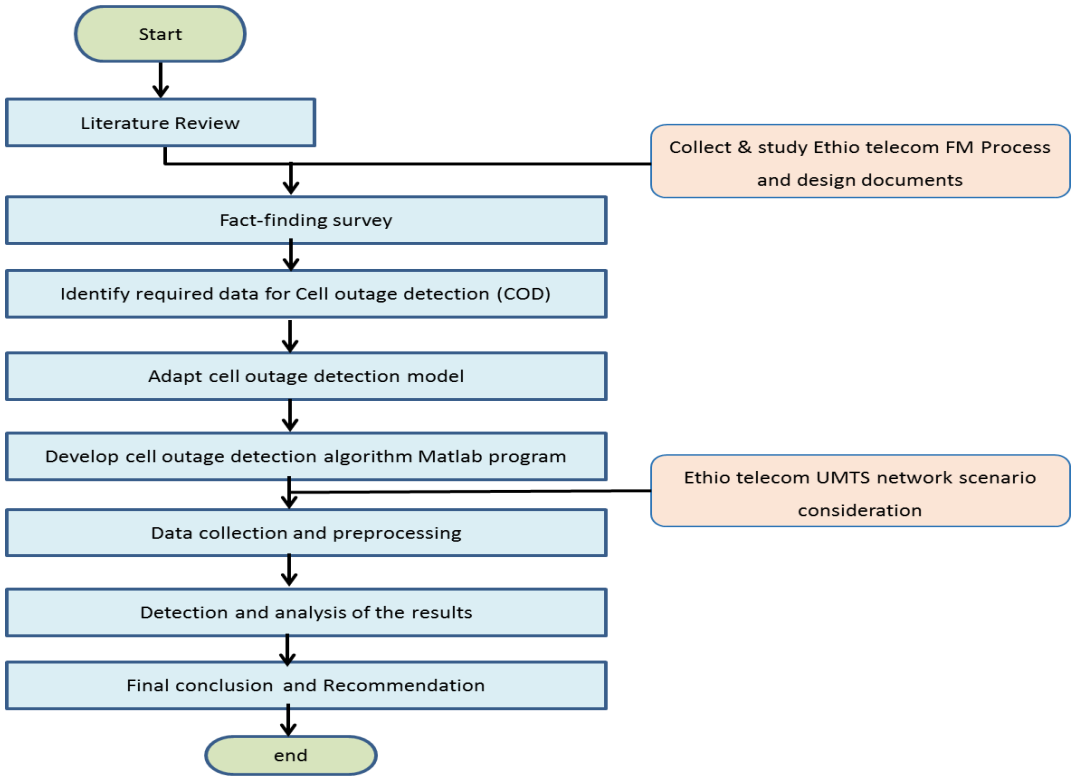

Figure 1. 2: Research framework.

1.7. Contribution of the Thesis

Most COD algorithms in the literatures used KPIs, alarms and signal parameter measurements for detecting cell outage. This only feasible as long as KPIs are available from Node-B of the cell in outage, but there are circumstances that affect Node-B. This thesis work contributes to overcome such problem by introducing COD which is based on number of incoming HO statistical data from neighboring Node-B cells that can tell the status of the cell in outage. Moreover in fault management perspective, it can be realized as a primary sleeping cell detection scheme in support of Ethio telecom fault management system.

Generally, this thesis work is important since its application can reduce the revenue losses of an operator due to longer cell outage period. Finally, this thesis can be an input for further research in the area of cell outage management (COM). Moreover, the output of the thesis also contributes to understand how huge and complex telecom data can be managed in a way to extract knowledge to solve problems.

1.8. Thesis Structure

Chapter 1 contains an introduction, problem statement, objectives and scope of the work and the thesis work approach. Chapter 2 presents a brief overview of UMTS network architecture and fundamental concepts of HO in UMTS as well as network management. Chapter 3 describes cell outage analysis techniques and discusses COD based on incoming HO statistics data. Brief description of data mining is presented in chapter 4. Chapter 5 is devoted to the proposed LOF and Fast Anomaly Detection given Duplications (FADD) algorithms. System model, experimental real network scenario, list of assumptions and analysis of the experimental results as well as evaluation of the proposed COD algorithms are presented in chapter 6. Finally, Chapter 7 summarizes the main conclusions of the thesis work and presents future lines of action.

Chapter II

2. Handover in UMTS

2.1. Overview

GSM specifications started with an objective of achieving a European mobile radio network in 1982 [19]. The work on the specification was continued until 1990 and then frozen with the development of the 2nd generation cellular wireless system (2G) which enabled voice traffic to go wireless. Then, the 3rd Generation Partnership Project (3GPP) finalized the first version 3rd generation mobile communication system which was known as UMTS following GSM. Wideband Code Division Multiple Access (WCDMA) is the main radio access technology that is used by the UMTS network in the air interface, and deployment has been started in Europe and Asia, including Japan and Korea, in the same frequency band, around 2 GHz [20]. WCDMA offers wider bandwidth and a different control and signaling channels [21].

To improve UMTS network performance for further mobile communication development towards LTE, new 3GPP framework was introduced with the concept of High Speed Downlink Packet Access (HSDPA) in Release 5 specification [21]. In order to make UMTS system capable of handling High Speed Packet Access and IP traffic in Core network (CN), High Speed Packet Access evolution (HSPA+) was introduces in Release 7 and beyond to transform UMTS to packet switched technology architecture [22].

A new standard called LTE is being deployed to cope with the enormous demand of capacity and services. The introduction of LTE defiantly improved system performance, higher data rates and spectral efficiency, reduced latency and power consumption,

enhanced flexibility of spectral usage and simplified network architecture [22]. LTE is specified with the intention to provide IP based packet switching and transmission solution in a network architecture. LTE has a capability to operate in Frequency Division Duplex (FDD) or Time Division Duplex (TDD).

2.1. UMTS Network Architecture

UMTS uses the same well-known architecture that has been used by 2nd generation systems [20]. It consists of a number of logical network elements that each has a defined functionality. In the standards, the network elements can be grouped based on similar functionality, or based on which sub-network they belong to. Such sub-systems in UMTS are UMTS Terrestrial Radio access network (UTRAN) which is in charge of handling all radio-related functionality, and the CN which is responsible for switching and routing calls and data connections to external networks. The UE is considered as part of UMTS sub-system that interfaces the radio access network with the user. The UMTS system architecture in Figure 2.1 below illustrates the network elements and interfaces of the UMTS architecture (Release 99).

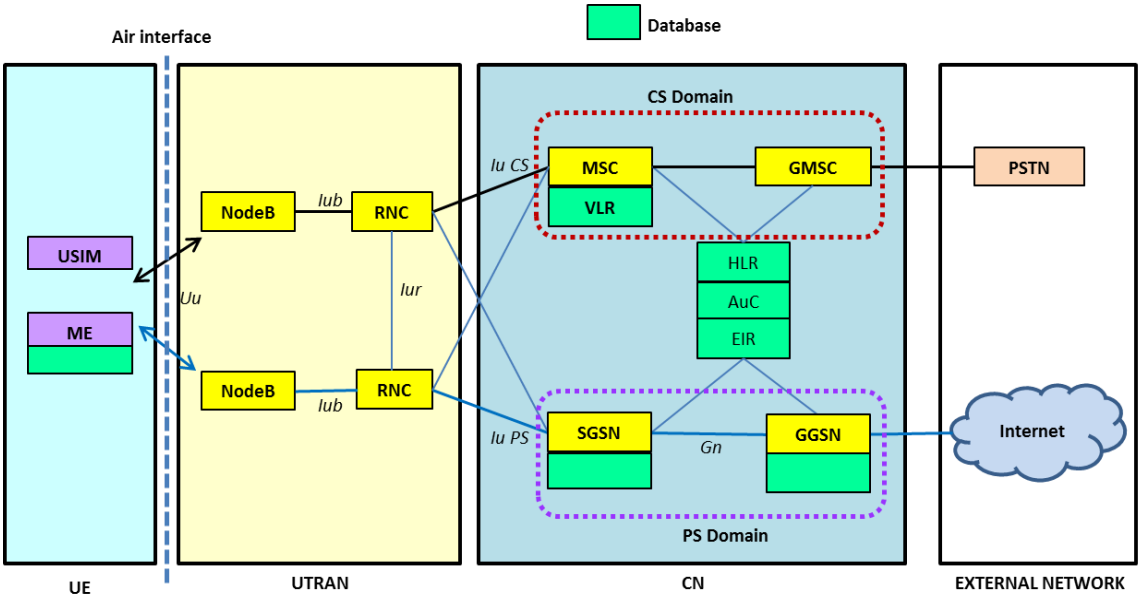

Figure 2. 1: UMTS architecture (Rel.99) [23].

The basic modulation technique in UMTS is Wideband Code Division Multiple Access (W-CDMA) but interworking with other technologies such as wireless local area networks and GSM/EDGE radio access technologies (GERAN) is also possible

2.2.1. User Equipment

The UE, as defined in [23] *[3G TS 23.002 V3.6.0 (2002-09)]* is the mobile equipment with one or several UMTS Subscriber Identity Modules (USIMs). The UE consists of two parts:

- The Mobile Equipment (ME) is the radio terminal that interfaces with the user.
- USIM is a smartcard that holds the subscriber identity, performs authentication algorithms, and stores authentication and encryption keys and some subscription information that is needed at the terminal.

2.2.2. UMTS Terrestrial Radio Access Network

The radio access network that is used in 3rd generation mobile network is generically known as 'Universal Terrestrial Radio Access Network (UTRAN). UTRAN is a logical grouping that includes two distinct elements [19]. One is Node-B that converts the data flow between the Iub and Uu interfaces and the second one is Radio Network Controller (RNC) which is the network element responsible for most of the Radio Resource Management in the UTRAN. Two main component of UTRAN are briefly described below.

- **Node-B:** The Node-B is similar in functionality to the BTS in GSM networks. It is the 3GPP term within UMTS to denote the base station transceiver. It is part of the UTRAN and contains the transmitter and receiver to provide the radio link between the UE within the cell and the UMTS network. The main function of Node-B is performing air interface processing such as channel coding and interleaving, rate adaptation and spreading [20]. It also collaborates with the RNC in the resource

management. The logical interface between UE and Node-B is specified by 3GPP and known as Uu interface. Node-B is connected to and controlled by RNC through the Iub interface which physically corresponds to a transmission link.

- **RNC:** The RNC is the UTRAN radio network subsystem responsible for most of the Radio Resource Management (RRM), some of the mobility management, softer HO (a special handover type when the UE is connected to cells belonging to the same node B) and controlling functions [21] such as congestion control, admission control and code allocation for the cells served by the controlled Node B's. It controls Node-Bs that are connected to it and used as access point for all services that UTRAN provides. It also performs data encryption / decryption to protect the user data from eavesdropping [21].

  The RNC takes care of HO decisions requiring signaling to the UE. In order to facilitate effective HO between Node-Bs under the control of different RNCs, the RNC communicates with the CN and neighboring RNC. RNCs can interconnect each other through the Iur logical interface. This can be conveyed over a direct physical connection between RNCs or through any appropriate logical transport network [20].

  A single Node-B can be connected to and controlled by two RNCs, RNC that has the active connection to the CN through the Iu-CS or Iu-PS interface is known as the Serving RNC (SRNC) and the other RNC (the one that just routes information) is known as Drift RNC (DRNC) [20].

As shown in Figure 2.2, UTRAN comprises of Radio Network Subsystems (RNSs) which communicate with the CN through the Iu interface. In turn a RNS contains a RNC and one or more Node-B which represent the UMTS base stations [19]. Within UTRAN, the RNCs communicate with each other over Iur interface and a Node-B connects to the RNC through the Iub interface. UTRAN can support either FDD or TDD or combined dual mode

operation. The UMTS network also supports both circuit switched and packet switched connections.

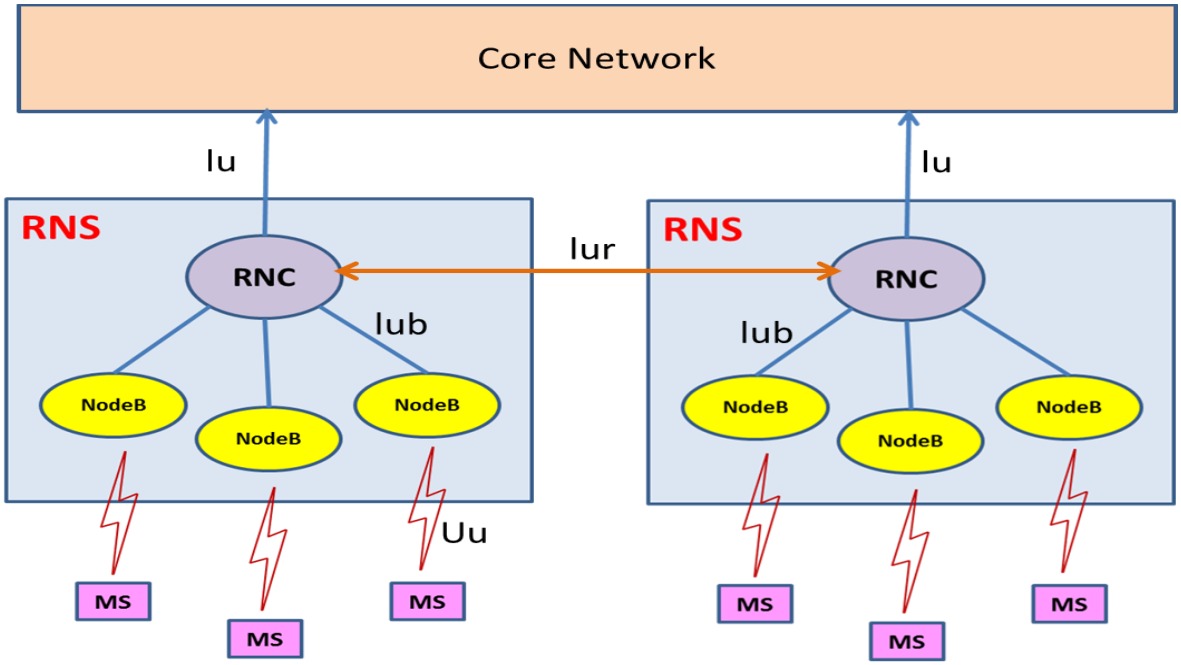

Figure 2. 2: UTRAN architecture [24].

2.2.3. Core Network

CN is the backbone of the mobile communication system and responsible for transport functions such as switching calls, routing data connections to external networks and tracking mobile users. CN function includes the transport functions, mobility management, handling subscriber databases with their information, service controlling functions, billing, etc. [19]. It is the "all-in-one" network that connects the UEs of a given UTRAN with other mobile or fixed users of other Networks. The main components of the CN (Rel-99) are:

- **HLR:** The Home Location Register is a database that contains the master copy of the user's service profile.
- **MSC/VLR:** The Mobile Services Switching Center (MSC) and Visitor Location Register (VLR) are the switches and database component of CN respectively.

- **GMSC:** The Gateway MSC is a MSC which interfaces with external circuit switched networks.
- **SGSN:** It is used to deliver data packets to and from the UE and can store user location information that recognizes the origination and termination point for a packet data request within its service area [21]. It is the counterpart of the MSC/VLR but in the packet switched domain.
- **GGSN:** The Gateway GPRS Support Node is the counterpart of the GMSC but in the packet switched domain.
- **The Authentication Centre (AuC):** AuC is defined in [23, Release 99] as it is an entity which stores data for each mobile subscriber to permit the International Mobile Subscriber Identity (IMSI) to be authenticated and allow communication over the radio path between the mobile station and the network to be ciphered.
- **The Equipment Identity Register (EIR):** It is the logical entity in the UMTS system which is responsible for storing the International Mobile Equipment Identities (IMEIs) in the network. The mobile equipment is classified as "white listed", "grey listed", "black listed" or it may be unknown as specified in TS 22.016 and TS 29.002 and may be stored in three separate lists [23].
- **External Networks:** These are the external networks that the UEs can communicate with and that the mobile network has to be connected to. It can be for example the public telephone network or any other GSM or UMTS network.

2.2.4. UTRAN Interfaces

In comparison to the GSM radio access network, several new interfaces such as Uu, Iu, Iub and Iur in the UTRAN as well as some new interfaces in the CN like Iu-PS and Iu-CS have been introduced. These interfaces are used to connect the logical network elements in the UMTS system and describe the mode of operation while internetworking with other network entities [21, 24]. UMTS Interfaces are described as follows:

- **Cu Interface** is an electrical interface which operates between the Universal Subscriber Identity Module (USIM) and the UE.
- **Uu Interface** is used by the UE to access the UTRAN. This is a typical UMTS open WCDMA radio interface through which mobile stations (MSs) are used to send and receive communication requests.
- **Iub** Interface is an open UMTS interface which links an RNC to a Node-B(s)
- **Iur** Interface is an over-the-air UMTS interface and it links different RNCs. This interface facilitates soft HOs in the presence of RNCs manufactured by different venders.
- **Iu** Interface is used to connect the RNC to the CN. This interface comes with two variants, Iu-CS and Iu-PS to facilitate a user's initiated requests for circuit and packet domains respectively.
    - **Iu–CS** interface for circuit-switched traffic, based on the ATM transport protocol.
    - **Iu–PS** interface for packet-switched traffic, based most likely on IP over ATM.

2.2. Handover in UMTS

Mobile stations in a cell border that are transmitting at their maximum power cannot increase their power levels. Hence, they may get disconnected if they move far from the serving base station further unless a HO takes place to other neighboring cell. To make a mobile station to be continuously connected to a mobile network when a mobile station user travels from one area of coverage or cell to another cell within call duration, HO should be initiated. This kind of HO can be related to mobility. Another condition that triggers HO is the malfunctioning or performance degradation of a serving cell. Due to this, mobile users who were previously connected to a current faulty cell are forced to connect to a nearby cell

with better signal level among other neighboring cells. Both mobility and malfunctioning cell can be cause of HO due to received signal strength level drops below a certain threshold value set by an operator. The strength of receiving signal is influenced by fading due to shadowing, and destructive interference (reflection, refraction, and scattering at small obstacles). Fading is a variation of the attenuation of a signal with various variables.

There are two main categories of HOs which are classified as inter-cell and intra-cell HOs [25]. The purpose of inter-cell HO is to keep the signal quality and coverage when the user moves to a new cell area whereas the purpose of intra-cell HOs is to change one channel inside the existing cell, due to fading or interference, to a new channel with better conditions.

Basic types of UMTS handover are hard handover, soft handover and softer handover. Other special kind of HO called inter-RAT handover is used between UMTS and GSM radio access technologies [26].

In this thesis, HO statistical data is the main input for the cell outage detection algorithm. Normally, the KPIs commonly monitored for HO performance includes number of HO attempt, HO success rate, HO failure rate, inter-cell HO, inter and intra-Node-B HO, number of successful and failed HO etc.. Hence, a brief description of the basic types of HOs is essential. The following sub-sections explain the types of HOs.

2.3.1. **Hard Handover**

As the name indicates, hard HO is a hard change of connection. It is essentially the same as that used for 2G networks where one link is broken and another established. It means that all the old radio links in the UE are removed before the new radio links are established (break-before-make) [25]. Hard HO can be seamless or non-seamless depend on its noticeability to the user call. If a HO requires change in carrier frequency (i.e., inter-

frequency HO), then it is always categorized as hard HO.

2.3.2. **Soft Handover**

Soft HO means that the radio links are added and removed in a way that the UE always keeps at least one radio link to the UTRAN (make-before-break). It occurs when a UE is in the overlapping coverage area of two cells. Soft HO is performed by means of macro diversity, in which several radio links are active at the same time. To implement soft HO it required that source and target cells need to operate on the same frequency. Figure 2.4 shows hard HO and soft HO scenarios.

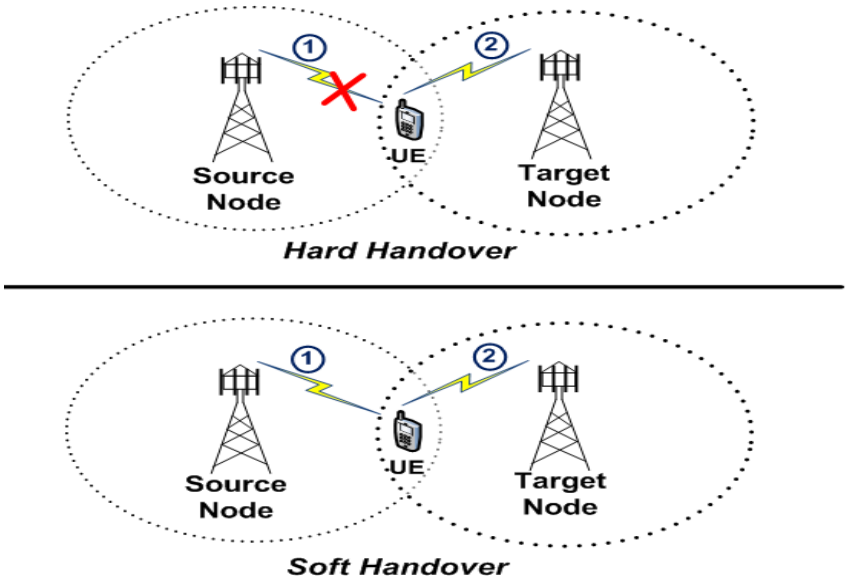

Figure 2. 3: Hard and soft HO types [25].

2.3.3. **Softer Handover**

Softer HO is special cases of soft HO where the radio links that are added and removed belong to the same Node-B (i.e. this occurs when several sectors may be served from the same Node-B) [25]. During softer HO, a UE is in the overlapping cell coverage area of two

adjacent sectors of a base station and it make concurrent communication with two air interface channels (one for each sector that carries two separate codes in the downlink direction). In softer HO, macro diversity with maximum ratio combining can be performed in the Node-B, whereas in soft HO on the downlink, macro diversity with selection combining is applied.

2.3.4. **Inter-RAT/Intersystem Handover**

In many situations UMTS and 2G GSM networks technology work together to provide services. In such cases it is necessary for the UMTS radio access network to HO to the 2G GSM network and visa verse. Such kind of HO is called Inter-RAT handover or Inter-system handover [26]. Inter-RAT handover is the most common HO that occurs between UMTS and GSM. The two types of inter-RAT handovers are UMTS to GSM handover and from GSM to UMTS handover. The HO from GSM to UMTS occurs to provide an improvement in performance.

2.3. Mobile Network Management System

The purpose of a NMS is to act as a human interface with a network for a network operator that maintains and operates the network [27]. To see the statues of the network remotely or centrally, NMS is essential for network quality maintenance activities. The four main functions of NMS are FM, configuration management (CM), performance management (PM) and security management (SM) [29].

3GPP in [3GP11e] standardized mobile management architecture based on strong hierarchical structure of the 3GPP network architecture in which several network elements are combined and managed through an EMS. In this network management architecture, a number of EMS together also managed by the so called Domain Managers (DMs). Central

NMS is the one responsible and controls the overall network at the higher hierarchic [28]. This kind of hierarchical deployment structure helps to manage different vendor's networks through a single NMS called unified NMS (UNMS).

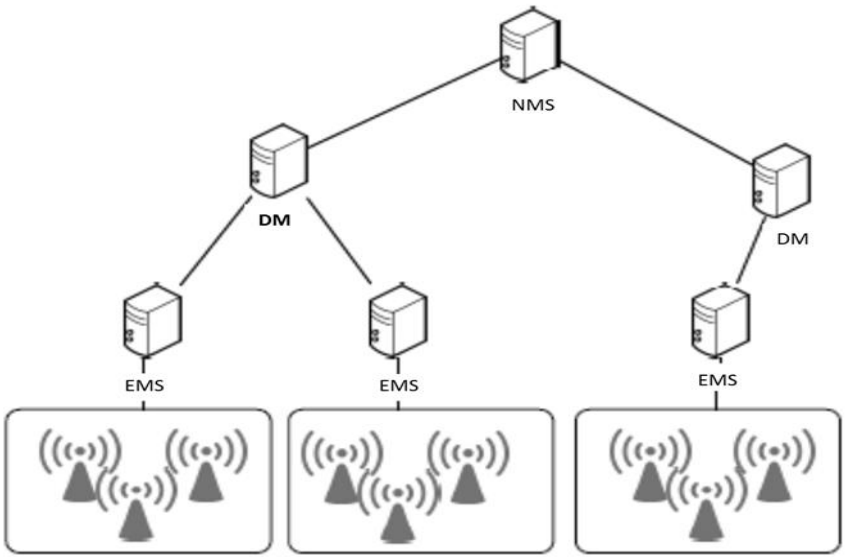

Figure 2. 4**:** 3GPP Network Management Architecture **[28]**.

FM system is one of the components of NMS that handles fault related issues. In a traditional network management, in order to detect network failures, it is important to know the state of the overall network and its individual network elements. Each of the network elements provides counters or performance measurement, which are used to represent its state. Instead of using raw performance measurement, KPIs that combined individual measurement values into more meaningful abstract values, are used to detect network element faults [28]. If the observed KPI crossed the predefined threshold value, alarm will be generated to indicate NE fault. The root cause analysis and interpretation of data for an observed behavior is complex so that the involvement of human operator who converts the available data into context.

Chapter III

3. Cell Outage Detection Techniques

3.1. Introduction

As the number of network elements in a mobile network increase, the probability of network outages increases proportionally. In order to address these network outages in a more proactive manner, 3GPP defined a 3 stage FM framework called self-healing. The first step in self-healing is to automatically detect a fault when a fault occurs in the network. Traditional detection of the fault can be done through measurement of KPIs and estimating or analyze of their values [13]. Fault diagnosis task, which is the second component of self-healing, involves analysis and determination of root causes of the problem. For such purposes, most diagnosis approaches currently depend on manual process which is time consuming and involves domain experts to perform a diagnosis task [4]. Corrective action is the third step in which fault is cleared from the network. This thesis focuses on the first stages of self-healing systems, which is specifically automatic cell outage detection in UMTS network.

A cell is closely related with a base station (BS), which is a radio tower creating coverage on a site by emitting radio signal [29]. The usual practice of mobile site deployment is that a BS's surrounding coverage is divided into three sectors of 120° each. Then, a single sector can be considered as a cell.

Automatic detection of malfunctioning cells is part of the self-healing mechanism and has been studied intensively [5, 11, 13, 15]. Generally, cell outage takes place due to multiple reasons: hardware or software failures, external failures of power supply, erroneous

configuration or even environmental changes. In this situation, cell cannot carry traffic and network operator needs to solve the problem sooner. The traditional detection methods are mostly manual that might takes longer period to detect the problem. With the increase in magnitude and complexity in the current and future mobile network, it is difficult to manage cell outage detection in a usual way.

Automated cells outage detection is also used for management of sleeping cell, which is a specific type of problem that can occur in the network. A sleeping cell is a special case of cell outage, which makes mobile service unavailable for users, but invisible for network operators via traditional alarms [11]. Generally, cell outage or sleeping cells can be classified into three groups [30]:

- *Impaired sleeping cell*: When a cell still carries traffic, but certain performance characteristics are slightly lower than expected..
- *Crippled sleeping cell***:** when a cell has severe degradation in its capacity due to a significant failure of a base station component.
- *Catatonic sleeping cell*: when a cell is completely out of service. This kind of sleeping cell problem leads to absence of services in the faulty cell area since it does not carry any traffic. Such kind of cell needs a timely detection and recovery action.

3.2. Cell Outage Detection Approaches

In regards to cell outage detection, several studies have been conducted to address the outage issue. Various solutions are also provided to effectively detect the outage with different approaches. Past studies used different detection algorithms beginning from KPI threshold comparisons all the way to more complex and advanced data mining techniques.

In most cases, the COD algorithm monitors and analyses KPIs and alarms reported by cells to determine cell outage [31]. But, some of the COD algorithms use statistical data to be

analyzed for the detection purpose. The main categories of machine learning algorithms are supervised, and unsupervised.

3.2.1. Supervised Learning

Supervised learning is a type of learning that requires a supervisor in order for the algorithms to learn their parameters. These algorithms are given a set of data which contains both input and output information with which learned model is determined. Supervised learning consists of several learning algorithms that each has its own application. K-nearest neighbor (KNN) and Support Vector Machine (SVM) are frequently used as supervised learning detection algorithms that apply in a number of cell outage detection studies. KNN algorithm is among the simplest of all machine learning algorithms in which an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors [32]. If k = 1, then the object is simply assigned to the class of its nearest neighbor. SVM is also a supervised machine learning algorithm that can be employed for both classification and regression purposes. SVMs are more commonly used in classification problems. Chernov et al., in [33], presents a data mining framework for the detection of sleeping cells. Its algorithm collects user data, preprocesses it and then performs, first, anomaly detection algorithm using KNN. Secondly the algorithm assigns sleeping cell scores to each cell and the higher score cells are considered as sleeping state. In [34], the authors consider a network scenario which has separate control and data planes. To perform cell outage detection, they used two distinct algorithms which are K-NN and Local Outlier Factor based Anomaly Detector (LOFAD). In [35], other authors also build a detection method based on K-NN and consider a heterogeneous network scenario which consists of macro-cells and pico-cells. Outage cell and neighbor cells work together in cooperation to detect outage cells.

In [36], Zoha et al. proposes a solution for automatically detect sleeping cells. They proposed two different algorithms KNN and LOFA in order to detect sleeping cells.

3.2.2. Unsupervised Learning

Unsupervised learning algorithm specifies a set of inputs and then its goal is to correctly conclude the outputs. Most popular unsupervised learning algorithms are K-Means, anomaly detection and Self Organizing Maps (SOMs): Zoha et al., in [37, 38], present and evaluate an outage detection framework. The study analyzes two different anomaly detection algorithms (LOFAD and One Class Support Vector Machine based Detector (OCSVMD)) in order to detect the outage. In [34, 36, 37, and 38], LOFAD algorithm, which is one of the anomaly detectors under unsupervised machine learning algorithm, is applied to detect cell outage. In another work of Chernov et al., in [30], applied and compared different anomaly detection algorithms such as KNN, Self-Organizing Maps (SOM), Local Sensitive Hashing (LSH) and Probabilistic Anomaly Detection (PAD). Sleeping cell detection solution provided by Chernogorov et al. in [39] used a Cluster Based Local Outlier Factor (CBLOF) for the classification of sleeping cell. A most popular dimension reduction technique called Principal Component Analysis (PCA) is also used. In [40], Ma et al. propose an unsupervised clustering algorithm in order to tackle the problem of outage detection. Unsupervised data mining techniques called K-means is used in [41] for classification.

Another approach for outage management involves in the analysis of statistical data. Such approaches exhibited in some recent researches. For example, Bandera et al. in [42] presented method to detect outage through the analysis of HO statistics. Similarly, Munoz et al. in [43] proposed a solution that detects degraded cells through the analysis of time-series evolution metrics. The solution compares the measured metric with a generated hypothetical degraded pattern. If the two are adequately correlated, then outage is detected.

3.3. COD Based on Incoming Handover Statistics

One of the use cases in Self-Healing is cell outage detection. A cell is in outage when it cannot carry traffic due to a failure [42]. In this situation, it is very important to identify the cell in outage as soon as possible based on the analysis of the performance of the problematic cell or its neighboring cells to minimize the effects in the network. Generally cell outage detection is an automatic way of discovering cells with services in outage [40].

Several approaches have been used to implement COD by monitoring KPIs and alarms reports, collecting and analysis of reference signals or HO statistical data. Cell outage detection based on HO statistical data is studied by different researchers with different detection techniques. For example, Bandera et al. in [42] present a method to detect outage through the analysis of HO statistics. The method uses two cases: in the first condition the detection is performed when the cell in outage is able to report performance indicators for detection analysis; the second condition is when these indicators are not available from cell in outage base station which is affected. In this case HO statistical data from neighboring cells is used for the detection of outage cell. Similarly, Peng Yu et al., in [44], proposed self-organized COD architecture and approach for cell outage management (COM) of heterogeneous cloud radio access networks (H-CRAN), which becomes one of the important components of 5G networks, based on HO statistics.

COD algorithm, which is proposed in this study enables to detect a cell in outage when Node-B is affected and there are no available KPIs or alarms from the cell in outage, is based on neighbor measurement HO statistical data. The proposed algorithm analyzes the number of incoming HO on a per-cell basis. The algorithm can consider a predefined time intervals called Granularity Periods (GPs) that determines the time interval between two consecutive data collection instance and algorithm executions. The algorithm can be executed every time with a periodicity of 5, 10, 15 minutes or one hour that the HO statistics are updated in the

OSS depending on operator configuration. In practice, there is no comprehensive real-time access to performance data at the OSS due to management link capacity limitations and statistical relevance of performance measurements. So, operator rather assigns GP for KPIs collection depending on importance of a particular performance measurement. If the number of incoming HO of a certain cell becomes zero in specific time interval of statistical data collection, the algorithm analyzes and then selects the cell as a candidate cell in outage.

The number of incoming HO on a per-cell basis is calculated based on HO statistics collected from OSS on per- neighboring basis using pre-processing methods such as data filtering, summarization, normalization and profiling. If it is the first measurement period in which the algorithm is activated, the data is kept as a reference and wait for the next measurement period to collect a testing incoming HO data, preprocess it and perform cell outage detection from the comparison of consecutive periods. Finally, the detected cell outages list is obtained and the algorithm is stopped until the next measurement period.

Chapter IV

4. Data Mining

4.1. Overview of Data Mining

The enormous amount of digital data is collected from different sources every day in the world and the amount is ever-increasing and there's no end [8]. It has been estimated that the amount of data stored in the world's databases doubles every 20 months that we can only related to the pace of growth qualitatively but not easy to justify it in quantitative term. As Jiawei Han et al. stated in their book [7] "The world is data rich but information poor". So the analysis of big data in different innovation and productivity research brings development and business competition. Therefore, the extraction and interpretation of hidden patterns, in other term called knowledge extraction has great importance.

In the case of telecom operators, various measurements are collected from all network devices and stored in a database which may include an enormous amount of records. The management of this huge database is a very challenging task for network operators, especially processing it and extracting useful information. Knowledge Discovery in Databases (KDD) is known to be an efficient tool for the processing and extraction of information from large volumes of data [7]. KDD is an umbrella term for machine learning and data mining techniques, which address anomaly detection, clustering, classification and other types of information retrieval.

In that sense, data mining is a modern tool, also popularly referred to as KDD [7], that aims to discover meaningful knowledge (patterns) from large datasets that are stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams[7]. Although, exact meanings of KDD and data mining terms differ from each

other, in most literatures they are used interchangeably. But, additionally, KDD is concerned about the evaluation and interpretation of discovered patterns.

4.1.1. Knowledge Discovery Process

The most popular methodologies utilized for data mining projects are Cross Industry Standard Process for Data Mining (CRISP-DM); Sample, Explore, Modify, Model, and Assess (SEMMA); and KDD [45]. Among these, KDD is the usual and mostly used in academia. However, all of the methodologies essentially follow similar standard data analysis (sequential and iterative data analysis).

KDD process is concerned about manipulation with massive data, scaling algorithms for better performance, proper interpretation of retrieved information, and human interaction with the overall process. KDD process is a sequential analysis that includes the following steps: selection, preprocessing, transformation, data mining, and information interpretation steps. The knowledge discovery process shown in Figure 4.1 illustrates the relationship between data mining and KDD process.

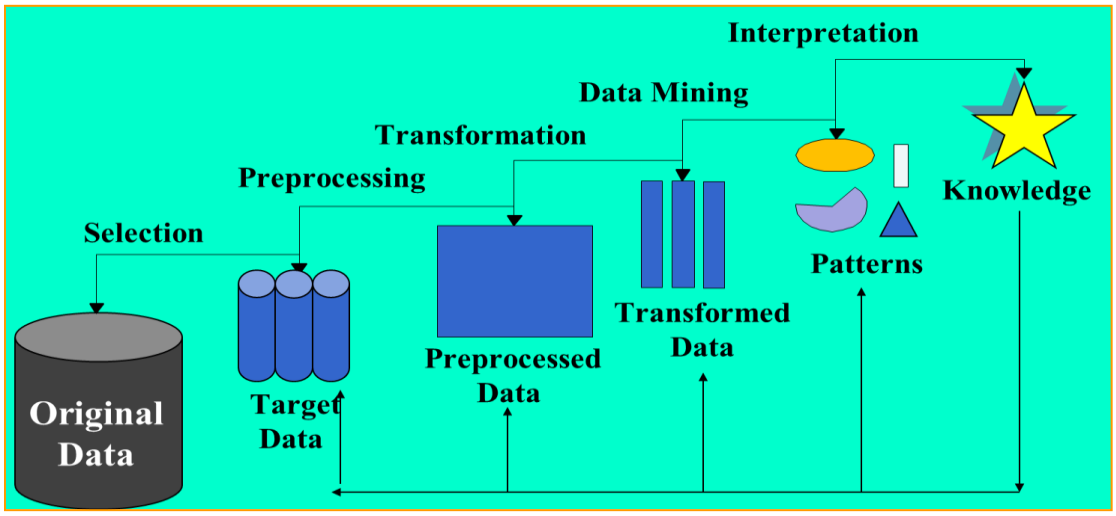

Figure 4. 1: Data mining step in the process of knowledge discovery [7].

- **Data selection**: at this stage data significant to the analysis purpose is retrieved from the database. This stage creates a target dataset, or data samples, on which discovery is to be performed by understanding the data.
- **Data Pre-processing:** it is concerned in removing noise or outliers if any, collecting the necessary information to model, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes. On top of these tasks, deciding on DBMS issues, such as data types, schema, and mapping of missing and unknown values are parts of data cleaning and pre-processing.
- **Data transformation:** data are transformed using dimension reduction and consolidated into forms suitable for mining by performing summary, normalization or aggregation operations. In some cases where there are large numbers of attributes in the database, reduction of dimension increases efficiency of the data-mining step with respect to the accuracy and time utilization.
- **Data mining**: in this stage which is an essential process step of KDD in which data mining techniques are applied to search for patterns of interest (knowledge). Data mining techniques include classification rules or trees, regression, clustering, sequence modeling, dependency, and line analysis. Therefore, selecting the right algorithm for the right area is very important.
- **Pattern evaluation:** the mined data is presented as interesting patterns representing knowledge to the end user in the form to viewable to human. This involves data visualization, where the user interprets and understands the discovered knowledge obtained by the algorithms.
- **Knowledge presentation:** knowledge representation and visualization techniques are used to present extracted knowledge to users.

The first three points above are different forms of data preprocessing, where data are prepared for mining.

4.1.2. Data Mining

To reveal valuable information from large amount of data which is stored in different sectors such as business, science and engineering, medicine, and almost every other aspect of daily life, it requires a tool that automatically uncovers the information [7]. So that, the development of information technology in data collection, database creation and data management leads to the birth of data mining, which is the process of analyzing large datasets and summarize them into useful information.

Database and information technology has grown from simple file processing systems to some sophisticated and powerful database systems since 1960s [7}. Later in 1970s, the development of database systems progressed to data modeling tools and relational database systems in which data are stored in relational table structures. In the late 80`s data mining term began to be known and used within the research community by statisticians, data analysts, and the management information systems (MIS) communities with the development of advanced database systems, data warehousing, and data mining for advanced data analysis.

Since data mining is interdisciplinary topic and can be defined in many different ways, many people treat data mining as a synonym for KDD, while others view data mining as merely an essential step in the process of knowledge discovery [7]. The later means data mining defined as the component of KDD process and deals with the exploration of internal patterns in databases.

It is believed that evolution of information technology is the basis of data mining. The data management approach evolved in the development of data collection and creation, data management (i.e. data storage, retrieval and database transaction processing), and advanced data analysis (data warehousing and data mining). In data mining, the data is stored electronically and the search is automated by computer [8]. In a simple term, data mining is

defined as a process of discovering patterns, which is meaningful, by analyzing data stored in databases. There are a number of data mining functionalities which can be categorized in two data mining methods called predictive and descriptive [46]. Some of the functionalities of the two methods are:

- characterization and discrimination (Descriptive)
- mining of frequent patterns, associations, and correlations (Descriptive)
- classification and regression (Predictive)
- clustering analysis (Descriptive)
- Outlier analysis (Predictive)

Details of the two data mining methods and associated functionalities are briefly described in the subsequent sub-section.

4.1.2.1. Data Mining Models

Data mining aim is to discover valid, potentially useful, and easily understandable correlations and patterns present in existing data and predict possible trends [47]. Its functionalities are also used to specify the kinds of patterns to be found in data mining tasks. Thus, such tasks can be classified into two major groups [7]. They are named predictive and descriptive mining groups. Predictive mining model perform induction on the current data in order to make predictions whereas descriptive mining model characterize properties of the data in a target dataset.

Predictive Model

The objective of predictive model is to predict the value of a particular attribute based on values of other attributes. It makes a prediction about values of data using known results found from different historical data [47]. Prediction methods use existing variables to

predict unknown or future values of other variables. Predictive model includes classification, prediction, regression and time series analysis tasks [7].

Classification aims to categorize unseen input data records into known classes. The assignment model or classifier learns from the training dataset, where the relationship between records and classes is provided. Classification is a method of plotting the target data to the predefined clusters or classes [47].

Prediction this task is the same as classification and estimation, except that the instances are classified according to some predicted future behavior or estimated future value.

Time series prediction it is also called forecasting, is a special case of prediction based on the assumption that values are dependent of previous observed values in the time series. Times series methods are either univariate or multivariate (when several variables are used to predict the dependent variable).

Regression aims to predict numerical values for input data records. The mapping function learns from the training dataset, where the relationship between records and their values is known.

Anomaly detection extracts points or outliers that are considerably different from the rest manifold of data points.

Descriptive Model

The objective of descriptive model is to identify patterns or relationships in datasets. It serves as an easy way to explore the properties of the data examined earlier and not to predict new properties [47]. Descriptive data mining tasks are often exploratory in nature and frequently require post processing techniques to validate and explain results. Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis.

According to Mortenson, Doherty, & Robinson (2014), descriptive analytics summaries and transforms data into expressive information for reporting and one-to-one care but also allows for thorough examination to answer questions such as "what has occurred?" and "what is presently bang up-to-date?".

Clustering: identifies many points called clusters with similar properties or behaviors from the given data.

Summarizations: Summarization is the process of giving summary information from the data and can be observed as squeezing a given set of relations into a smaller set of designs while recollecting the supreme likely information.

Association analysis: Association mining is a two-step process. One is finding all frequent item sets and the second is generating strong association rules from the frequent item sets. Generally, it discovers relationships between records within the same dataset.

Sequence analysis: The goal is to model the states of the process generating the sequence patterns or to extract and report deviation and trends over time.

4.2. Data processing in data mining

Both data mining in engineering and data mining in science often collect massive amounts of data, so that the practices require data preprocessing, data warehousing, and scalable mining of data [7]. Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. The purpose of data preprocessing is to clean selected data for better quality. The data selected may have different formats as the selection could be from different sources.

Major steps involved in data preprocessing are, namely, data cleaning, data integration, data reduction, and data transformation. Sometime data collection categorized as part of

preprocessing.

Data Collection, which includes selecting a dataset or focusing on a subset of variables or data samples from many available databases is the beginning of data mining process.

Data cleaning is the first step in preprocessing stage to clean the data by filtering, smoothing noisy data, aggregating, and fill in missing values [7]. By filtering data, the selected data are examined for outliers and redundancies. Outliers differ greatly from the majority of data, or data that are clearly out of range of the selected data groups.

Data integration is one of the required tasks in data mining, especially, in data preprocessing. Its task is to merge data from multiple data stores. Proper integration of data from different sources, which may include multiple databases, data cubes, or flat files, can help to reduce and avoid redundancies as well as inconsistencies in the prepressed dataset [7]. Generally, data integration play an important role in data mining process in regards to improve accuracy and speed.

Other task of data preprocessing is **data reduction**, which includes finding useful features to represent the whole data or reducing representation of the dataset that is much smaller in volume, but produces almost the same analytical results. Useful feature representation of data can be done using data reduction strategies such as *dimensionality reduction* and *numerosity reduction* [7]. In dimensionality reduction, representation of the original data can be done by applying data encoding schemes to obtain a reduced or "compressed" one. But in case of numerosity reduction the data are replaced by alternative, smaller representations using parametric models or nonparametric models (e.g. *Regression, linear model, clustering, sampling, or data aggregation*). Data compression techniques can be an example of data reduction.

The **transformation** of data using dimension reduction or transformation methods is done at this stage so that the resulting mining process may be more efficient, and the patterns

found may be easier to understand. Usually there are cases where there are large numbers of attributes in the database for a particular case. With the reduction of dimension there will be an increase in the efficiency of the data-mining step with respect to the accuracy and time utilization [7]. Strategies for data transformation include smoothing, attribute construction, aggregation, normalization and discretization.

Chapter V

5. Outliers detection Techniques

5.1. Introduction

A database may contain points that do not comply with the general behavior or model of the data. These data objects are outliers. Outliers are values that differ greatly from the normal distribution of an attributed dataset [7]. Outliers can be considered anomalous due to several causes. The analysis of outlier data is referred to as outlier detection or anomaly detection (mining). Many studies are using unsupervised approaches such as anomaly and outlier detection to detect suspicious activity [8]. An early and widely referenced definition of outlier that is defined by Grubber (1969) stated as

> "An outlying observation or "outlier" is one that appears to deviate markedly from other members of the sample in which it occurs". (*Grubbs, 1969*) [48]

Outlier detection is an important branch in data pre-processing and data mining. It is one of the learning algorithms, which is being used to differentiate between data that appears normal and abnormal with respect to the distribution of the training data. Outlier detection techniques are used in many applications, such as network intrusion, medical diagnosis, public safety and security, fraud detection and etc. [7].

The traditional outlier detection methods can be classified into four main approaches: statistical-based, distance-based, density-based and clustering-based [8, 9]. Statistical methods are model-based methods in which data that do not following the model are outliers. Distance-based and density-based outlier detections are the two major types of proximity-based outlier detection methods. A point is an outlier if the proximity of the point

to its neighbors significantly deviates from the proximity of most of the other points to their neighbors in the same dataset [7]. A distance-based outlier detection method checks the neighborhood of a point, which is defined by a given radius and then labeled as an outlier if its neighborhood does not have enough other points. Most of the distance-based methods are designed with the use of Euclidean distances. Moreover, distance-based outlier detection methods don't capture local outliers rather they are used for global outliers. A density-based outlier detection method is also proximity-based, which compares the density of a point and that of its neighbors. If its density is relatively much lower than that of its neighbors, then a point can be considered as an outlier [7]. The last category of outlier detection is clustering-based methods which consider that outliers belong to small or sparse clusters [7]. In a typical density-based clustering method, there are two parameters that define the notion of density: *(i)* a parameter MinPts specifying a minimum number of data points; and *(ii)* a parameter specifying a volume [49]. These two parameters determine a density threshold and used to detect density-based outliers by comparing the densities of different sets of data points.

Outlier detection algorithms could either be global or local. Global approaches refer to the techniques in which the anomaly score is assigned to each instance with respect to the entire dataset. On the other hand, the anomaly score of local approaches represent the outlier-ness of the data point with respect to its direct neighborhood. The local approaches can detect outliers that are ignored using global approaches, especially in case of a varying density within a dataset [49].

The outputs produced by outlier detection techniques are in the form of scores or labels [50]. Scoring techniques assign an outlier score to each instance in the test data based on measuring "outlier-ness" of the instance [7]. So, the output will be a ranked list of outliers that enable to select the top few anomalies using specific threshold value. Labeling techniques, on the other hand, assign a binary label (normal or abnormal) to each test instance.

The outlier detection extension contains two categories of approaches as stated above: nearest-neighbor based and clustering based algorithms [50]. Algorithms in the nearest-neighbor category assume that outliers lie in sparse neighborhoods and that they are distant from their nearest neighbors. The second category algorithm assumes that normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster. In this paper only nearest neighbor based outlier detection is briefly explained.

5.2. Nearest neighbor based outlier detection

The concept of nearest neighbor analysis has been used in several outlier detection techniques. Nearest neighbor based outlier detection techniques require a distance (or similarity measure) defined between two data instances. Distance (or similarity) between two data instances can be computed in different ways. Such techniques are based on the key assumption that Instances of outliers are located at large distances from their closest neighbors, while normal data occur in dense neighborhoods [51]. Nearest-neighbor based algorithms assign the anomaly score of data instance relative to their neighborhood.

The outlier-ness of the data points or instances is mostly dependent on the assumptions, good understanding of the data itself and the circumstances of the analysis. Aggarwal in [51] clearly defined the situation as *"The core principle of discovering outliers is based on assumptions about the structure of the normal patterns in a given dataset. Clearly, the choice of the "normal" model depends highly on the analyst's understanding of the natural data patterns in that particular domain*". This implies that analyst to have a good understanding of the data representation.

Nearest neighbor based approaches of outlier detection has extension such as LOF, connectivity based outlier factor (COF), local outlier probability (LoOP), influenced outlier-ness (INFLO) and Cluster-Based Local Outlier Factor (CBLOF). LOF is used to find

anomalous data points by measuring the local deviation of a given data point with respect to its neighbors [49]. Based on the issue with LOF effectiveness, various techniques have been proposed as a modification of LOF. INFLO was introduced using a symmetric nearest-neighbor relationship based on the nearest-neighbor distance as well as the reverse nearest-neighbor distance in order to define the local outliers [51]. Similarly, COF was proposed to be able to identify outliers in low density regions effectively. The major difference between LOF and COF is that COF defined the neighborhood of a data point incrementally by adding the closest point to the current neighborhood set. Another LOF approach is that it can be combined with other clustering techniques in which anomalies are defined based on the combination of both local distances to nearby clusters and the size of the clusters to which the data point belongs. Such technique is called Cluster-Based Local Outlier Factor. Data points in small clusters and that are at a large distance to nearby clusters are identified as outliers.

The distance-based and density-based are the ones applicable to the problem of COD. For the accuracy in most COD studies, the density-based approach usually results in a higher accuracy than the distance-based approach. The density-based approach is capable of solving the multi-cluster outage detection problem effectively [49]. The LOF is an effective method to find an outlier and it is actually based on the concept from the distance-based approach. However, the LOF algorithm generates a relative density value instead of a distance value.

This thesis focuses on density-based original LOF outlier detection algorithm and some LOF based fast anomaly detection algorithm called FADD. Fast Anomaly Detection given Duplications (FADD) has been introduced to overcome LOF duplicates data point analysis problem.

5.3. Local Outlier Factor

The nearest-neighbor based methods in outlier detection attract much attention and enjoy much popularity due to the natural connection between neighborhoods and outlier-ness. Breunig et al. (2000) presented one outlier detection method called the LOF that is widely used [49]. This method calculates the degree of being an outlier for each instance based on the local density around it. Generally, LOF compares the density of data instances around a given instance X with the density around X's neighbors by using the k-nearest neighbor technique [52]. If the density of data instances X is low compared to X's neighbours, then it means that X is relatively isolated or it is called an outlier. Such outliers are considered anomalous. The LOF algorithm computes the so-called LOF score value for each instance, which is a measure of how anomalous the instance is. The LOF outlier factor value is set to the ratio of the local density of the data instance to the average local density of its neighbors. Outliers generally are data points, whose densities differ much from their neighbors' densities.

Most of the studies that are considered in [49] show that dataset has got a binary property, which is either an outlier or not being an outlier. This situation is not suitable for many applications, which is more complex in nature. So, it becomes more meaningful to assign a degree of being an outlier for each instance in dataset.

In LOF, the outlier factor is local in which only a limited neighborhood number is considered for each point for outlier analysis [49]. The advantages of LOF include the ease of the interpretation of its outlier factor and its capability to identify outliers which were unable to be detected or hidden by the global approaches outlier detection.

5.2.1. Mathematical description of LOF

The method used here is the relative density of a data point against its neighbors as the

indicator of the degree of the point being outliers. Hence, Figure 5.1 illustrates how the k-distance, the key proximity in LOF, is defined as the max distance of k nearest neighbor points.

A. Definition of k-distance of a point p

For any positive integer *k*, the *k*-distance to a data point *p*, denoted as *k-dist*(*p*), is defined as the K^{th} smallest distance (nearest neighbor) to a data point p (as shown in Figure 5.1). *K* is a user-specified parameter.

$$\boldsymbol{k - dist(p) = dist(p,q)} \quad , \quad \boldsymbol{where\ \ k = 3}$$

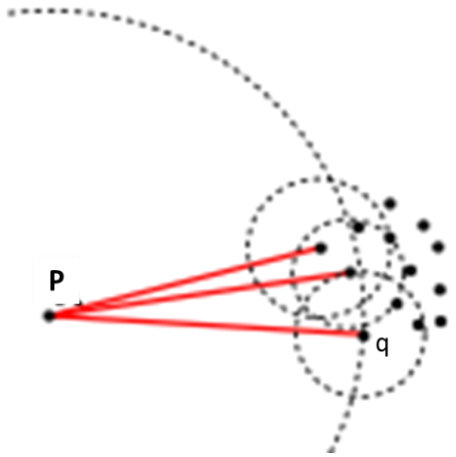

Figure 5. 1: *k*-distance of a point p, where *k*=3[49].

B. Definition of k-distance neighborhood of p

Given the *k*-distance of p, the *k*-distance neighborhood of p contains every point whose distance from p is not greater than the *k*-distance, i.e. $N_{k\text{-dist(p)}}(p) = \{ q \in D \setminus \{p\} \mid d(p, q) \leq k\text{-dist}(p) \}$. These points q are called the *k*-nearest neighbors of p. Just to simplify our notation to use $N_k(p)$ as a shorthand for $N_{k\text{-dist(p)}}(p)$. Note that the *k*-distance (p) is well defined for any positive integer *k*, although the point q may not be unique. $N_k(p)$ could be bigger than *k* since multiple points may have identical distance to p (i.e. The *k*-distance neighborhood $N_k(p)$, contains all the instances that are closer to p than its *k*-dist(p) value).

C. Definition of reachability distance of a point p

A reachability distance is considered as the intermediate parameter for LOF outlier factor calculation and it is expressed as equation 5.1 based on Figure 5.2, which illustrates the idea of reachability distance with *k*=3.

$$\boldsymbol{rdis_k(q,p) = Max\{dist(p,q), k-dist(q)\}} \qquad (5.1)$$

Where *q* is a target point and *p* is the current data point.

As Figure 5.2 shows if point q_2 is far away from *p*, then the reachability distance between the two is simply their actual distance *[dist(p, q_2)]*. But, if the two point's *p* and *q* are close as that of *p* and q_1, the reachability distance will be the *k*-distance of *p*.

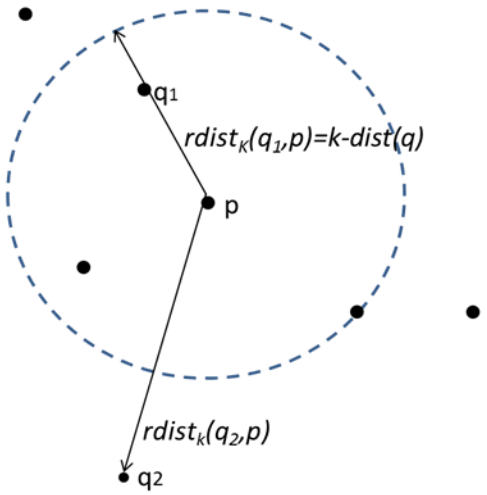

Figure 5. 2: rdist(q_1,p) and rdist(q_2,p), for K=3 [49].

The LOF algorithm uses reachability distance instead of Euclidean distance *dist(q, p)* to reduces the statistical fluctuations of *dist(q, p)* considerably for all the *q's* close to *p.* The strength of this smoothing effect can be controlled by the parameter *k* [49]. If the value of *k* is higher, then the reachability distances for points within the same neighborhood becomes similar.

For outlier's detection purpose, we consider a specific value of *k* and MinPts that is the minimum number of points in a typical density-based clustering algorithm and compare the densities of different sets of points dynamically. Therefore, we keep MinPts as the only parameter and use the values $rdist_{MinPts}(q, p)$, for $q \varepsilon N_{MinPts}(p)$, as a measure of the volume to determine the density in the neighborhood of a point p.

D. Definition of local reachability density of a point p

Local reachability density of a point p is the inverse of the average reachability distance based on the MinPts-nearest neighbor of p. A local reachability density is defined as

$$\boldsymbol{lrd(p)} = \frac{|\boldsymbol{N_k(p)}|}{\sum_{\boldsymbol{q \in Nk(p)}} \boldsymbol{rdist(q,p)}} \qquad (5.2)$$

Where $N_k(p) = \{q \mid dist(q, p) < dist(q)\}$. It calculates the average reachability distance of *k* neighbors.

E. Definition of LOF

LOF of a point p is the average of the ratio of local reachability of p and those of *q*'s *k*-nearest neighbors. LOF is defined as:

$$\boldsymbol{LOF_k(p)} = \frac{\sum_{\boldsymbol{q \in Nk(p)}} \frac{\boldsymbol{lrd_{k(q)}}}{\boldsymbol{lrd_k(p)}}}{\boldsymbol{|N_k(p)|}} \qquad (\mathbf{5.3})$$

Based on Equation (5.3), if its *lrd(p)* is relatively small compared to its neighborhoods *lrd's*, then the value of LOF is higher and point p is nominated as an outlier. That is, a point whose density is different from the densities of its neighbors is likely to be an outlier [52].

5.4. Anomaly Detection given Duplications

Data size that is stored and used for analysis purpose, now a day, is dramatically increasing throughout the world in every domain such as medical, telecommunication, agriculture and

space science etc. [52]. Such condition builds big data that may contain duplicate data points. Duplicate data points in big data with the same value or coordinate creates new data analysis problem since traditional outlier detection algorithms do not seriously consider duplicate data point problems because they only deal with smaller or medium dataset with some duplicating data points. To solve this problem Jay-Yoon Lee et al. in [52] proposed a modified version of LOF algorithm called Fast Anomaly Detection given Duplications (FADD).

In FADD identical data points in n-dimensional space are considered as a super node with their duplicate count information (C_i). FADD deals with M unique super nodes $SN(u_i)$ rather than visiting all N of the data points separately [52]. $SN(u_i)$ is the set of x_j that have the same coordinates as u_i ε U (where U is the set of all unique points in n-dimensional space). With this super node scheme, for the *k*-nearest neighbors' method, local reachability density of a duplicate point $x_j \varepsilon SN(u_i)$ with more than *k* duplicates is defined as [52].

$$\boldsymbol{lrd_k(x_j) = |SN(u_i)|/\beta} \qquad (5.4)$$

Where β is a constant designate infinitesimal artificial distance, which is the sum of the distances in the super node. The average distance is also calculated as the sum β divided by the duplicate count C_i.

This new definition of density follows naturally from equation (5.2) which defines density approximately as average distance to neighbors. In the case of points with many duplicates, the sum of the distances to other points in the neighborhood is 0. In FADD, instead of 0, β is used to avoid 0 values in the denominator of local reachability density.

The super node $SN(u_i)$ to have higher local reachability density, the β value should be smaller than the sum of distances of any other neighborhood set. To achieve this the β should satisfy the following condition: $\beta \leq k \times \min(d(x_i; x_j))\ \forall x_i; x_j \varepsilon X; i \neq j$ [52]. The C_i information differentiates the density of one duplicate point to another.

By re-defined the local reachability density as equation 5.4, it is possible to compute the $LOF_k(x_j)$ for the duplicate point x_j. $LOF_k(u_i)$ for a point with more than *k* duplicates will have higher value since the neighborhoods are just identical points with identical local densities.

Jay-Yoon Lee et al. in [52] remarked two benefits of super node scheme. First it enables to define the duplicate point's local density and outlier score. Second, it reduces the number of points from whole N data points to M unique super nodes, and avoids the computations for duplicate points in $SN(u_i)$ that have C_i larger than *k* (i.e. it reduces runtime complexity significantly).

It can be noted that the LOF score of a point with more than *k* duplicates should be a threshold value, because all *k* neighbours and the point itself are at the same point [52].

5.5. Performance evaluation of LOF

To evaluate the performance of the LOF algorithm, both normal and anomalous instances are needed. In this thesis the concept of the receiver operating characteristic (ROC) curve has been used to evaluate the performance of LOF data mining detection algorithm. The ROC curve plots the true positive rate (TPR or sensitivity) vs. the false positive rate (FPR or 1 – specificity) at all possible thresholds [8]. The TPR is the fraction of instances correctly classified as normal among all the truly normal ones. The FPR is the fraction of instances incorrectly classified as normal among all the truly anomalous ones. TPR takes the vertical axis against the FPR, which is on the horizontal axis of the plot. Equation 5.5 and 5.6 mathematically define the TPR and FPR of ROC curve [8]. TP, FP, TN, and FN are the number of true positives, false positives, true negative, and false negatives respectively.

$$\boldsymbol{Sensitivity = TPR = \frac{TP}{(TP + FN)} 100\%} \qquad (5.5)$$

$$\boldsymbol{Specificity = TNR = 1 - FPR = 1 - \frac{FP}{(FP + TN)} 100\%} \qquad (5.6)$$

Figure 5.4 illustrates the sample ROC curve plot. The diagonal line of the ROC curve divides the ROC space. Curves above the diagonal indicate a model is useful classifier; and curve below the diagonal is a misleading one.

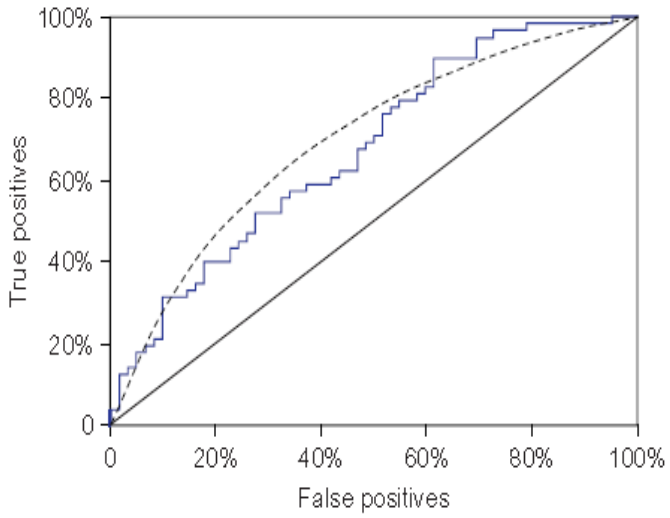

Figure 5. 3: Sample ROC curve [8].

An Area under ROC curve (AUC) metric is a threshold-independent measure of the performance and used for model comparison. AUC value of 1 or close to it indicates that the target algorithm under investigation has a higher classification power [8].

Chapter VI

6.Experimentation and Result Analysis

6.1. System Model

As it has been indicated in this thesis objective, data mining technique is used to adapt a model that proactively detects cell outage in UMTS mobile network using incoming HO statistical data as an input to the model. So, cell outage detection model has been used by implementing the selected LOF detection algorithm.

The thesis considered a UMTS network scenario in Figure 6.1 that illustrates the conceptual network scenario of 3G UMTS network with tri-sectored cells. The network consists of mobile user's equipment, Node-Bs, RNCs, and CN elements. The network also illustrates the mobile HO scenario. The HO process in this scenario is considered as a repeating procedure in which the serving, receiving cells and management network elements follow a certain organized manner to provide services to the mobile users during their HO. The operational units of the HO process are UEs, serving and receiving cells of Node-B and mobility management unit. UEs move through the cells of Node-Bs; serving cell of Node-B senses and decides based on the received information from the other network units; receiving cell receives the HO request and issue the acknowledgement signal, and the mobility management unit that is the control node responsible for idle UE and retransmission in HOs.

Considering Figure 6.1, we can deduce that the network explain the HO scenario. When there is cell outage in cell ① coverage, all UEs which were getting service from cell ① are forced to get connected to other neighboring cells that have strongest RSRP and HO margin which is set by operator. This HO scenario simply exhibits:

- Cells ②, ③, ④, ⑤, ⑥ and ⑦ incoming HO dramatically increased if the number of UEs is large.
- Cell ① stops giving service to new connecting users as well as UEs in transition to HO from neighboring cells to cell ①.
- Based on the operator HO data collection interval time, HO statistical data which consists of granularity period, number of incoming HO statistical data which includes HO attempts, failure and success, and location information is sent to OSS of the network management center as a statistical data.

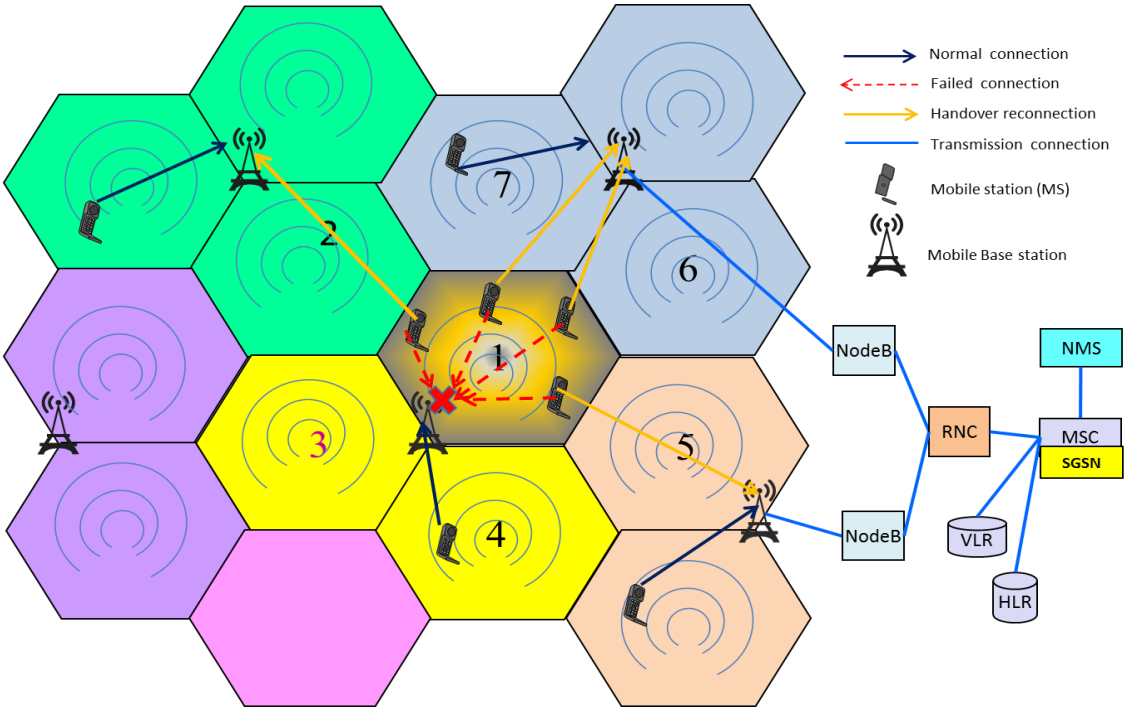

Figure 6. 1: Cellular principles and handover scenario.

HO data is normally used as an input for network design, performance analysis and optimization purposes. In this thesis, HO statistical data primary preprocessed and then the output is used as input data for data mining detection algorithms in order to identify cell outage and localize the position of the cell in outage. The performance of the detection algorithms is evaluated to get a reasonable accuracy and comparison of them is made.

To conduct this thesis, four-step process model illustrated in Figure 6.2 is implemented. The four steps LOF based COD model include data collection, preprocessing, data mining and pattern evaluation & localization. In this model, HO statistical data is collected from the OSS system per cell base. Then the HO data preprocessed using different procedures which are explained later in this chapter. Preprocessed output will be fed into a data mining detection algorithm, in this case LOF, which analyze data and perform the detection automatically. In this way, the anomalous cells can be detected timely so that appropriate correction measure can be taken by operator to minimize down time and service degradation of the network. The brief description of each process steps are explained below.

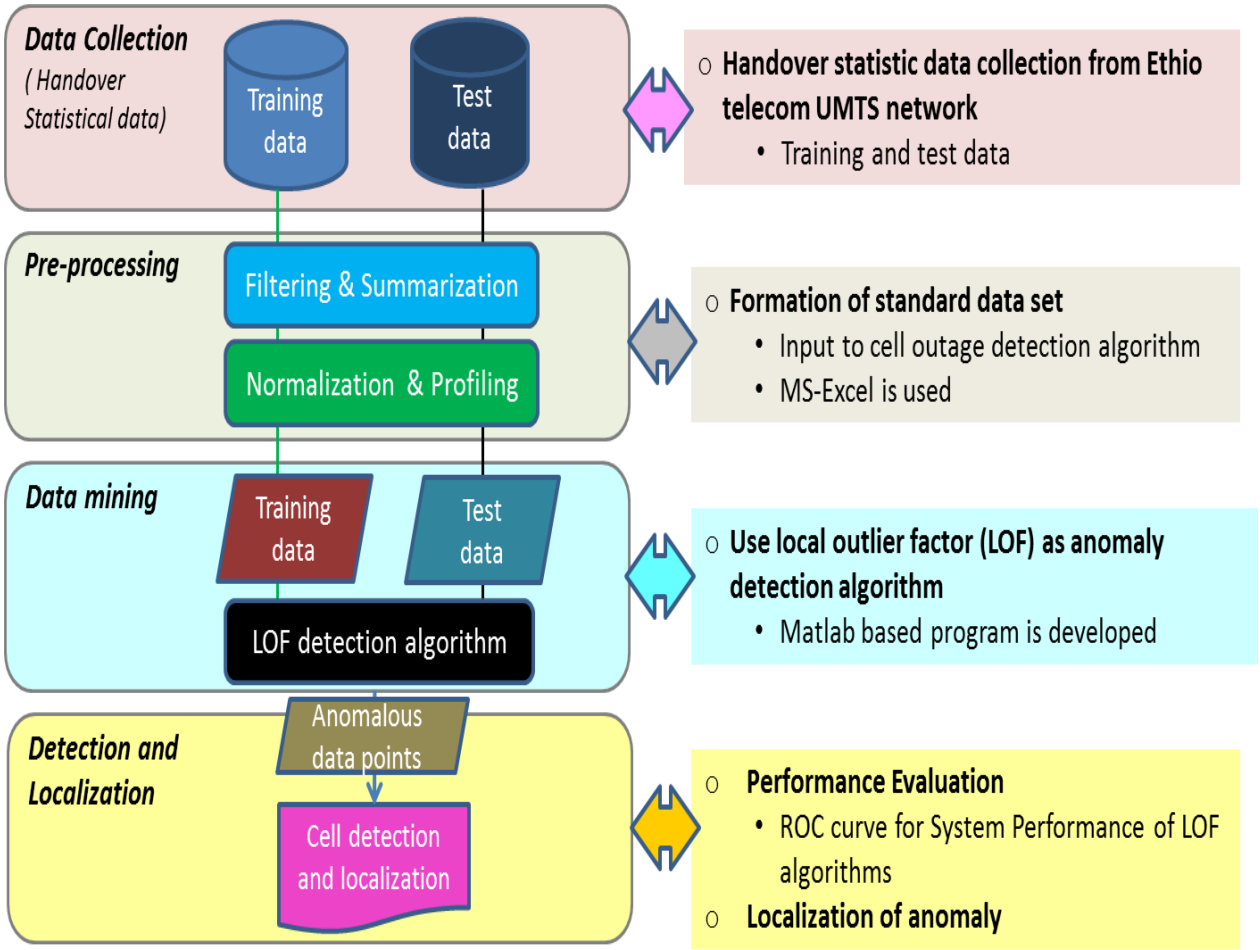

Figure 6. 2: Four steps process model.

Data Collection

The primary source of data to conduct this thesis is Ethio telecom UMTS mobile network incoming HO statistical data that is collected and reported by a base station (Node-B) periodically to OSS system. Since the data is very huge and contains information that is not significant for this thesis purpose, it needs to be preprocessed before using it as an input to LOF detection algorithm. The data extracted from the OSS system is in CSV format which is not suitable for window base manipulation therefor it is needed to be converted to MS-Excel format.

The incoming HO data stream from the system is normally cell-level data (data per cell) which contains information tabulated in Table 6.1. The thesis considers the cell outage as site outage if all cells in the Node-B are out of service.

CELL-LEVEL DATA	
Measurements	***Description***
Time	Granularity Period
Cell Localization	Longitude and Latitude Information
Cell inHO	Cell Incoming Handover

Table 6. 1: Structure of the cell level HO information.

Cell level incoming HO is a general term that represents different HO types such as hard and soft HOs. The total number of incoming HO for a given cell consists of both hard and soft HO statistics as well as circuit switching and packet switching traffic HO types.

Preprocessing

The first step in the cell outage detection process is to make ready the data in the reference (training) and the testing set. Depending on the problem and the applied data mining algorithms, preprocessing phase may contain several kinds of tasks. The major data preprocessing tasks considered in this thesis are data cleaning, filtering, aggregation and finally normalization. Profiling of data and then arranging in one MS-Excel format file that is suitable for the experiment is the prerequisite for data mining. Data is cleaned by

removing the records that have incomplete (invalid) data or missing values under each column. It is assumed that records with this nature are few and their removal does not affect the entire dataset. Filtering is one of the preprocessing tasks, which removes the unnecessary contents or errors from processing data and enables the data to be significantly reduced without losing the important features of the data [7]. Data aggregation is a process during which data is gathered and presented in a summarized format without loss of information necessary for the analysis task [7]. The other preprocessing task, which grouped in data transformation, is normalization. Normalization involves transforming the data to fall within a smaller or common range, such as [-1, 1] or [0.0, 1.0]. The final task in preprocessing is profiling. Data from incoming HO statistical data and geo-location database information are correlated to create as a final database for testing data.

MS-Excel is used for data preparation, pre-processing and analysis task because it has the capability of filtering attribute with different values and summarizing different values of the same attribute. Besides, it is also very important to make ready the data and easily convert into of the file that accepted by detection algorithm.

Data mining

The data mining technique in the third major step of process model in Figure 6.2 performs the analysis of the incoming HO statistical data to detect cells in outage. For this purpose LOF detection data mining algorithm is considered. The algorithm compares the density of data instances around a given HO instance X with the density around X's neighbors. If HO instance X is low density compared to X's neighbors, the so-called LOF score that is a measure of how anomalous the instance is, will have a higher value. It means that X is relatively isolated (an outlier). Such outliers are considered cells in outage. The mathematical descriptions of reachability distance, local reachability density and LOF factor have been described in section 5.3 of equations 5.1, 5.2 and 5.3 respectively.

The algorithm for LOF detection method is listed below. By iterating over the instances, the algorithm calculates the LOF score of each instance for all K values.

Algorithm: Local Outlier Factor Detection Model

LOF (Input data set $X = \{x(t)\}_{t=1}^{N}$ *, Minimum neighbors* k_{min}*, Maximum neighbors* k_{max}*)*
for t = 1 to N do
for k = k_{min} *to* k_{max} *do*
Find k-dist (x(t))
Find k-distance neighborhood, N_k *of x(t)*
Calculate $rdist_k(x(t); x(i))$ *for instance pairs x(t) and* $x(i) \in N_k(x(t))$ *.. from Eqn. 5.1*
Calculate $lrd_k(x(i))$ *for* $x(i) \in \{x(t)\} \cup N_k(x(t))$ *……..from Eqn. 5.2*
Calculate $LOF_k(x(t))$*……..from Eqn. 5.3*
end for
$LOF(x(t)) = max(LOF_k min; .... ; LOF_k max)$
end for
return LOF

Localization and Performance evaluation

During the profiling phase, geo-location information is used to locate each and every cell or site in the dataset. The location of cell or site can be established by correlating the geo-location information and HO statistical data based on cell ID which is a common parameter for both. When the COD algorithm detects the cells in outage, the malfunctioning cells in a Node-B are easily identified and located. Consequently, the cells in outage change the original normal color to different color in order to distinguish them from operating cells. The detection algorithm program utilizes different colors to differentiate a complete site outage and partial site outage (a number of cells in a site are out of service).

In this thesis, the quality of the COD model (LOF) is evaluated using Receiver Operating Characteristic curve analysis.

6.2. Discussion and Experimental Results

The thesis considered a real UMTS mobile network scenario and statistical HO data of the system for the analysis of COD with LOF data mining algorithm. For simplicity, part of Addis Ababa UMTS Node-Bs, which are controlled almost by a single RNC, were selected. We apply the proposed LOF and modified LOF (FADD) algorithms to the preprocessed incoming HO statistical data which was collected from Ethio telecom OSS system.

6.2.1. Experimental Real Network Scenario

Ethio telecom UMTS network comprises of 7500 number of base stations (Node-Bs) throughout the country. Among them, 744 Node-Bs that incorporate 7159 cells have been deployed in Addis Ababa. Even if considering all UMTS sites for the analysis of cell outage detection is possible, it is better to select a given UMTS network deployment sites and perform the cell outage detection analysis in order to display the results in a graphical representation. So that, for this thesis analysis purpose, a central part of Addis Ababa Node-B sites was nominated to create a real network scenario for the experiment. The nominated area has been selected with the assumption that the number and mobility of the mobile users are adequate to acquire HO statistical data for the analysis. To do that, four geographical coordinate points, which are depicted in Table 6.2, were designated to mark the borders of the selected region. Based on the selected Longitude and Latitude, a total of 35 Node-B sites that contain 401 sectoral cells have been identified. 34 of the Node-Bs are controlled by RNC-101 and the remaining one Node-B is controlled by RNC-103.

	UMTS Node-B sites in Addis Ababa	Selected Node-Bs
Longitude	*38.645469 – 38.940409*	*38.740000 – 38.760000*
Latitude	*8.818630 – 9.0943901*	*9.010000 – 9.040000*
Total No. of RNCs	*5 (RNC101- RNC105)*	*2 (RNC 101+ RNC 103)*
Total No. of sites	*744*	*35 (34+1)*
Total No. of cells	*7159*	*401 (389+12)*

Table 6. 2: Geographical coordinates to selected UMTS Node-B sites in Addis Ababa.

Addis Ababa UMTS network base stations selection process outcome is illustrated in Figure 6.3. 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.

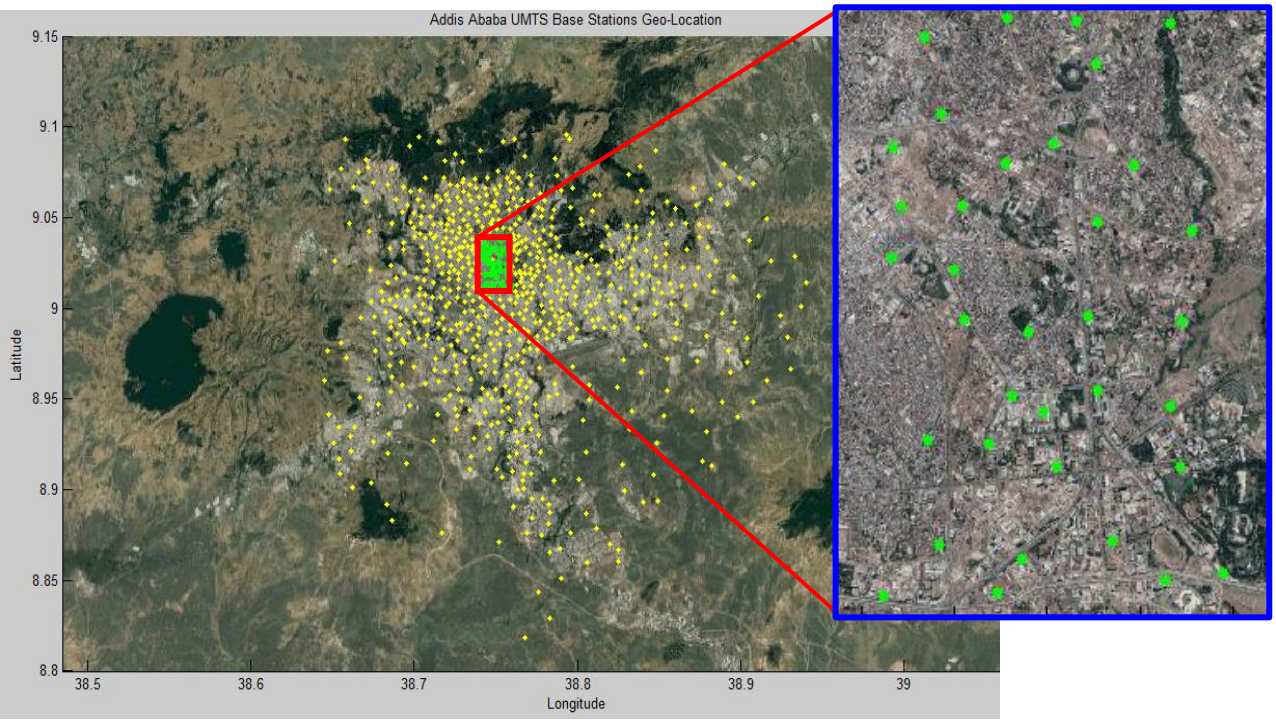

Figure 6. 3: Geographical area in Addis Ababa and UMTS Node-Bs distribution.

Based on the selected region, geo-location information, site ID, cell ID and controller RNC's attribute values of each cell were filtered and organized in Table 6.3 fashion from the

general Node-B site location information database. It also used as reference information to filter out the number of HO per cell data from huge incoming HO statistical data collected from OSS.

Index	Cell ID	Site ID	Location Name	UMTS Cell	Longitude	Latitude	RNC ID
1	30191	111019	CAAZ_Rail Way Station	111019_1	38.75633521	9.01169456	101
2	30192	111019	CAAZ_Rail Way Station	111019_2	38.75633521	9.01169456	101
3	30193	111019	CAAZ_Rail Way Station	111019_3	38.75633521	9.01169456	101
4	30194	111019	CAAZ_Rail Way Station	111019_4	38.75633521	9.01169456	101
5	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
400	55792	112119	WAAZ_AA,W01, Merkato,Sobla Construction & Real Estate BLDG	112119_12	38.74214183	9.03022007	101
401	55793	112119	WAAZ_AA,W01, Merkato,Sobla Construction & Real Estate BLDG	112119_11	38.74214183	9.03022007	101

Table 6. 3: Sample selected Node-B sites with their cell IDs and location information.

Usually, real world data are typically contain noisy and inconsistent data as well as it could be originated from heterogeneous sources [7]. Such unclean data may cause difficulty for the data mining process. So that data preprocessing, as a primary task, should be considered prior to data mining implementation. Hence, in this thesis work such data preprocessing step has been taken as a major task for the preparation of suitable incoming HO dataset that has been used as an input for LOF detection algorithm.

6.2.2. Data Preparation and Parameter Analysis

As primary procedure, incoming HO statistical data was collected from the Ethio telecom OSS system at different time instances to prepare reference and testing datasets. Originally, one time collected HO statistical data has about 60,000 records with 52 attribute values each. The attributes of the statistical data comprises of number of HO attempt, failure and success both in hard HO and soft HO considering packet and circuit switching traffic separately. The next step after collecting the HO data was to perform data transformation which includes filtering, aggregation, profiling and normalization the dataset. So that incoming

HO data of all cells that controlled by RNC-101 and a single cell site which also controlled by RNC-103 were filtered out from originally collected incoming HO statistical data and then aggregated the number of HOs per cell. After filtering and aggregation preprocess have been done, a total of 401 datasets have been found for the purpose of conducting this study. Moreover, the number of initial attributes reduced to 7 most suitable attributes.

To get a normalized final input dataset, the Node-B site location information and the incoming HO statistical data need to be correlated, profiled and normalized. So, normalization step is used to standardize the range of values to a certain range, often 0 to 1. Min-max normalization is used for LOF detection scenario to scale the given **X** which is the set of x's. In equation 5.7 describes min-max normalization.

$$x_i' = \frac{x_i - \min(X)}{\max(X) - \min(X)} \qquad (5.7)$$

Where **X**= $\{x_1,\ldots, x_n\}$ and x_i' is new ith normalized data.

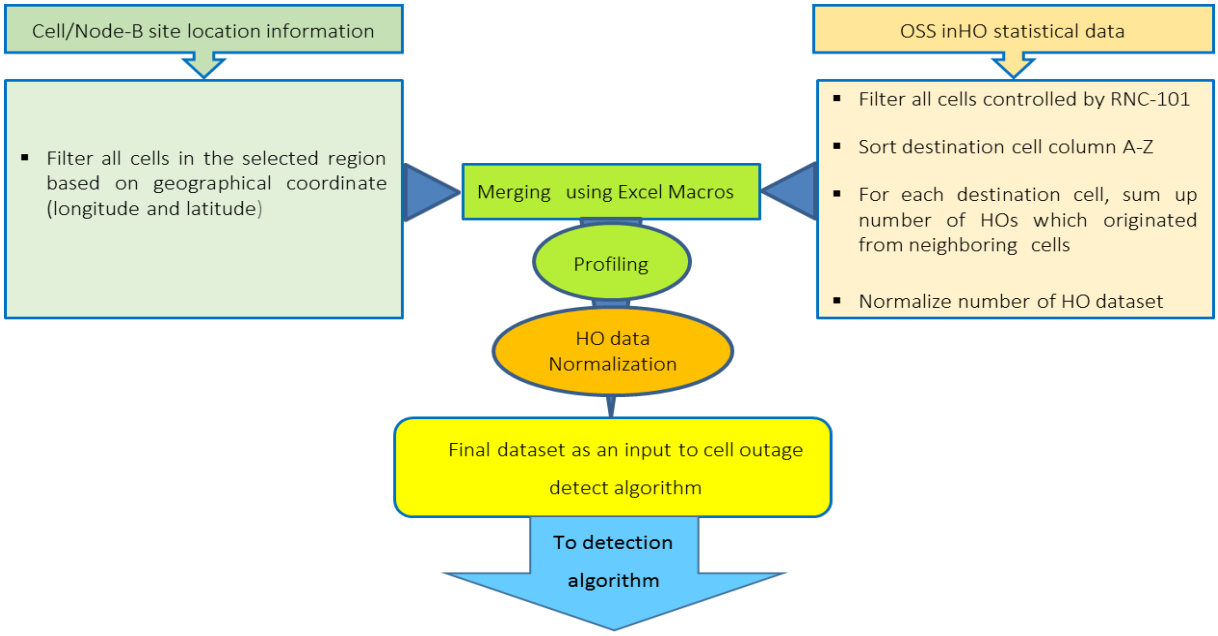

Figure 6. 4: Handover and Geo-location data correlation procedure for normalization.

The procedural steps of Figure 6.4 were generally applied to get the normalized reference and test dataset that was used as a final input datasets of the detection algorithm.

For our real network experiment of this thesis, LOF and FADD programing in Matlab have been adapted and they have also been applied to real Ethio telecom UMTS network scenario based on the description in sections 5.3 and 5.4 respectively. The cell statistical incoming HO data which was pre-processed in Section 6.2.1 is fed into the detection algorithms which are designed to detect cell in outage. The two LOF version detection algorithms compared testing HO dataset against the reference normal pattern. In this case, the deviation of the test data from the normal pattern is expressed in terms of the LOF score that discriminate outliers from normal.

The accuracy of LOF data mining algorithm is dependent on the input parameter *K* which is the number of nearest neighbors and the selection of the threshold value of LOF score. A bad selection of *k* value can easily mask the structure in the dataset so that the algorithm may not detect potential outliers. Therefore, different *k* values were used to select the optimal values of *k* with which the LOF algorithm can detect the most accurate estimation of cell outages. Due to the high computational requirements of the LOF algorithm, step size increments of *k* from arbitrary k_{min} to k_{max} has been used to get the optimum value of *K*. Based on the repeated trials and detection result's evaluations, the optimum value of *k* is nominated as 16.

Another consideration in LOF detection algorithm is setting the threshold score value of LOF that determines the outlier-ness of the dataset. Normally the algorithm calculates the LOF score of each dataset for each value of *k*. After finding the score of each dataset, the next step should be grouping of LOF scores to binary labels (normal and abnormal dataset). To do that the best way is to set a threshold depending on the specific application that the LOF is applied to. Datasets with LOF scores larger than the threshold are considered as

anomalies and the rest as normal. With this concept, the threshold LOF score has been considered in this thesis work to detect normal cell and cell in outage.

In the case of FADD algorithm, one additional parameter β has been used as constant designate infinitesimal artificial distance, which is the sum of the distances in the super node, to avoid zero value of the denominator or undefined result of local reachability density in equation (5.2) as a result of duplication data points (i.e. where $C_i > K$). The average distance of the super node is also calculated as the sum β divided by the duplicate count C_i.

In this thesis, the performance of the targeted detection algorithms was evaluated using ROC curve analysis. An Area under ROC curve (AUC) metric is used for comparison purposes of the two versions of LOF detection algorithms. In this case true positive rate (sensitivity) and false positive rate (1-specificty) have been considered as performance comparison measurement for the two detection algorithms.

6.2.3. Results and Discussion

This section of the thesis presents the results obtained in the LOF and FADD COD algorithms analysis. A number of experiments were carried out to evaluate the attainments of the proposed algorithms and compared it with each other. As pointed out in section 6.2.1, the thesis work considered real network scenario of Ethio telecom Addis Ababa UMTS network and the cell level periodic incoming HO statistical data obtained from OSS for cell outage identification purpose.

Both normalized reference and test cell incoming HO statistical data which obtained in different granularity (data collection frequency) periods have been fed into the LOF data mining procedure. Normally, GP of HO statistic data collection in Ethio telecom UMTS network is 30 minutes. This is due to the huge HO data generation from cell sites at a time and unavailability of adequate data storage capacity, which directly associated with cost.

The automatic COD is strongly influenced by the GP as it determines the availability of HO data that is required to analyze the cell in outage. Hence, the successful and automatic COD analysis at OSS level is only possible when all required HO statistical data is available with shorter granularity periods (i.e. 5 or 10 minutes). So, to implement the adopted COD in operators' network, like Ethio telecom, the GP of HO statistical data collection should be shorter than the actual 30 minutes.

The LOF normalized dataset distribution is illustrated in figure 6.5. This normalized date sets are used for both LOF and FADD algorithms to analysis and subsequently detect anomalous behaviors of the HO dataset.

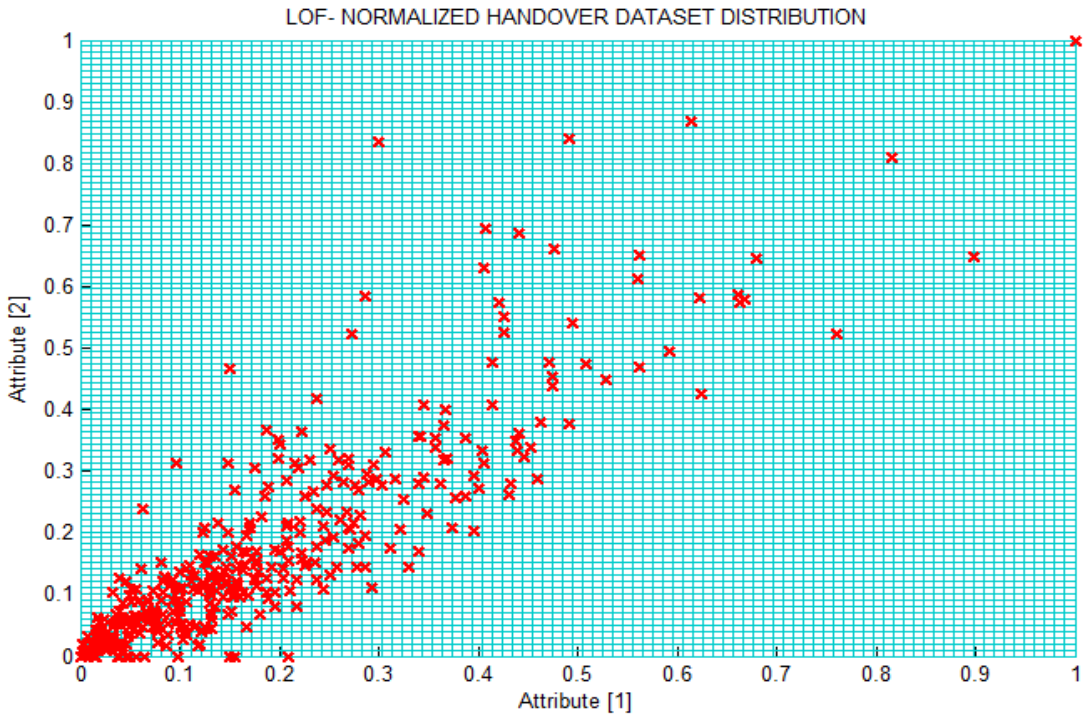

Figure 6. 5: Normalized dataset distributions.

Considering the four steps process model in section 6.1, LOF detection algorithms take a finite number of dataset, in our case 401 data points, and consider user defined parameters such as *K* and LOF score threshold values and then generates a list of LOF score values.

6.2.3.1. Original LOF detection result

The LOF detection is based on spatial study of the different cells at a single temporal instance. Based on the real network scenario in section 6.2.1 and preprocessed input dataset, the LOF scores of each datasets of the 401 cells, as it is show in Figure 6.6, has been obtained as a graphical description result after LOF detection process. It can be observed that LOF score values of cells with greater than a threshold value, which is 2, have been identified as cells in outage whereas the rest cells with lower LOF score compare to threshold have been considered as normal cells. Appendix-A shows the result of each cell LOF score in tabular form that represents a graphical output in Figure 6.6. The yellow line is the threshold LOF score value and the x axis represents each cell ID and y axis represents LOF score value.

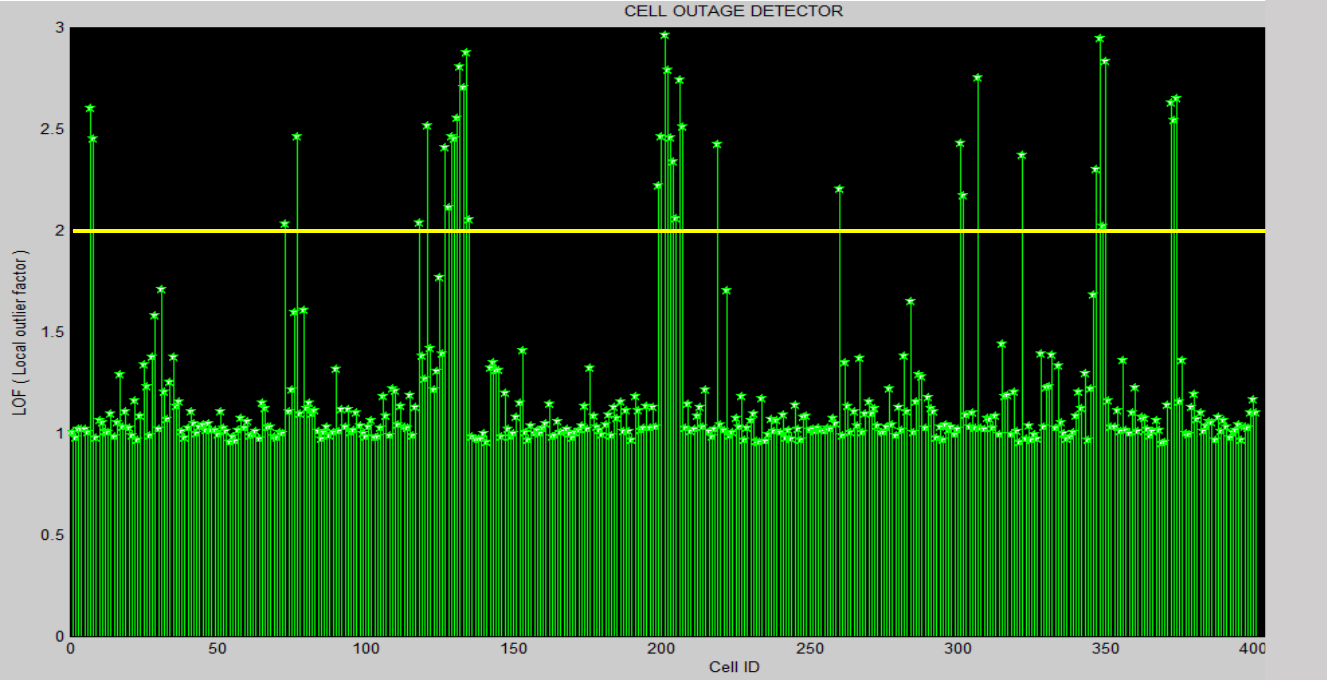

Figure 6. 6: LOF scores of each cell under investigation.

In total, 37 targeted cell outage situations are observed from a total of 401 cells under investigation in LOF detection algorithm analysis. 364 have been identified as normal cells. Among 37 targeted cells, 27 cells were detected as outage cells and 10 of them are normal but detected as outage. Moreover, among 364 cells that has been considered as normal, only 355 are actually normal cells. The result of LOF detection is shown in Table 6.4. The FADD

COD algorithm has also been used to assess this situation and the results has been described in section 6.2.3.2.

Parameters	Results
K	*16*
Threshold value	*2*
Total cells number	*401*
Normal cells	*364*
Targeted cells	*37*
Active cells but detected (potential false positives)	*10*
Outage cells but Undetected (potential false negative)	*9*
Cells in Outage detected	*27*

Table 6. 4: Original LOF COD results.

In order to localize a cell whether it is in outage or normal one, cell ID, LOF score and the geographical location of a map need to be correlated. To visualize that, as it is illustrated in Figure 6.7, google map has been used to mark the cell location with its status. The LOF scores associated with their status are indicated by the shape and color of the point on the map. The outage cells are marked by red star with yellow circle whereas normal cell sites are marked with green point with yellow circle. When all cells of the site are in outage, the location is marked with red star.

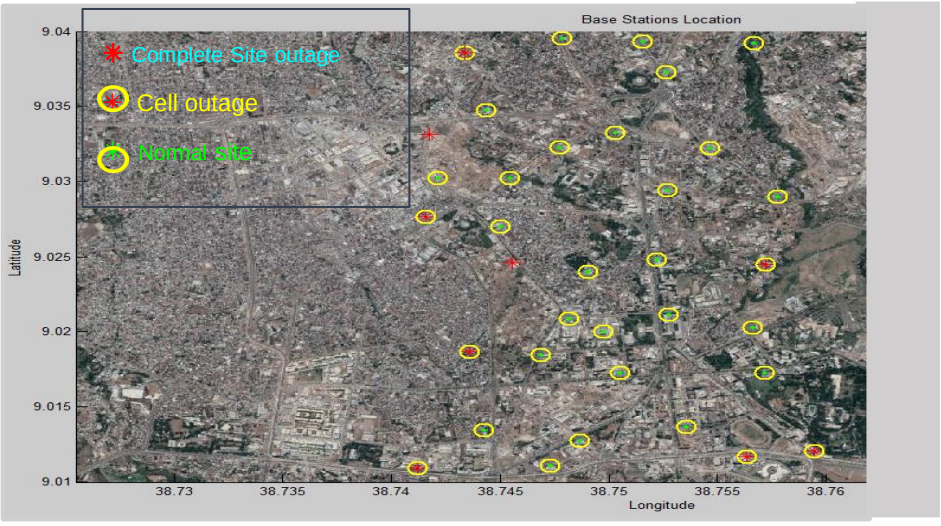

Figure 6. 7: Location and detected cell status (using original LOF).

Table 6.4 designates that the proposed original LOF algorithm is able to detect most cell outages. The percentage false positive rate of LOF detection algorithm in this case is 0.0275. FPR is calculated as the number of normal cells, classified as outlier, divided by the number of normal cells. The best false positive rate is 0.0 whereas the worst is 1.0. The TPR of the LOF detection algorithm is 0.750. Generally, LOF detection algorithm exhibits 75% detection capability to the desirable accuracy of 100% (1.0).

6.2.3.2.Modified LOF (FADD) detection result

In FADD detection algorithm, the same COD procedure has been followed as that of LOF COD. The modified LOF (FADD) tries to overcome dataset duplication problem of LOF when number of duplication points are greater than the value of *K*. Consequently, outage cells can be easily distinguished from the normal cells. According to the graphical output of FADD LOF in Figure 6.8, cells with larger FADD LOF score values than the normal threshold value have been classified as cells in outage. Moreover, cells with indexed as 7, 8, 260, 301, 302, 307 and 322 which were detected as outage cells by LOF, have scored lower values and considered as normal cells by FADD LOF. In addition to that 9 undetected outage cells in the case of LOF reduced to 4 during FADD LOF detection. The rest of the cells with lower FADD LOF score compare to threshold have been identified as normal cells like that of LOF detection algorithm..

Parameters	Results
K	*16*
Threshold value	*2*
Total cells number	*401*
Normal cells	*362*
Targeted cells	*35*
Active cells but detected (potential false positives)	*3*
Outage cells but Undetected (potential false negative)	*4*
Cells in Outage detected	*32*

Table 6. 5: FADD COD results.

Appendix-A shows the result of each cell FADD LOF score in tabular form for Figure 6.8 graphical presentation

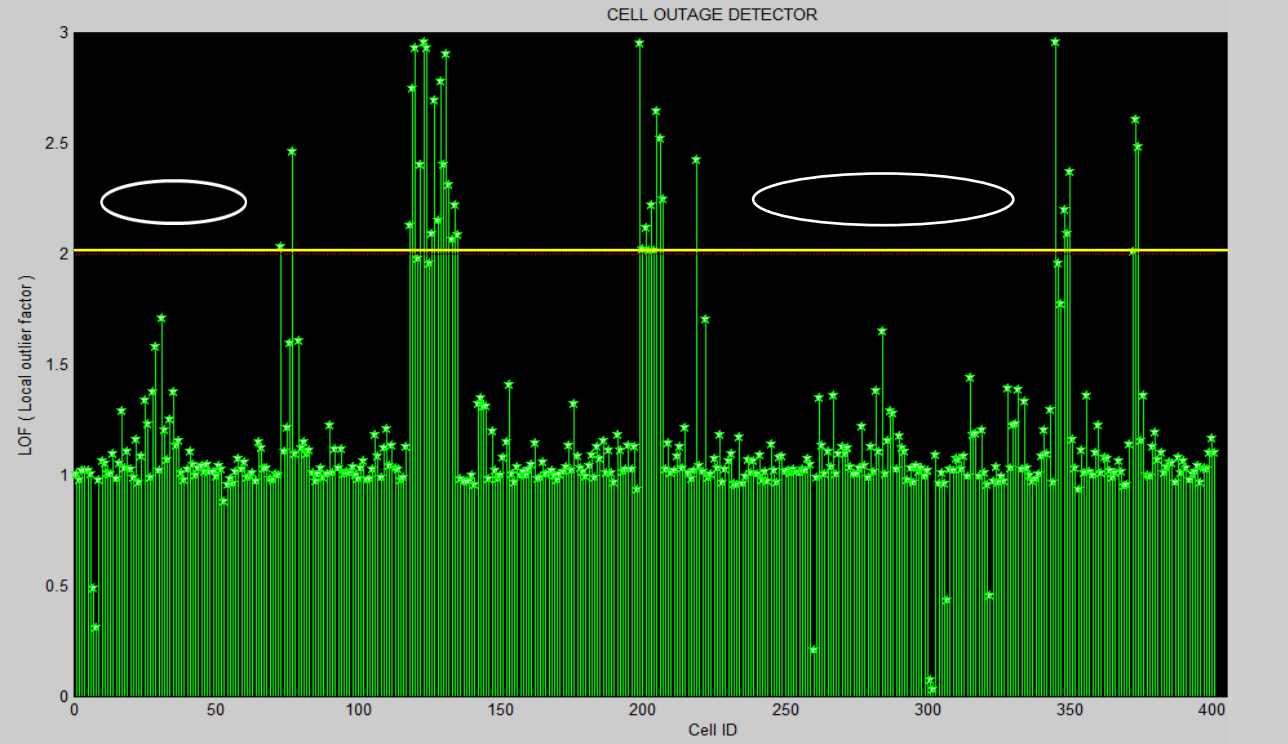

Figure 6. 8: FADD LOF scores of each cell under investigation.

Table 6.5, which is the result of FADD LOF detection, indicates that a total of 35 target cells as outage and 366 cells as normal have been identified based on the value of each FADD LOF score. The FPR of FADD LOF detection algorithm is 0.0082 that is relatively smaller compare to original LOF. Moreover, its TPR is 0.889 which is much better than original LOF.

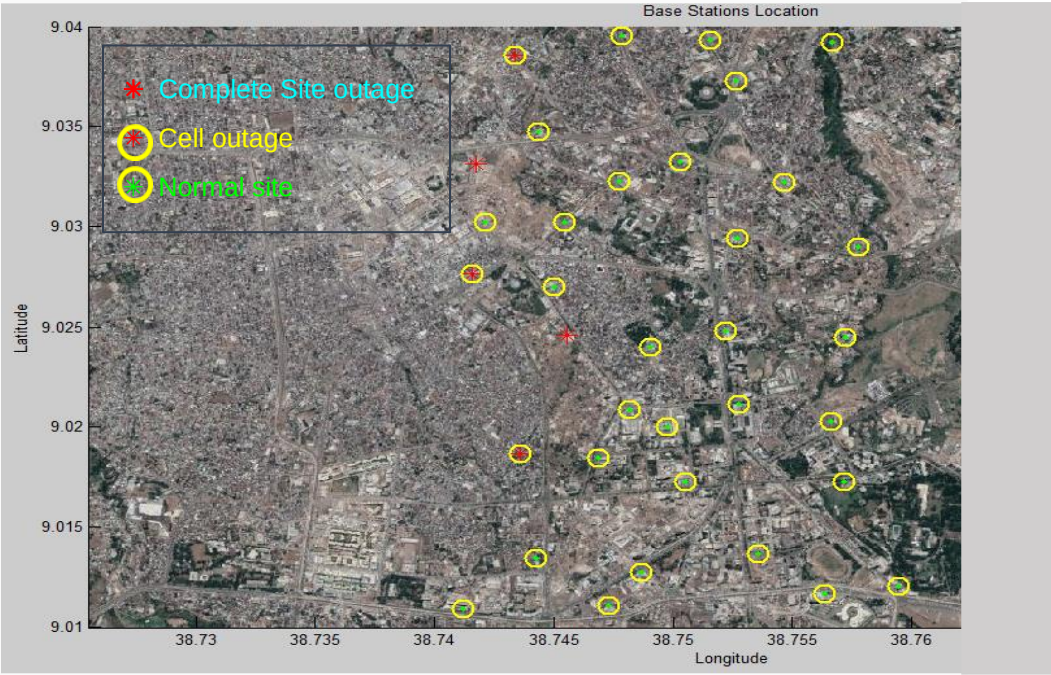

Figure 6. 9: Site location and detected cell status (using FADD).

Locations of cells in outage as well as normal cells were displayed on the google map based on the FADD LOF score and related cells status. It is illustrated in Figure 6.9.

6.3.Evaluation of the Detection Algorithms

The two common methods that represent the performance of binary classifiers are ROC and precision-recall curves. The ROC curve is a well-known graphical method that can display the accuracy of a binary classifier by plotting the TPR against the FPR at various threshold settings, whereas precision-recall shows the relation between precision and recall. The ROC curves also contain the operating point which indicates the actual algorithm's performance. A better algorithm has a higher precision and recall and a lower FPR.

- Precision is the number of abnormal data points, which are classified as outliers, divided by the total number of outlier data points.
- Recall or TPR is the number of abnormal data points, which are classified as outliers, divided by the total number of abnormal data points.
- FPR is the number of normal data points, which are classified as outlier, divided by

the total number of normal data points.

	Original LOF	FADD LOF
TPR	75%	89%
FPR	2.74%	0.82%

Table 6. 6: Performance evaluation.

After analyzing the proposed COD algorithms, it is possible to make a comparison between them based on the area under the ROC curve (AUC) that can display the performance accuracy of a binary classifier (detector). In this thesis, we evaluated the performances of both LOF and FADD LOF algorithms using the AUC. The larger the area under the curve, the better accuracy detection can be achieved. As shown in Figure 6.10, the FADD LOF approach outperforms the original LOF in terms of accuracy. The AUC values of FADD LOF and original LOF detection algorithms are 0.889 and 0.750 respectively. It seems that the difference is small in value, but it is very significant in practice when we see it in network availability and revenue generation.

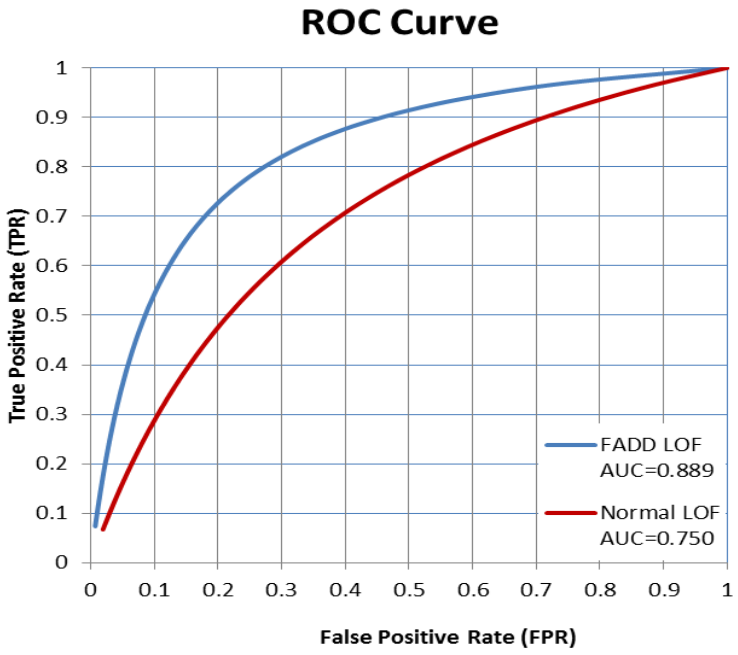

Figure 6. 10: ROC curve comparison between original LOF and FADD LOF.

Both detection algorithms have achieved over 75% for FPR of 0.5. The LOF terms have a 75% TPR and 2.74% FPR (the false positive cases are 10 out of 401 datasets) whereas for the FADD LOF have obtained a 89% TPR and less than 1% FPR (3 false positive out of 401). The result indicates that FADD LOF preforms better with 0.89 AUC unit value than the original LOF.

Chapter VII

7. Conclusion and Future works

7.1. Conclusion

In this thesis work, a proactive COD scheme which is based on LOF detection techniques has been proposed and the number of cell based UMTS incoming HO statistical data that was available as a counter and collected from Ethio telecom OSS system has been utilized. LOF as one of the detection algorithms uses the similarity measure between different points and assigns each point a degree of being an anomaly. Most COD algorithms implemented in the literatures are based on detecting cells in outage by monitoring KPIs and signal measurements from the cell in outage. All these methods can be represented by KPIs available approach in the network. This situation occurs in UMTS when the outage does not affect the Node-B and send KPIs to the OSS, whereas this thesis approach is detecting cell in outage differently by using HO information from its neighboring cells and reveal the status of the cell in outage. It means that the detection model identifies a cell outage without considering alarms or KPIs from the affected Node-Bs.

Real network Ethio telecom Addis Ababa UMTS network scenario and incoming HO statistical data have been considered for the investigation. Data has been collected, preprocessed and tests have been carried out to evaluate the proposed detection algorithms. The experiment has shown that original LOF and modified LOF (FADD) detection algorithms can be used for COD by determining the nominal value of LOF score which can be set to limit the level of normality or abnormality of the cell. Moreover, ROC curve is applied for the purpose of performance evaluation of the detection techniques.

The test results show that both the proposed COD algorithms are able to detect most outage situations which can be associated with various network elements faults. Moreover Based on the numerical analysis FADD LOF algorithm has performed better than the original LOF detection algorithm.

7.2. Future works

In the future, this thesis work can be extended to detect cell outage by incorporating some signal parameter measurements such as RSRP, RSRQ, and neighbours cell list together with incoming HO statistical data to increase the performance of the proposed cell outage detection algorithms. The parameters can be sorted out from the literatures based on the objective and approach of the study. Moreover the aim of this thesis is only to detect cell outage using LOF algorithm. But, the detection algorithm can be considered and used, as one possible line of study, to detect and diagnosis other kind of network faults.

References

[1]. GSMA, "The Mobile Economy 2018", Report [online], P. 12, 2018. Available:*https://www.gsma.com/mobileeconomy/wp-content/uploads/2018/05/The-mobile-Economy-2018.pdf.*

[2]. U. R. Kamboh, Q. Yang, M. Qin "Impact of Self-Organizing Networks Deployment on Wireless Service Provider Businesses in China", *Int. Journal Communications, Network and System Sciences*, Vol.10, pp. 78-89, May 26, 2017.

[3]. M. Amirijoo et al., "Cell Outage Management in LTE Networks", *IEEE/ ISWCS-2009*, Siena, Italy, pp. 600-604, February 2009.

[4]. R. Barco, P. Lázaro, and P. Muñoz, "A unified framework for self-healing in wireless networks", *IEEE Communications Magazine*, vol. 50, pp 134-142, December 2012.

[5]. A. Zoha, A. Saeed, A. Imranz , M. A. Imrany and A. Abu-Dayya, "Data-Driven Analytics for Automated Cell Outage Detection in Self-Organizing Networks", in Design of Reliable Communication Networks (DRCN), *11th international Conference 2015*, pp. 203-210, March 2015.

[6]. F. Wanrong, T. Yinglei, M.Yi, S. Mei, " Cell Outage Detection Based on Improved BP Neural Network in LTE System "Department of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China, 2015.

[7]. J. Han, M. Kamber and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann, 2012.

[8]. I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, *Data Mining Practical Machine Learning Tools and Techniques*, 4th ed., Cambridge: Morgan Kaufmann, 2017.

[9]. Ethio telecom Press Release, "Ethio telecom ranked first in Africa" [November 10, 2017], Available:*http://www.ethiotelecom.et/?q=node/968948, Accessed: December 21, 2017 09:18*

[10]. Ethio telecom National Network Operation Center six month report "Network Trouble Tickets Management Report ", January 10, 2018

[11]. S. Hämäläinen, H. Sanneck and C. Sartori, *LTE Self-Organizing Networks (SON): Network Management Automation for Operational Efficiency*, ISBN 9781119970675. John Wiley & Sons, 2012.

[12]. F. Nisa, S. Haryadi, "Simulation of The Fault Management with Self-Healing Mechanism (Case study: LTE Network in Banda Aceh Area)", *IEEE 10th International Conference on Telecommunication Systems Services and Applications (TSSA)*, Bali, Indonesia, pp 1-6, 2016.

[13]. P. Szil´agyi and S. Nov´aczki , "An Automatic Detection and Diagnosis Framework for Mobile Communication Systems", *IEEE/IFIP transactions on Network and Service Management*, vol. 9, no. 2, pp 184—197, June 2012.

[14]. S. Rezaei, H. Radmanesh, P. Alavizadeh, H. Nikoofar and F. Lahouti, "Automatic Fault Detection and Diagnosis in Cellular Networks Using Operations Support Systems Data", *IEEE/IFIP Network Operations and Management Symposium, Istanbul*, Turkey , pp 468-473, 2016.

[15]. I. De-La-Bandera, R. Barco, P. Mu˜noz, and I. Serrano, "Cell Outage Detection Based on Handover Statistics," *IEEE Communications Letters,* vol. 19, no. 7, pp. 1189–1192, 2015.

[16]. E. Rozaki, "Design and implementation for automated network troubleshooting using data mining", *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol.5, no.3, pp 9-27, May 2015.

[17]. T. Zhang, L. Feng, P. Yu, S. Guo,W. Li, and X. Qiu, "A handover statistics based approach for Cell Outage Detection in self-organized Heterogeneous Networks," in *Proceedings of the 15th IFIP/IEEE International Symposium on Integrated Network and Service Management, IM 2017*, pp. 628–631, May 2017.

[18]. Y. Ma, M. Peng, W. Xue, and X. Ji, "A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks," in Proceedings of the 2013 *IEEE Wireless Communications and Networking Conference, WCNC 2013*, pp. 2266–2270, chn, April 2013.

[19]. Y. Zaki, *Future Mobile Communications*, Advanced Studies Mobile Research Center Bremen, DOI 10.1007/978-3-658-00808-6_1, © Springer Fachmedien Wiesbaden 2013.

[20]. H. Harri and T. Antti, *WCDMA for UMTS – HSPA evolution and LTE*, 4th Edition, John Wiley & Sons Ltd, Chichester, England, 2007.

[21]. F. Khan Muhammad, "Femto-cellular Aspects on UMTS Architecture Evolution", AALTO University, Espoo, Finland, April 20, 2010

[22]. I. de-la Bandera, P. Munoz, I. Serrano, and R. Barco, "Improving cell outage management through data analysis," *IEEE Wireless Communications*, vol. PP, no. 99, pp. 2–8, 2017.

[23]. ETSI TS 123 002 V3.6.0 (2002-09). Digital cellular telecommunications system (Phase 2+); UMTS; Network Architecture (3GPP TS 23.002 version 3.6.0 Release 1999), 2002.

[24]. Jonathan P. Castro. *The UMTS Network and Radio Access Technology: Air Interface Techniques for Future Mobile Systems*. Wiley, West Sussex, England, 2001.

[25]. Behjati M, Cosmas JP, Nilavalan R, Araniti G, Condoluci M."Self-organising comprehensive handover strategy for multi-tier LTE-advanced heterogeneous networks" *IET Science, Measurement & Technology*. 2014; 8(6):441-451.

[26]. Ian Poole (*Radio electronics.com),* Available:*https://www.radioelectronics.com/info/cellulartelecomms/umts/umts-wcdma-handover-handoff.php*" UMTS *WCDMA Handover" 08/09/2018 4:45.*

[27]. K. Kawamura et al., "Management System for Mobile Networks", *FUJITSU Sci. Tech. J.*, Vol. 48, No. 1, pp. 47–53(January 2012).

[28]. T. Bandh, "Coordination of autonomic function execution in Self-Organizing Networks," PhD Thesis, Technische Universität München, Germany, April 2013.

[29]. T. Anttalainen, *Introduction to Telecommunications Network Engineering*, 2nd ed., Artech House, London, 2003.

[30]. S. Chernov, M. Cochez, and T. Ristaniemi, "Anomaly detection algorithms for the sleeping cell detection in LTE networks," *in 2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2015.

[31]. E. J. Khatib, R. Barco, P.Muñoz, I. de-la-Bandera and I. Serrano, "Self-healing in mobile networks with big data", *IEEE Communications Magazine*, vol. 54, no. 1, pp. 114-120, January 2016.

[32]. S. Robinson, "K-Nearest Neighbors Algorithm in Python and Scikit-Learn", [February 15, 2018],Available:*https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/, Accessed: 08/06/2018 12:05.*

[33]. S. Chernov, F. Chernogorov, D. Petrov, and T. Ristaniemi, "Data mining framework for random access failure detection in LTE networks," *IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC),* pp. 1321–1326, IEEE, 2014.

[34]. O. Onireti, A. Zoha, J. Moysen, A. Imran, L. Giupponi, M. A. Imran, and A. Abu-Dayya, "A cell outage management framework for dense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 2097–2113, April 2016.

[35]. W. Xue, M. Peng, Y. Ma, and H. Zhang, "Classification-based approach for cell outage detection in self-healing heterogeneous networks," *in 2014 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 2822–2826, April 2014.

[36]. A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A SON solution for sleeping cell detection using low-dimensional embedding of mdt measurements," *in 2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC)*, pp. 1626–1630, Sept 2014.

[37]. A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "Data driven analytics for automated cell outage detection in self-organizing networks," *in Design of Reliable Communication Networks (DRCN), 2015 11th International Conference*, pp. 203–210, March 2015.

[38]. A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A learning-based approach for autonomous outage detection and coverage optimization," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 3, pp. 439–450, 2016.

[39]. F. Chernogorov, T. Ristaniemi, K. Brigatti, and S. Chernov, "N-gram analysis for sleeping cell detection in LTE networks," *in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4439-4443, May 2013.

[40]. Y. Ma, M. Peng, W. Xue, and X. Ji, "A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks," *in 2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 2266–2270, April 2013.

[41]. S. Chernov, D. Petrov, and T. Ristaniemi, "Location accuracy impact on cell outage detection in LTE-A networks," *in 2015 International Wireless Communications and Mobile Computing Conference (IWCMC),* pp. 1162–1167, Aug 2015.

[42]. I. de-la Bandera, R. Barco, P. Munoz, and I. Serrano, "Cell outage detection based on handover statistics," *IEEE Communications Letters*, vol. 19, pp. 1189–1192, July 2015.

[43]. P. Mu˜noz, R. Barco, I. Serrano, and A. G´omez-Andrades, "Correlation based time-series analysis for cell degradation detection in SON," *IEEE Communications Letters*, vol. 20, no. 2, pp. 396–399, 2016.

[44]. Peng Yu, Fanqin Zhou, Tao Zhang, Wenjing Li, Lei Feng, and Xuesong Qiu "Self-Organized Cell Outage Detection Architecture and Approach for 5G H-CRAN", *Hindawi Wireless Communications and Mobile Computing*, Volume 2018, Article ID 6201386, May 2018.

[45]. KDnuggets, "What main methodology are you using for your Analytics, data mining, or data science projects?", *ULR:https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html*, Accessed on 03.09.2018, 5:50.

[46]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and T. Widener, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27-34, 1996.

[47]. K. B. Agyapong, J.B Hayfron-Acquah, M. Asante, "An Overview of Data Mining Models (Descriptive and Predictive)", *International Journal of Software & Hard ware Research in Engineering*, Volume 4, pp. 53–60, May, 2016.

[48]. F. E Grubbs, "Procedures for Detecting Outlying Observations in Samples", *Tech-nometrics*, Vol. 11, No. 1, pp. 1, 1969.

[49]. M. M Breunig, H. P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density based Local Outliers", *SIGMOD Record*, Vol. 29, pp. 93-104, 2000.

[50]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey", ACM Computing Surveys, Vol. 41, no.3, pp. 15:1-15:58, 2009.

[51]. C. C. Aggarwal, *Outlier Analysis*, 2nd ed., Springer International Publishing, Cham, Switzerland, 2017.

[52]. J-Y. Lee, U. Kang, D. Koutra, C. Faloutsos, "Fast anomaly discovery given duplicates," *Carnegie-Mellon University, School of Computer Science*, Dec. 2012, CMU-CS-12-146.

Appendix A

Result of LOF and FADD detection algorithms cell scores.

Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score
1	1.002215	1.002215	51	1.104828	1.042113	101	1.027525	1.027525	151	1.075476	1.075476
2	0.975115	0.975115	52	1.02628	1.02628	102	1.064442	1.064442	152	1.149294	1.149294
3	1.017219	1.017219	53	1.008327	0.878266	103	0.974981	0.974981	153	1.40457	1.40457
4	1.017137	1.017137	54	0.955582	0.955582	104	0.980966	0.980966	154	1.005208	1.005208
5	1.020897	1.020897	55	0.979625	0.979625	105	1.026734	1.026734	155	0.9633	0.9633
6	1.004459	1.004459	56	0.961912	0.961912	106	1.180171	1.180171	156	1.037316	1.037316
7	2.52085	0.48816	57	1.013984	1.013984	107	1.081144	1.081144	157	0.996141	0.996141
8	2.219077	0.307789	58	1.071503	1.071503	108	0.985158	0.985158	158	1.01423	1.01423
9	0.975097	0.975097	59	1.026025	1.026025	109	1.219529	1.121945	159	0.997846	0.997846
10	1.061349	1.061349	60	1.057934	1.057934	110	1.209459	1.209459	160	1.021674	1.021674
11	1.0505	1.0505	61	0.988685	0.988685	111	1.040813	1.040813	161	1.043465	1.043465
12	1.006516	1.006516	62	0.990113	0.990113	112	1.132028	1.132028	162	1.140224	1.140224
13	1.000228	1.000228	63	1.005544	1.005544	113	1.030394	1.030394	163	0.983364	0.983364
14	1.091893	1.091893	64	0.969759	0.969759	114	1.02391	1.02391	164	0.986973	0.986973
15	0.980706	0.980706	65	1.145342	1.145342	115	1.184303	0.973503	165	1.054909	1.054909
16	1.052446	1.052446	66	1.123618	1.123618	116	0.987679	0.987679	166	1.001854	1.001854
17	1.287988	1.287988	67	1.029537	1.029537	117	1.126493	1.126493	167	1.007365	1.007365
18	1.028413	1.028413	68	1.028411	1.028411	118	2.842388	2.123729	168	1.020148	1.020148
19	1.103264	1.103264	69	0.980349	0.980349	119	1.376829	2.742381	169	0.998994	0.998994
20	1.02282	1.02282	70	0.975757	0.975757	120	1.265399	2.927996	170	0.973276	0.973276
21	0.986937	0.986937	71	1.004702	1.004702	121	2.662933	1.977872	171	1.000196	1.000196
22	1.160054	1.160054	72	0.998346	0.998346	122	1.414479	2.398926	172	1.013851	1.013851
23	0.96391	0.96391	73	2.027329	2.027329	123	1.210302	2.956242	173	1.034481	1.034481
24	1.081535	1.081535	74	1.103833	1.103833	124	1.304834	2.927996	174	1.132779	1.132779
25	1.33548	1.33548	75	1.213795	1.213795	125	1.766635	1.953069	175	1.020961	1.020961
26	1.226225	1.226225	76	1.595176	1.595176	126	1.389928	2.088852	176	1.318457	1.318457
27	0.984485	0.984485	77	2.457483	2.457483	127	2.816235	2.981176	177	1.081454	1.081454
28	1.375097	1.375097	78	1.096247	1.096247	128	2.793878	2.86199	178	1.026911	1.026911
29	1.580528	1.580528	79	1.602644	1.602644	129	2.469105	2.083821	179	1.011249	1.011249
30	1.017474	1.017474	80	1.12255	1.12255	130	2.309525	2.337712	180	0.991176	0.991176
31	1.709082	1.709082	81	1.14634	1.14634	131	2.687579	2.236129	181	1.041936	1.041936
32	1.202849	1.202849	82	1.093298	1.093298	132	2.986852	2.317805	182	1.090257	1.090257
33	1.067278	1.067278	83	1.112117	1.112117	133	2.769934	2.984448	183	0.988506	0.988506
34	1.25011	1.25011	84	1.008188	1.008188	134	2.829581	2.548251	184	1.128009	1.128009
35	1.371906	1.371906	85	0.969405	0.969405	135	2.706085	2.749251	185	1.074107	1.074107
36	1.133664	1.133664	86	1.004566	1.004566	136	0.979706	0.979706	186	1.151015	1.151015
37	1.154556	1.154556	87	1.029905	1.029905	137	0.968611	0.968611	187	1.009603	1.009603
38	1.007819	1.007819	88	0.988268	0.988268	138	0.969978	0.969978	188	1.11064	1.11064
39	0.97615	0.97615	89	1.0042	1.0042	139	0.969854	0.969854	189	1.007821	1.007821
40	1.022825	1.022825	90	1.316935	1.225115	140	0.996448	0.996448	190	0.967132	0.967132
41	1.107162	1.107162	91	1.010409	1.010409	141	0.955236	0.955236	191	1.181624	1.181624
42	1.043183	1.043183	92	1.114988	1.114988	142	1.317861	1.317861	192	1.110193	1.110193
43	0.998641	0.998641	93	1.031876	1.031876	143	1.34673	1.34673	193	1.012794	1.012794
44	1.028316	1.028316	94	1.117613	1.117613	144	1.308279	1.308279	194	1.021907	1.021907
45	1.042367	1.042367	95	1.002289	1.002289	145	1.308482	1.308482	195	1.132778	1.132778
46	1.011769	1.011769	96	1.012436	1.012436	146	0.978963	0.978963	196	1.024948	1.024948
47	1.044883	1.044883	97	1.097579	1.005804	147	1.195491	1.195491	197	1.12385	1.12385
48	1.012491	1.012491	98	1.03456	1.03456	148	1.01779	1.01779	198	1.029602	0.930714
49	1.012426	1.012426	99	0.999006	0.999006	149	0.981189	0.981189	199	2.595336	2.841852
50	0.99059	0.99059	100	0.981091	0.981091	150	0.997034	0.997034	200	2.752874	2.16689

Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score	Cell ID	LOF Score	FADD Score
201	2.752874	2.16689	251	1.020711	1.020711	301	2.724166	0.075038	351	1.158221	1.158221
202	2.496722	2.903098	252	1.009189	1.009189	302	2.613682	0.030132	352	1.027698	1.027698
203	2.865133	2.105124	253	1.023678	1.023678	303	1.086645	1.086645	353	1.029879	0.931961
204	2.068028	2.745093	254	1.01302	1.01302	304	1.028971	0.961737	354	1.111565	1.111565
205	2.968546	2.729372	255	1.008477	1.008477	305	1.097421	1.005658	355	1.006171	1.006171
206	2.098756	2.71747	256	1.02507	1.02507	306	1.024348	0.958755	356	1.355808	1.355808
207	2.546977	2.133432	257	1.017813	1.017813	307	2.782968	0.432736	357	1.013755	1.013755
208	2.40297	2.445789	258	1.074984	1.074984	308	1.025087	1.025087	358	0.998629	0.998629
209	1.024251	1.024251	259	1.046116	1.046116	309	1.018966	1.018966	359	1.100186	1.100186
210	1.143255	1.143255	260	2.10704	0.20816	310	1.06992	1.06992	360	1.221078	1.221078
211	1.006842	1.006842	261	0.988909	0.988909	311	1.069267	1.069267	361	1.00927	1.00927
212	1.016504	1.016504	262	1.347558	1.347558	312	1.026238	1.026238	362	1.079707	1.079707
213	1.082912	1.082912	263	1.132725	1.132725	313	1.083203	1.083203	363	1.072536	1.072536
214	1.127206	1.127206	264	1.000971	1.000971	314	0.991053	0.991053	364	1.017843	1.017843
215	1.027582	1.027582	265	1.106684	1.106684	315	1.436528	1.436528	365	0.984654	0.984654
216	1.211509	1.211509	266	1.034934	1.034934	316	1.178698	1.178698	366	1.007581	1.007581
217	1.009956	1.009956	267	1.370515	1.356866	317	1.187274	1.187274	367	1.059932	1.059932
218	0.982247	0.982247	268	1.001206	1.001206	318	0.990107	0.990107	368	1.011615	1.011615
219	1.020302	1.020302	269	1.092027	1.092027	319	1.201279	1.201279	369	0.951575	0.951575
220	2.421746	2.421746	270	1.153963	1.12543	320	1.005606	1.005606	370	0.953579	0.953579
221	1.040311	1.040311	271	1.095593	1.095593	321	0.956075	0.956075	371	1.13523	1.13523
222	1.0151	1.0151	272	1.119085	1.119085	322	2.566621	0.454915	372	2.871452	2.677725
223	1.703597	1.703597	273	1.037276	1.037276	323	0.970908	0.970908	373	2.507602	2.805838
224	0.98732	0.98732	274	1.001992	1.001992	324	1.035175	1.035175	374	2.788823	2.531243
225	1.004261	1.004261	275	1.002418	1.002418	325	0.966681	0.966681	375	1.154788	1.154788
226	1.074479	1.074479	276	1.027563	1.027563	326	0.993427	0.993427	376	1.35833	1.35833
227	1.02886	1.02886	277	1.216826	1.216826	327	0.967869	0.967869	377	0.999761	0.999761
228	1.181833	1.181833	278	1.040613	1.040613	328	1.38917	1.38917	378	0.991379	0.991379
229	0.963199	0.963199	279	0.984367	0.984367	329	1.030572	1.030572	379	1.124296	1.124296
230	1.023194	1.023194	280	1.126271	1.126271	330	1.223296	1.223296	380	1.191187	1.191187
231	1.060974	1.060974	281	1.011185	1.011185	331	1.231028	1.231028	381	1.068584	1.068584
232	1.091484	1.091484	282	1.380461	1.380461	332	1.385032	1.385032	382	1.100491	1.100491
233	0.959879	0.959879	283	1.103706	1.103706	333	1.022821	1.022821	383	1.006413	1.006413
234	0.954542	0.954542	284	1.649655	1.649655	334	1.329283	1.329283	384	1.034679	1.034679
235	1.170023	1.170023	285	1.004456	1.004456	335	1.050218	1.030655	385	1.055605	1.055605
236	0.957661	0.957661	286	1.152393	1.152393	336	0.999893	0.999893	386	1.052011	1.052011
237	0.990564	0.990564	287	1.285665	1.285665	337	0.970964	0.970964	387	0.965125	0.965125
238	1.064862	1.064862	288	1.275024	1.275024	338	0.981987	0.981987	388	1.078507	1.078507
239	1.010661	1.010661	289	1.025953	1.025953	339	1.001404	1.001404	389	1.00595	1.00595
240	1.060681	1.060681	290	1.176449	1.176449	340	1.08425	1.08425	390	1.062046	1.062046
241	1.005296	1.005296	291	1.121327	1.121327	341	1.200605	1.200605	391	1.027484	1.027484
242	1.087877	1.087877	292	1.103759	1.103759	342	1.120836	1.093069	392	0.9739	0.9739
243	0.97511	0.97511	293	0.978354	0.978354	343	1.291189	1.291189	393	1.00212	1.00212
244	1.011025	1.011025	294	1.02952	1.02952	344	0.964145	0.964145	394	1.012514	1.012514
245	0.972828	0.972828	295	0.965647	0.965647	345	1.216002	2.956242	395	1.039333	1.039333
246	1.137807	1.137807	296	1.041133	1.041133	346	1.678479	1.953069	396	0.962656	0.962656
247	1.020637	1.020637	297	1.011202	1.011202	347	2.29867	1.774058	397	1.028904	1.028904
248	0.965062	0.965062	298	1.027912	1.027912	348	2.811319	2.508787	398	1.026272	1.026272
249	1.079152	1.079152	299	0.994577	0.994577	349	2.576776	2.53049	399	1.09882	1.09882
250	1.083548	1.083548	300	1.017227	1.017227	350	2.944029	2.859717	400	1.16279	1.16279
									401	1.100747	1.100747