A MASTER'S THESIS SUBMITTED TO JIMMA UNIVERSITY

INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

IMPROVED CELLULAR NETWORK FAULT PREDICTION USING MACHINE LEARNING: IN THE CASE OF ETHIO TELECOM NODE B NETWORK, ETHIOPIA

By:

ABDI SHESHIFA ABDURAHMAN

A THESIS SUBMITTED TO THE FACULTY OF COMPUTING AND INFORMATICS OF JIMMA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR DEGREE OF MASTER OF SCIENCE IN COMPUTER NETWORKING

Dec 2021

JIMMA, ETHIOPIA

APPROVAL SHEET

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

As thesis research advisor, we hereby certify that **Abdi Sheshifa** thesis work entitled" **IMPROVED CELLULAR NETWORK FAULT PREDICTION USING MACHINE LEARNING: IN THE CASE OF ETHIO TELECOM NODE B NETWORK, ETHIOPIA**" is under our guidance. I recommended that it be submitted as fulfilling the thesis requirement

Fisseha Bayu (**Ph.D. Cand**) ____________________ ____________________

Advisor Date Signature

Mr. Tesfu Mekonen ____________________ ____________________

Co-advisor Date Signature

As a member of the board of examiners of the Master of Science thesis open defense examination, we examined the thesis prepared by **Abdi Sheshifa** and certified the candidate. We recommended that the thesis be accepted as fulfilling the thesis requirement for Degree of Master of Science in Computer Networking.

____________________ ____________________ ____________________

External examiner Date Signature

Kebebew Ababu (**Asst. Prof**.) ____________________ ____________________

Internal examiner Date Signature

____________________ ____________________ ____________________

Chairperson Date Signature

DECLARATION

I declare that the thesis work entitled **“IMPROVED CELLULAR NETWORK FAULT PREDICTION USING MACHINE LEARNING: IN THE CASE OF ETHIO TELECOM NODE B NETWORK, ETHIOPIA”** is my original work and under the guidance of **Fisseha Bayu (Ph.D. Cand)** and **Mr. Tesfu Mekonen**. All sources of materials used for the thesis have dully acknowledged and are being submitted to the Jimma University in partial fulfillment of the requirements for the award of Master of Computer networking.

I also hereby declare that this work in part or full has not been submitted to any other university for any Degree or Diploma.

Name: Abdi Sheshifa ________________

Signature

Place: Jimma

Date of Submission: _________________

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

Fisseha Bayu (PhD Cand) _________________ _________________

Advisor Signature Date

Mr. Tesfu Mekonen _________________ ________________

Co-advisor Signature Date

ACKNOWLEDGMENTS

Firstly, I would like to acknowledge everyone who played a role in my academic accomplishments, Especially My Advisor, Fisseha Bayu (Ph.D. Candidate), for his unwavering encouragement, input, and revisions. I'd also like to thank Mr. Tesfu Mekonen (M.Sc.), my co-advisor, for his continuous, Advice and feedback on my thesis.

Secondly, I'd like to express my gratitude to Mr. Kebebew Ababu (Assistant Prof.) and Mr. Getamesay Haile (M.Sc.) for their valuable feedback on my thesis. Each of whom has provided patient and guidance throughout the research process. Thank you all for your unwavering support.

Lastly, I want to thank all those who, in one way or another, understood how tedious the four-year program was for me and sacrificed their time with me, allowing me to finish my research, who supported me with love and understanding, Especially to My Wife.

Abstract

In today's telecommunications settings, operators must deal with rapid technological developments while improving operational efficiency, which is, lowering operational costs while maximizing network performance. The growing amount of services available and subscribers has put cellular network service providers under a lot of strain. Subscribers to cellular networks have varying wants and requirements. Ethio telecom is one of the biggest telecom service providers in Africa and also the only one in Ethiopia. Its vision is to become a world-class telecom service provider.

The primary aim of the research in this study is to relook at the concepts of cellular Network Fault Management proactive measure from a new perspective. The fault happened daily and need to rectify, so fault management regarding the faults of the cellular network needs proper attention and urgent solution. Because the company is currently using traditional methods of Network Management System (NMS), but in this thesis from drive test data, to improve ensure customer satisfaction, Cellular network fault of cellular must be managed properly. A Viable Solution to enhance its service available proactively is needed. It is, thus, with this intent that the researcher is motivated and decided to figure on the problem during this thesis.

So far, most studies on Cellular network fault have focused on existing measurement of data available. Fault predicts were made by using real Ethio telecom EMS (Element Management System) data and achieved Previously done fault predicted using Time series considering down site NN(Neural Network) based and root cause analysis, however, the study had gaps, so the researcher was motivated to fill the gap.

In this study, the researcher has tried to focus on the issue rather than existing measurement data from EMS (Element Management System) based to focus on No direct measurement data available. So improved prediction of Node B network fault empowered drive test. The result from DT (Drive test) data collected for two months shows that the performance of the Naïve Bayes with an accuracy of 98%.

Keywords: Cellular Network Fault, Machine Learning, Naïve Bayes, Bayesian Network, drive test, Node B, Predictive model

Table of Contents

List of Figure

List of Table

List of Acronyms

BBU	Baseband Unit
BTS	Base Transceiver Station
CDP	Conditional Probability Distribution
CPT	Conditional Probability Table
CNMS	Cellular Network Management System
CN	Core Network
CRISP-DM	Cross-industry standard process for data
DS	Down Site
DT	Drive Test
EE	Engineering Error
EMS	Element Management System
FDS	Frequently down site
RBS	Radio base Station
UMTS	Universal mobile telecommunication system
Qos	Quality of service
TT	Trouble Ticket
OSS	Operation support system
SMC	Service Management center
SON	Self-Organizing Network
O&M	Operational and Maintenance
Node B	UMTS Standard
UE	User Equipment’s
RAN	Radio Access Network
3GPP	3rd Generation Partnership Project
ME	Mobile Equipment’s
USIM	UMTS Subscriber Identity Module
HSDPA	High Speed Downlink Packet Access

HSUPA	High speed Uplink Packet Access
KPI	Key performance Indicator
RSCP	Received Signal Call Power
UNC	Undefined Neighbor cell
SL	Signal Level
SQ	Signal Quality
HoF	Handover Failure
INT	Interference
PP	Pilot Pollution
Pip	Ping Pong
SS	Same scrambling code
Ovs	Overshooting
Cov	Coverage
MN	Missing Neighbor
ML	Machine Learning
RCA	Reversely Connected Antenna
RC	Root cause
NV	Naïve Bayes
SVM	Support Vector Machine
UTRAN	Terrestrial Radio Access Network

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the study

The growing amount of services available and subscribers has put cellular network service providers under a lot of strain. Subscribers to cellular networks have varying wants and requirements. This necessitates optimal network performance at all times to attract and keep consumers while also ensuring service quality. This can happen if the network is properly operated and maintained [1]. In today's, Telecommunication networks ensuring excellent service quality and avoiding service disruptions is important to play due to a major role in society as they support the transmission of information between businesses, governments, and individuals, Hence For this purpose, fault management is critical. It consists of fixing network problems and detecting, isolating, a task that requires considerable resources and complexity led for large networks, to improve various aspects of fault management an emerging research area is to develop machine learning and data mining-based techniques [9].

The technique for avoiding catastrophic network failures is to foresee cellular network faults, where these faults are estimated before they occur. Cellular faults must be carefully managed to improve service quality. The goal of fault prediction is to forecast when systems or components are about to fail [2]. Network faults can be classified into two groups, malfunctions and outages [16]. Malfunction is said to occur when the active network elements (NE) may be working with some errors in some sense but not well, while outages occur when the active network elements are completely knocked out and do not work at all. Malfunctions are normally characterized by performance degradation in various performance parameters, for example, unclear reception, increase in noise level, increase in delay, Failure prediction approaches have been divided into three types, including experience-based, condition-based, and model-based methods [57]…etc.

So to intervene in the fault, Considered Models of propagation in cellular systems that are tailored to certain locations. When these models are applied in different environments, their efficiency suffers. Cellular mobile network's capacity, Quality of Service (Qos), and coverage are affected by shift inefficiency. (DT) Drive testing is a technique for determining and assessing the quality of service (Qos) of a mobile network and identifying and correcting faults. Different sites of recent technology 2G, 3G, HSPA, WCDMA, and Long Term Evolution (LTE), these different sites analyses with different parameters like Rx Lev, Rx Qual, and BER, etc… [3]

The Telecom networks sectors contain thousands of various hardware elements of various types: BS(base station), routers, servers, switching units, modems, cables, fittings, Air Conditioners, etc. a stimulating cause analysis of mobile site outage was conducted using the Bayesian Networks Model by Mesfin Geremew and Ephrem Teshale. (i.e., Base stations, servers, and routers, switching units, cables, cooling elements) aren't well-defined explicitly. For example, the paper emphasized: The relationship between everyone the hardware installed within the room and the cooling system that controls an area temperature isn't defined through a failure will probably affect a smooth running of the hardware within the air conditioning. A reason is needed to link the network element to another existing incident and an incident produced, creating a child-parent relation. The outcome of Gere mew's analysis is the BTS mobile site network outage Model-based Root Causes Analysis."[4]

The model can help inform field technicians of the real extent of failures and the probable existence of root problems, which optimizes resources and reduces restored time, but not consider network fault at locations or times from which no direct measurement real data is available.

There are various research conducted in Cellular Network fault like Bayesian network consists of the evaluation of the probabilities associated with the occurrence of the fault of one or more, based on the collected data from the system under investigation The alerts created during the operation of the controlled network elements, or information received as a result of prior correlation operations, make up the majority of this data. [15-16].

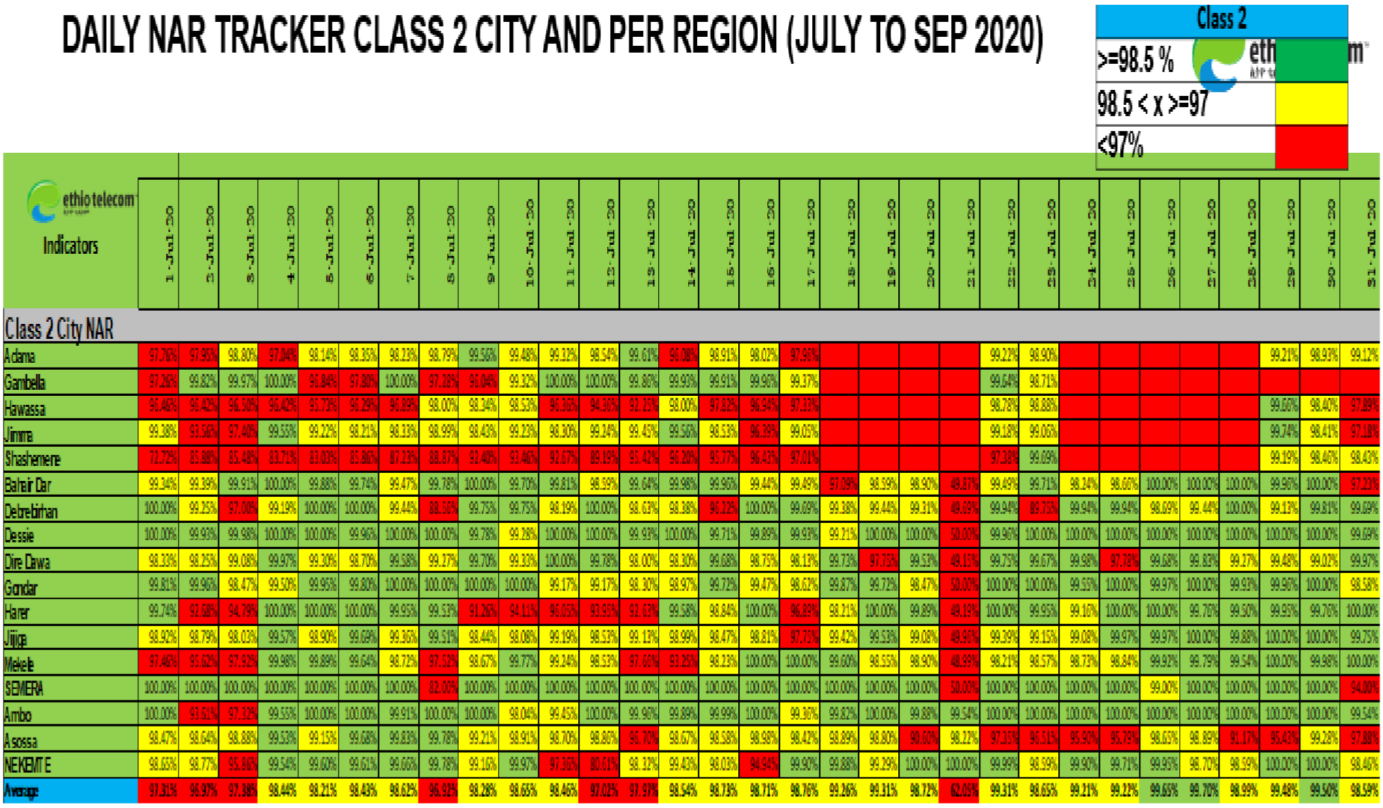
In this thesis, consider ML from the normal approach to assessing signal quality parameters in a wireless network that requires performing what is known as a driving test. A driving test consists of driving a vehicle equipped with a wireless measurement toll box around the area of interest, to take measures of several wireless signal parameters. The slowness of the process makes this technique impractical when the goal is to measure the impact caused by minor changes on the network; when you need something like a trial-and-error approach; or if you are acting proactively were trying to have an early alert on signal degradation before clients complain start.

Other conducted BS Level fault Prediction did but analyzing the core network metrics to determine the specific status of the site because if the site is active, it will interact with the core network. Such prediction needs evidence [1-2]. This kind of fault prediction scenario is the most challenging and requires statistics technique analysis rather than an alarm. The researcher started by assuming site fault predicting the fault's inter-arrival time value to predict fault, but he, finally, predicts the target site status as fault based on the predicted history data as objective[5][6].

However, the author recommended the inclusion of site-level statistical data to achieve a higher level of accuracy than what was attained by the study. Thus, it is with this intention that this thesis is designed to predict fault by applying machine learning technique NN on the Ethio telecom Node B network by considering fault data to predict the fault in a pre-set time interval. So the objective of this thesis is to the improved prediction of Node B network fault empowered by drive test data which is no direct measurement is available in the OSS, so this motivated the researcher to fill the gap.

Radio Base stations, backhaul, and transmission networks (Tx) all work together to deliver the essential service in a cellular network. These networks administer employs Element Management Systems (EMS) for site-based Follow-up. Fault management checks the status of the network and reports those faults that occur in it such as unclear reception of Error, BS outage, and equipment failure. This can be established by different alarms generated from the NE's like SO-ME, log file, and DT data or the observed abnormal behavior of the network [28].

Besides below Fig 1.1 and above stated management and daily dispatched Availability rate, hence due to that the availability degraded focus on the reason of that degraded NUR only but not the area where data available.



DAILY NAR TRACKER CLASS 2 CITY AND PER REGION (JULY TO SEP 2020)

Class 2	
>=98.5 %	🟩
98.5 < x >=97	🟨
<97%	🟥

Indicators	1-Jul-20	2-Jul-20	3-Jul-20	4-Jul-20	5-Jul-20	6-Jul-20	7-Jul-20	8-Jul-20	9-Jul-20	10-Jul-20	11-Jul-20	12-Jul-20	13-Jul-20	14-Jul-20	15-Jul-20	16-Jul-20
Class 2 City NAR																
Adama	97.76%	97.95%	98.80%	97.04%	98.14%	98.35%	98.23%	98.79%	99.56%	99.48%	99.32%	98.54%	99.61%	96.08%	98.91%	98.02%
Gambella	97.26%	99.82%	99.97%	100.00%	96.84%	97.80%	100.00%	97.28%	96.04%	99.32%	100.00%	100.00%	99.86%	99.93%	99.91%	99.96%
Hawassa	96.46%	96.42%	96.50%	96.42%	95.73%	96.29%	96.89%	98.00%	98.34%	98.53%	96.36%	94.36%	92.25%	98.00%	97.82%	96.94%
Jimma	99.38%	93.56%	97.40%	99.55%	99.22%	98.21%	98.33%	98.99%	98.43%	99.23%	98.30%	99.24%	99.45%	99.56%	98.53%	96.39%
Shashemene	72.72%	85.88%	85.48%	83.71%	83.03%	85.86%	87.23%	88.87%	92.40%	93.46%	92.67%	89.19%	95.42%	96.20%	95.77%	96.43%
Bahir Dar	99.34%	99.39%	99.91%	100.00%	99.88%	99.74%	99.47%	99.78%	100.00%	99.70%	99.81%	98.59%	99.64%	99.98%	99.96%	99.44%
Debrebirhan	100.00%	99.25%	97.00%	99.19%	100.00%	100.00%	99.44%	88.56%	99.75%	99.75%	98.19%	100.00%	98.63%	98.38%	96.22%	100.00%
Dessie	100.00%	99.93%	99.98%	100.00%	100.00%	99.96%	100.00%	100.00%	99.78%	99.28%	100.00%	100.00%	99.93%	100.00%	99.71%	99.89%
Dire Dawa	98.33%	98.25%	99.08%	99.97%	99.30%	98.70%	99.58%	99.27%	99.70%	99.33%	100.00%	99.78%	98.00%	98.30%	99.68%	98.75%
Gondar	99.81%	99.96%	98.47%	99.50%	99.95%	99.80%	100.00%	100.00%	100.00%	100.00%	99.17%	99.17%	98.30%	98.97%	99.72%	99.47%
Harer	99.74%	92.68%	94.79%	100.00%	100.00%	100.00%	99.95%	99.53%	91.26%	94.11%	96.05%	93.95%	92.63%	99.58%	98.84%	100.00%
Jijiga	98.92%	98.79%	98.03%	99.57%	98.90%	99.69%	99.36%	99.51%	98.44%	98.08%	99.19%	98.53%	99.13%	98.99%	98.47%	98.81%
Mekele	97.46%	95.62%	97.92%	99.98%	99.89%	99.64%	98.72%	97.52%	98.67%	99.77%	99.24%	98.53%	97.66%	93.25%	98.23%	100.00%
SEMERA	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	82.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Ambo	100.00%	93.61%	97.32%	99.55%	100.00%	100.00%	99.91%	100.00%	100.00%	98.04%	99.45%	100.00%	99.96%	99.89%	99.99%	100.00%
Asossa	98.47%	98.64%	98.88%	99.53%	99.15%	99.68%	99.83%	99.78%	99.21%	98.91%	98.70%	98.86%	96.70%	98.67%	98.58%	98.98%
NEKEMTE	98.65%	98.77%	95.86%	99.54%	99.60%	99.61%	99.66%	99.78%	99.16%	99.97%	97.36%	80.61%	98.32%	99.43%	98.03%	94.94%
Average	97.31%	96.97%	97.38%	98.44%	98.21%	98.43%	98.62%	96.92%	98.28%	98.65%	98.46%	97.02%	97.97%	98.54%	98.73%	98.71%

Indicators	17-Jul-20	18-Jul-20	19-Jul-20	20-Jul-20	21-Jul-20	22-Jul-20	23-Jul-20	24-Jul-20	25-Jul-20	26-Jul-20	27-Jul-20	28-Jul-20	29-Jul-20	30-Jul-20	31-Jul-20
Class 2 City NAR															
Adama	97.96%					99.22%	98.90%						99.21%	98.93%	99.12%
Gambella	99.37%					99.64%	98.71%								
Hawassa	97.33%					98.78%	98.88%						99.66%	98.40%	97.89%
Jimma	99.05%					99.18%	99.06%						99.74%	98.41%	97.18%
Shashemene	97.01%					97.38%	99.09%						99.19%	98.46%	98.43%
Bahir Dar	99.49%	97.09%	98.59%	98.90%	49.87%	99.49%	99.71%	98.24%	98.66%	100.00%	100.00%	100.00%	99.96%	100.00%	97.23%
Debrebirhan	99.09%	99.38%	99.44%	99.31%	49.69%	99.94%	89.75%	99.94%	99.94%	98.69%	99.44%	100.00%	99.13%	99.81%	99.69%
Dessie	99.93%	99.21%	100.00%	100.00%	50.00%	99.96%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.69%
Dire Dawa	98.13%	99.73%	97.75%	99.53%	49.15%	99.75%	99.67%	99.98%	97.78%	99.68%	99.83%	99.27%	99.48%	99.02%	99.97%
Gondar	98.62%	99.87%	99.72%	98.47%	50.00%	100.00%	100.00%	99.55%	100.00%	99.97%	100.00%	99.93%	99.96%	100.00%	98.58%
Harer	96.89%	98.21%	100.00%	99.89%	49.19%	100.00%	99.95%	99.16%	100.00%	100.00%	99.76%	99.50%	99.95%	99.76%	100.00%
Jijiga	97.75%	99.42%	99.53%	99.08%	49.96%	99.39%	99.15%	99.08%	99.97%	99.97%	100.00%	99.88%	100.00%	100.00%	99.75%
Mekele	100.00%	99.60%	98.55%	98.90%	48.99%	98.21%	98.57%	98.73%	98.84%	99.92%	99.79%	99.54%	100.00%	99.98%	100.00%
SEMERA	100.00%	100.00%	100.00%	100.00%	50.00%	100.00%	100.00%	100.00%	100.00%	99.00%	100.00%	100.00%	100.00%	100.00%	94.00%
Ambo	99.36%	99.82%	100.00%	99.88%	99.54%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.54%
Asossa	98.42%	98.89%	98.80%	90.60%	98.22%	97.35%	96.51%	95.90%	95.79%	98.65%	98.89%	91.17%	95.43%	99.28%	97.88%
NEKEMTE	99.90%	99.88%	99.29%	100.00%	100.00%	99.99%	98.59%	99.90%	99.71%	99.95%	98.70%	98.59%	100.00%	100.00%	98.46%
Average	98.76%	99.26%	99.31%	98.72%	62.05%	99.31%	98.65%	99.21%	99.22%	99.65%	99.70%	98.99%	99.48%	99.50%	98.59%

Figure 1. 1 Prepared by SMC | November 2020 |Corporate dashboard

Despite centrally dispatched TT, the NE availability also encompasses DT analysis affected area concerning Customer complaints the design target of network availability of mobile network sites is 99.978% [29]. Figure 1 indicates the operational performance of mobile sites in terms of availability. There is a performance gap from the design target due to faults and resulting in high revenue loss happened per year.

This thesis case focus due to fault frequently registered and raised complaint from the customer side It requires a lot of resources to intervene in the issue, hence, high cost due to repetitive maintenance without knowing the fault, while network planning and defining neighbor relationship configuration; BTs activation requires manual software loading and data configuration mistakes might be quite often while the neighbor configuration is made manually; cellular network fault is made manually causing delays in recovering the fault. However, these weaknesses of the traditional system can be minimized by deploying SON and automating it. But need to consider the No direct real data measurement available.

The motivation of the study is The Mobile Site seems working well from the management side, but customers keep complaining the NE is running however, it hinders both client and operator, the cellular network is a gateway to human operation. And also the motivation of conducting this research is, encompassing locations or times from which no direct measurement available that affects the user. When radio base station network fault happens like call and internet issue customer may not get desired service quality identifies so in this thesis proposed work Using DT analysis and apply ML that is different from EMS because customer uses the degraded service, they will wait until service resumed the faults and the fault persists for a long time, which has a significant impact on companies image and revenues.

1.2 Statement of the Problem

Ethio telecom is one of the biggest telecom service providers in Africa and the only service provider in Ethiopia. Its vision is to become a world-class telecom service provider. Recently, the Ethiopian government has decided to liberalize the arena and invited other operators to hitch the business. Thus, it's believed that other national and/or international telecom operators will join the service market shortly. To become competent within the sector, Ethio telecom is functioning hard towards tackling its challenges and thereby improving the availability and Qos of those sites. One of the most challenges that Ethio telecom faces is that the issue related to its Cellular Network Management System (CNMS) with handling Customer Complaints.

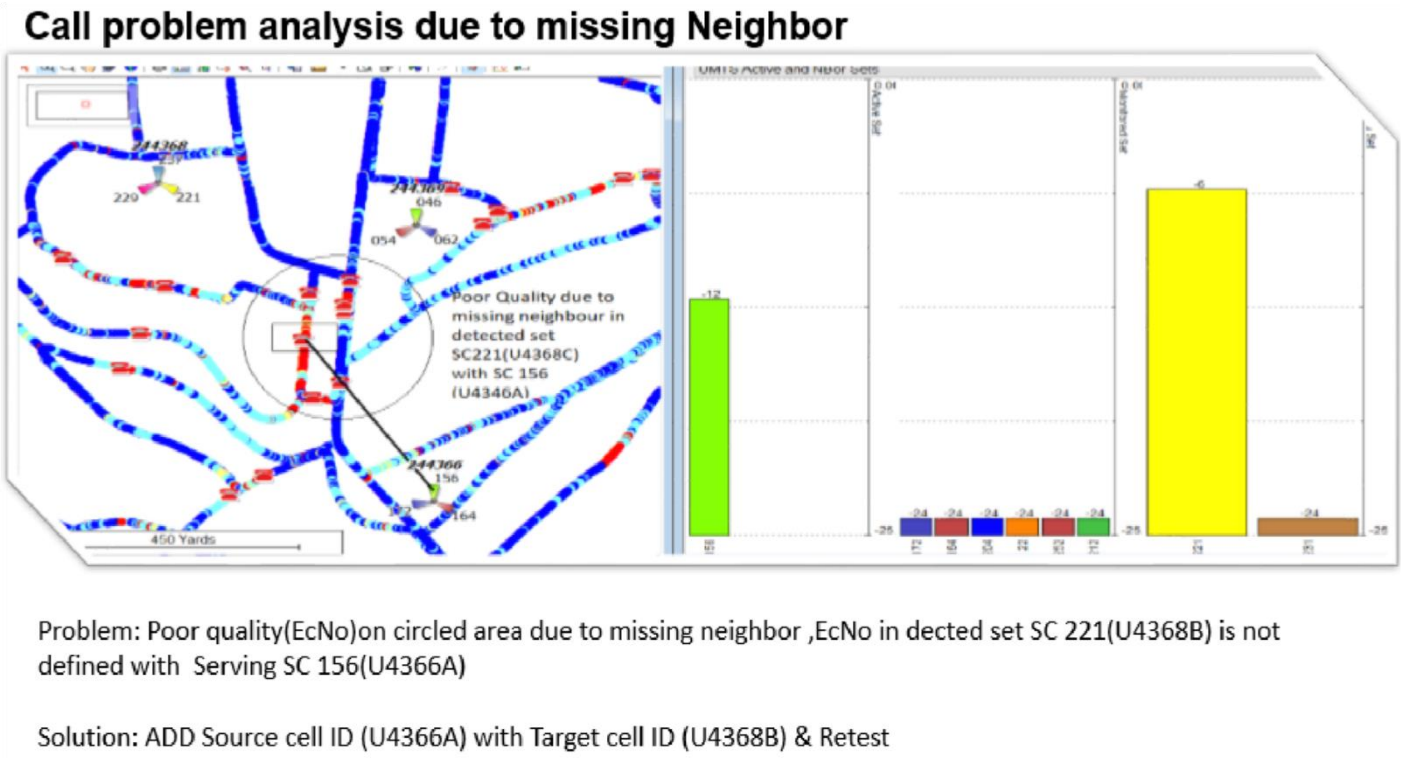
Among the functions of which is fault management regarding the faults of cellular network that needs proper attention and urgent solution. Because the company is currently using traditional methods of Network Management System (NMS),

Using ML from Drive test, especially which of its fault-related issue may be a viable solution to enhance its service availability proactively and ensure customer satisfaction. It is, thus, with this intent that the researcher is motivated and decided to figure on the problem during this thesis. To improve service quality, the fault of cellular must be managed properly.

In another study, fault prediction was made by using real Ethio telecom EMS data and achieved 71.7% accuracy [1] using NN. Previously done fault predicted using only down site NN based, however the study as the fault was made based direct measurement data.

Synchronization problems may be a symptom of fault on the backhaul rather than an indication that the site is down. This type of symptom requires other evidence to learn the nature of the problem. This type of symptom requires other evidence to learn the nature of the problem.

However, the author recommended the inclusion of Cell level statistical data to achieve a higher level of accuracy than what was attained by the study. Thus, it is with this intention, thesis is designed to apply the machine learning BN technique on the Ethio telecom Node B network by considering fault complained by a user using DT fault analyses fault with RC. The objective of this thesis is to improve the prediction of Node B network fault using a Bayesian network through NB empowered by drive test logs. Most Customer complaint fault ET handled by Using the below sample Missing Neighbor Fig.1.1



Call problem analysis due to missing Neighbor

Poor Quality due to missing neighbour in detected set SC221(U4368C) with SC 156 (U4346A)

Problem: Poor quality(EcNo)on circled area due to missing neighbor ,EcNo in dected set SC 221(U4368B) is not defined with Serving SC 156(U4366A)

Solution: ADD Source cell ID (U4366A) with Target cell ID (U4368B) & Retest

Figure 1. 2 Call Problem due to Missed Neighbor 2020 [Ac tix Analyzer]

Network quality issues can be exacerbated by an insufficient neighbor cell plan. The impact of a missing neighbor cell on network performance is dependent on the coverage provided by other cells in the same area. A missing neighbor definition might have a major performance impact in certain situations, as per the above-shown SC221 (U4368B), especially if few other cells give coverage and the lost adjacency would have been the main HO target. These facts point to the requirement for a proactive mechanism that examines existing neighbor cells and flags any that are missing for every operating cell.

So considering the huge number (above 9000) of mobile sites infrastructure the Ethio telecom which includes new sites (2G, 3G, and LTE), and the significant number of daily mobile site faults, the quality of service is negatively affected. This implies, there is a need for improvement of mobile sites by digging the reason why the fault is happened frequently and not consistence the availability of the mobile network. Thus, it is essential for researchers on improving mobile sites' fault reason analysis, which helps to mitigate the faults by implementing proactive maintenance techniques. This, as a case including drive test analysis and, is focused on sensitivity to faults area.

1.3 Objective of Study

1.3.1 General Objective of Study

The general objective of this research is to improve cellular network fault Prediction DT based using machine learning.

1.3.2 Specific Objective of Study

To achieve the general objective of the study, the following specific objectives are identified:

- To explore related works that have more understanding of cellular network fault issue approaches used in the ML environment.
- Identify Best Algorithm
- Prepare and preprocess data sets for simulation.
- Train the dataset by using the algorithms
- Evaluate the model.
- Design new architecture for ML Environments.

1.4. Methodology

1.4.1 Literature Review

To achieve the desired result, use the following methods. Different kinds of literature which were done on the wireless network on fault Cellular network area reviewed to know the state of the art and to find the best methods that can help us to solve the problem.

1.4.2 Simulation tools

To do this research, work used different tools than necessary to collect the data and implement the proposed system. Using Act ix software, Mobile and dongle Google Earth are selected to get organized the desired data. MATLAB (R2021a): MATLAB is a high-performance numerical computation programming language. MATLAB allows for matrix manipulations, function and data plotting, and algorithm implementation, and it comes with several toolboxes that can be used in algorithm development, data acquisition, modeling, simulation, prototyping, math and computation, data analysis, exploration, and visualization, a detailed description of these tools is presented below.

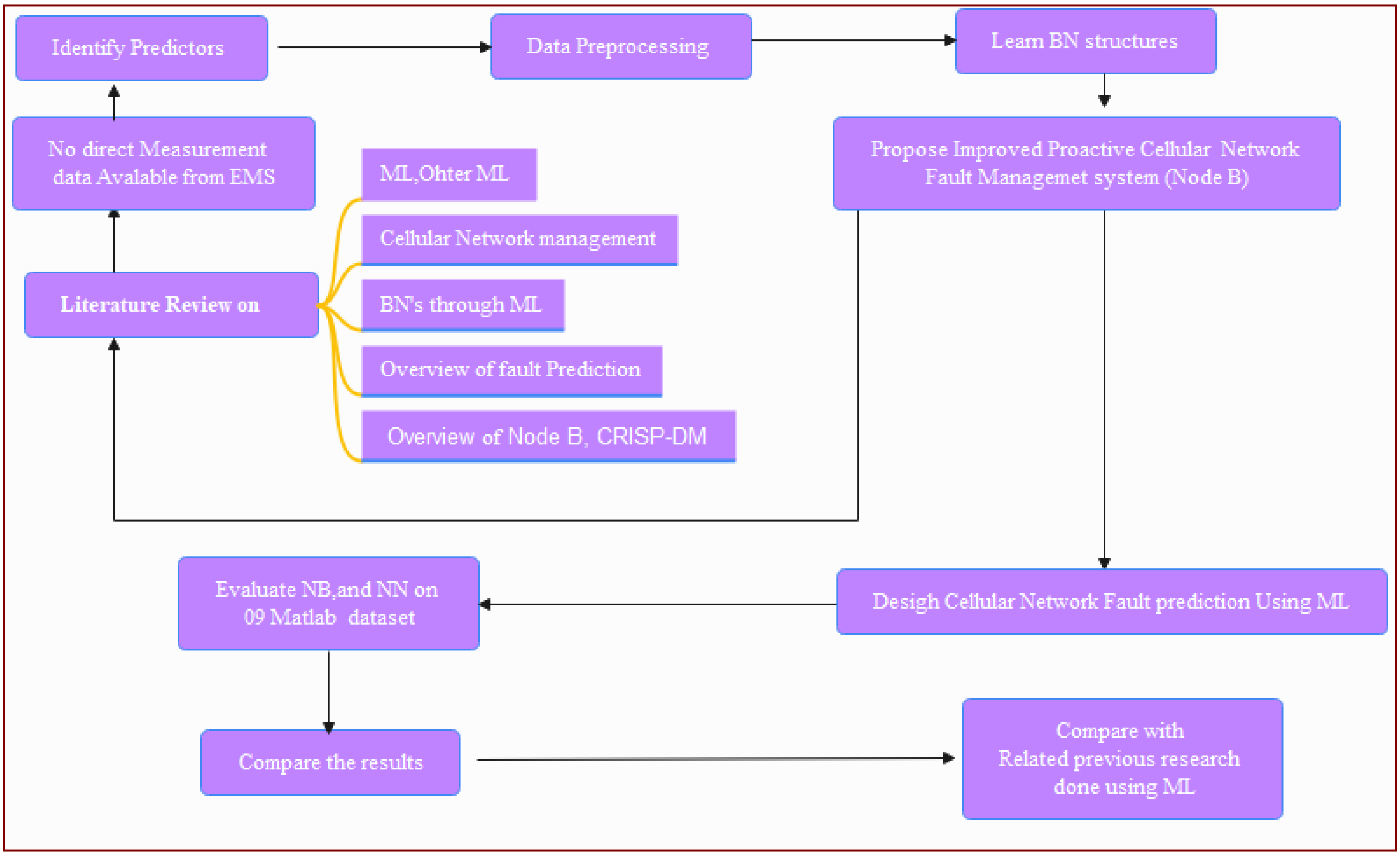

Identify Predictors
Data Preprocessing
Learn BN structures
No direct Measurement data Avalable from EMS
ML,Ohter ML
Propose Improved Proactive Cellular Network Fault Managemet system (Node B)
Cellular Network management
Literature Review on
BN's through ML
Overview of fault Prediction
Overview of Node B, CRISP-DM
Evaluate NB,and NN on 09 Matlab dataset
Desigh Cellular Network Fault prediction Using ML
Compare the results
Compare with Related previous research done using ML

Figure 1. 3 Research Methodology

1.5 Scope and Limitation of the study

The scope of this research is to improve a predictive model analysis that identifies cellular network fault from Ethio telecom mobile from DT call-related issues using DT Node B parameters. Cellular network services 3G (Third Generation) call-related issue based on Assumption like each feature equal contribution or Independent.

Also didn’t take into account the Backhaul RC status.

1.6. Application of Results

Engaging fault proactively should help both operator and customers for service delivery or revenue and ensure customer satisfaction, respectively. Maximize the chances of success for some predefined goal, resource utilization and proactively address the issue, and at the same time increase revenue for the company to react proactively.

1.7 Organization of the Thesis

This thesis is organized into six chapters. The first chapter briefly discusses the background to the problem area, states the problem statement, objective of the study, research methodology, scope and limitation of the study, and significance of the results. The second chapter presents a Literature review and Cellular network fault. The third chapter, related work, discusses, Why BN was selected? BN, Cross-industry standard process. In this chapter, the general cellular network architecture and discussed. The fourth chapter deals with the Design of the Proposed System and result interpretations. In this chapter, the building of the model with training datasets and validating the result with testing datasets, and interpretation of the result of the experimentation was a major concern. Chapter, five Implementation, Result, and Discussion, The last chapter, chapter six is devoted to conclusion and Future work forwarded based on the research findings of the present study.

1.9 Significance of the study

The result of this study helps the network providers to improve fault management by engaging proactive measures to satisfy customers. This research also helps the researchers as a ground route to study cellular issues related to DT. Generally, this study is important since its application can dig the issue and provide for the upcoming vendor to compute, deliver acceptable service and ensure customer satisfaction not only this in income generation by reducing customer complaints.

CHAPTER TWO

2. Literature Review

2.1 INTRODUCTION

This chapter examines the Applicable sources and facts, laying the groundwork for future investigation. Analyze these works critically in connection with the research topic under consideration while reviewing them. There are various sections to the review. Overview of Faults Prediction, Overview of Cellular Network Management (CNM), Overview of Machine Learning, Bayesian Networks (BN's) through Machine Learning (ML), Other ML and CRISP-DM (Cross-industry standard process for data mining) Considering Business Aspect, Then go over the Overview of Node B.

2.2 Overview of faults prediction

Prediction refers to the methodical process of determining what will occur under given circumstances. A telecommunications fault is an abnormal operation that severely reduces or interrupts the functioning of an active entity in the network. All errors are not faults because most protocols can deal with them. In general, abnormally [24], [25]

As a result, fault prediction is the technique of predicting which telecommunication problem will occur in a given set of circumstances. In the last few years, some progress has been achieved in this field in terms of study. This includes a classifier training approach for detecting anomaly faults, the application of reinforcement learning for proactive network management, and fault management in communication. Networks in [26].

Because of its ability to prevent failures and maintenance costs, failure prediction is critical for predictive maintenance. At present, mathematical and statistical modeling are the prominent approaches used for failure predictions. These are based on equipment degradation physical models and machine learning methods, respectively. None of these systems guarantees failure predictions long ahead of time, allowing for proactive treatment of potential reasons [53].

The study Bayesian belief network for intelligent fault management systems [27] looked at how to use the Bayesian Belief Network in a fault prediction intelligent monitoring system that uses adaptive statistical approaches to detect defects before they occur, but it does not connect the two. Can detect faults before they occur but do not relate these faults to services. This research presented this prediction for performing using Drive test data through ML.

In thesis focused on due to fault frequently registered and raised complaint from customer side It requires a lot of labor and resource to intervene the issue, hence, high cost due to repetitive maintenance without knowing the fault, while network planning and define neighbor relationship configuration; BTS activation requires manual software loading and data configuration mistakes might be quite often while the neighbor configuration is made manually, and cellular network fault is made manually causing delays in recovering the fault. However, these weaknesses of the traditional system can be minimized by deploying SON and automating it, but need to consider No direct real data measurement available from the existing Managements.

2.3 Overview of Cellular Network Management

In recent years, the heterogeneity and increasing flexibility of resources offered in cellular networks mobile network management systems have seen various developments and advancements, owing to network function virtualization (NFV) [41] allow big data under the paradigm of the software-defined networks (SDN) orchestration integrating tools into operational support systems (OSS). Thus, the functions of self-organizing networks (SON) [42] that are intended to automate many of the tasks performed by network engineers are becoming reasonable increasingly. Despite all these advancements, some issues persist, making cellular network management a difficult undertaking. That is the case for sleeping cells, i.e., this fact is evidenced by incidents that monitoring systems may not notice in the cellular network. Areas that, either due to interference, low coverage, or any other radio problem, cause the low quality of experience (QoE) for the user. Such unnoticed incidents are especially significant in places where events with large concentrations of people. Furthermore, without a user terminal to identify the deep causes and depend just on (KPI) network performance indicator traces, troubleshooting issues in such locations may be challenging. Even though adding user traces to any network operations, administration and maintenance (OAM) activity are very helpful [43], there has yet to be a systematic compilation to serve as a regular input for the SON system. In this context, drive tests used to map network deployments reveal the network's state for a single period and do not capture network change. Although minimizing drive test procedures (MDT) [43] can assist to reduce the high cost of these tests, periodic monitoring that accurately reflects the conditions at any one time is still highly costly. Physical exploration or examination, on the other hand, is preferred by operators. Approaches that did not necessitate active user Participation [40], 2021.

In today's culture, telecommunication networks are essential because they allow information to move to businesses, governments, and individuals. As a result, maintaining excellent service quality and avoiding service interruptions are crucial. In this case, critical fault management. It entails identifying, isolating, and resolving network faults, a difficult task for big networks that often necessitates significant resources. As a result, developing machine learning and data mining-based strategies to improve fault management in various components of the process is becoming an emerging research field [37].

As a result of numerous service interruptions due to fault, periodic maintenance and system updates, and internet, 3G Networks were reportedly unavailable [20](2014). So, wireless networks have the potential to significantly reduce the operational costs for network operators under preferred Optimization and fault management, as well as user experience improve in terms of quality. Currently, Cellular network management involves human interaction ranging from conducting drive tests to evaluate the network fault and coverage, and performance to diagnosing customer complaints [9], (2018). That machine learning (ML) techniques can assist in reducing the need for drive tasks even before they noticeably degrade the quality of service of the network users by helping to predict and diagnose network fault.

The most important concern in the Telecom sector is a cellular network failure. Due to the direct influence on Customers, profits, companies are seeking to develop tools to predict likely client behavior, particularly in the telecom business. As a result, it's critical to identify variables that contribute to client churn and take the required steps to reduce it. Our work's key contribution is the development of a churn prediction model that helps telecom carriers estimate which customers are most likely to leave. The model developed in this work employs machine learning techniques on a big data platform to develop a fresh approach to engineering and feature selection. To measure the performance of the model, the Area under Curve (AUC) standard measure is adopted, and the AUC value obtained is 93.3%. Working on a massive dataset obtained by translating big raw data provided by Syria Tel Telecom Company, the model was constructed and evaluated using the Spark environment. The dataset contained all customers' information over 9 months, and was used to train, test, and evaluate the system at Syria Tel [38], 2019.

This study provides a novel approach for wireless network measurement and fault resolution based on a piece of software called Mobile Qos Agent. According to Soldani's explanation [39], this agent runs on regular mobile phones and measures the quality of service. Based on a configurable profile put on the phone, the mobile agent can perform a variety of tests. Because these measurements are taken on the ground and with actual customer terminals, they provide an authentic reflection of the service quality that customers have experienced. The technique demonstrated that it is possible to turn tens of thousands of customer phones into quality control stations. The approach provided in [39] is intended to be used as a network design and optimization tool for mobile network operators. They stated "the measuring results dramatically minimize the needs of typical drive or walk tests" [39] as one of the advantages of that strategy.

2.4. Overview of Machine learning

By learning from historical data, machine learning allows computer systems to do complicated tasks such as prediction, diagnosis, planning, and recognition. The performance of machine learning models is dependent on data and techniques. Machine learning models can be improved by using high-quality data and big data sets. It's also critical to use the right methods to address various issues, especially those involving various datasets. An overview of the machine learning types and the tasks which can be addressed by machine learning algorithms are presented in Fig. 2.1 and Fig. 2.2, respectively. Machine learning types are explained as follows:

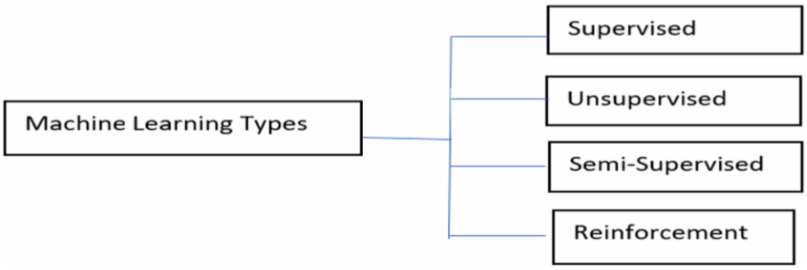


Figure 2. 1. Overview of machine learning types.

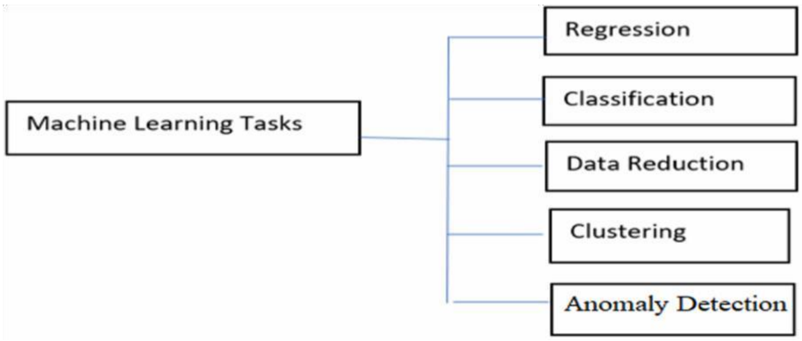


Figure 2. 2. Overview of machine learning tasks.

- **Supervised Learning**: The computer program creates a function between the input(s) and output(s) based on a set of labeled training data (s). Human intervention is very important in supervised learning. People not only label the training set's output, but they also choose features, methodologies, and even control Parameters for algorithms based on a variety of assumptions. A human has specific knowledge and expertise for a model in the vast majority of circumstances. The supervised learning method expects greater data processing for feature selection and anticipates parameter improvement for better algorithm construction.

- **Unsupervised Learning**: This machine learning method does not require labeled data, unlike supervised learning. When the associations between input variables are unknown, unsupervised learning is used. Unsupervised learning, unlike supervised learning, does not produce an output value; instead, it presents a pattern of input variables and, in most cases, different clusters based on the input data.
- **Semi-supervised Learning**: Labeled data is used to train the model in supervised learning, which makes the model more resilient and accurate; nevertheless, the data labeling procedure is costly. Only a small percentage of the data is labeled in several cases, leaving the rest of the data points unidentified. Semi-supervised learning algorithms may train a model with both labeled and unlabeled data, which can improve accuracy over supervised learning algorithms that employ only a small amount of labeled data.
- **Reinforcement Learning**: In reinforcement learning, agents observe their surroundings, take some actions, and receive rewards (negative or positive) based on the actions they take, after which the model is updated. A feedback mechanism is used in reinforcement learning to reward positive behavior and punish poor behavior. Self-driving automobiles and online games like backgammon, for example, use this strategy.

Machine learning techniques are applied in numerous areas to solve different kinds of tasks. Five common tasks are explained as follows [66]:

- **Regression**: The input features are mapped to a numerical continuous variable via regression, also known as value estimation. To obtain a minimum error in the prediction, machine learning methods are utilized to optimize the coefficients of each independent variable. An integer or a floating-point number can be used as the output variable.
- **Classification**: Input features are mapped to one of the discrete output variables by classification. The output variable represents the underlying problem's class. The output variable for binary classification can only be one or zero. The output variable for multi-class classification can have multiple classes.
- **Clustering**: Clustering divides data points into meaningful groups. The similarity pattern between data points is used to group the data. Similar points are clustered together to give data scientists with useful information.
- **Data Reduction**: Due to noisy data instances or repeating data points, data reduction operations can lower the number of numbers and also delete some rows (i.e., data points). Some of the features that are strongly associated or are not very relevant may be deleted from the dataset to create models faster.

This work is mostly used as a supplement to other machine learning tasks like regression and classification.

- **Anomaly Detection**: Unsupervised learning methods are commonly used to solve anomaly detection problems. Anomaly detection algorithms arrange the samples in a similar way to clustering algorithms.

Anomaly detection methods are used to discover outliers in the dataset.

2.4.1 Requirements of Creating Good Machine Learning Systems

As such, what are the requirements for developing these machine learning systems? The following are the requirements for developing such machine learning systems:

- Data: Input data is required for predicting the output.
- Algorithms – Machine Learning is dependent on certain statistical algorithms to determine data patterns.
- Automation: It is the ability to make systems operate automatically.
- Iteration: The complete process is iterative i.e., repetition of the process.
- Scalability: The capacity of the machine can be increased or decreased in size and scale.
- Modeling: The models are created according to the demand by the process of modeling. Machine Learning methods are classified into certain categories these are Fig 2.1

2.5 Bayesian Networks through Machine Learning

Machine learning is a branch of artificial intelligence that allows a system to learn from data. As they learn from the training data, machine learning algorithms will generate increasingly accurate predictive models. Bayesian network learning entails inferring a model, structure, and associated parameters from data. This is naturally separated into two parts: (1) structural learning, which entails determining the network's structure or topology; and (2) parametric learning, which entails determining the associated probabilities given the structure [45].

Machine learning techniques are designed to recognize complicated patterns in a data set automatically, allowing inference or prediction in fresh data sets [46]. Bayesian modeling, on the other hand, is widely used to create algorithms for learning from data. It occurs at the interface of statistics and computer science, with a focus on efficient computer algorithms [46]. In this way, Bayesian networks are statistical techniques that evolved in the field of artificial intelligence and allow us to deal with research circumstances involving numerous variables and complex interactions where uncertainty exists. [46, 47] Pearl

Bayesian networks (BNS) are a simple mathematical language for expressing unambiguous relationships between variables. The benefit of a BN is that it may contain multiple variables, and all nodes and probability tables can be evaluated about the domain; uncertainty can also be controlled, resulting in an explanatory environment that aids decision-making. BNS has comprehensive knowledge of the system's state and can make observations (get evidence) and update the system's probabilities [32]. As a result, a BN can be used to handle a model that characterizes causality in terms of conditional probabilities that represent the conditional independencies of a variable. These in dependencies simplify the representation of knowledge (fewer

Parameters) and reasoning (propagation of probabilities).

2.6 Other Machine learning (ML)

Machine learning (ML) methods are a subfield of artificial intelligence (AI) The ML methods are mainly examined in three main categories as semi-supervised, supervised, and unsupervised algorithms [35]. Supervised learning methods aim to make predictions about unknown situations. Classification, similarity detection, and regression are among the most common tasks of supervised machine learning methods [34-35].

Use Bayesian Networks to investigate and discuss the following common supervised machine learning techniques: Artificial Neural Networks, Nave Bayes Algorithm, Support Vector Machines, and K-Nearest Neighbor Algorithm. Look at each machine learning method as much as possible to compute the best outcomes.

Neural networks (NN), also give good results, but they require large databases with numerous instances, and they ignore the experts' uncertainty and knowledge. Alternatively, the classical model, on the other hand, does not allow for estimates in complex models or small sample numbers [33–34].

The Naive Bayes (NB) Algorithm is a popular machine learning approach based on the Bayes Rule. This strategy is based on the independence of variables and is a classic Bayesian network. The classes that will be approximated using the NB approach must be unrelated to one another. One of the supervised learning algorithms is this one. Despite its simplicity, it gives excellent results in a variety of applications [48].

Support Vector Machines (SVMs) are statistical algorithms that use statistical learning theory to produce a consistent estimator using available data. It tries to divide the data into two basic categories. The n-dimensional hyperplane is produced for this reason. Linear separation of data is possible, system optimization is done in the linear SVM. If not possible, quadratic optimization is provided with the non-linear SVM. Models use kernel functions for this. The selected kernel function affects the performance of the system. Different results can be obtained with different kernel functions [51]

The k-Nearest Neighbor Algorithm (k-NN) determines how data is classified based on its nearest neighbors. This is one of the most commonly used algorithms in data mining. It is preferred because of its simplicity and ease of comprehension. The performance of the algorithm is impacted by the similarity function and the k parameter value. Based on the state of its nearest neighbor, it calculates the chance of a datum being included in the class of its neighbors. It outperforms NN, which is a completely black box in this regard. However, determining the distance between neighbors is difficult [52].

The main distinction between our work and several previous research is that some papers deal with the study of wireless fault networks employing fault like time series analysis on the Management Server side, fault collects from a customer perspective, and other TT or EMS that fetched from Server but no consider and involve using existing measurement data network cause of fault where direct measurements are not available.

In our proposed work focused on relevant issues while considering the Node B network Cell Level.

2.7 CRISP-DM Process Model

CRISP-DM *(*Cross-industry standard process for data mining) describes six major iterative phases, each with its own defined tasks and set of deliverables such as documentation and reports

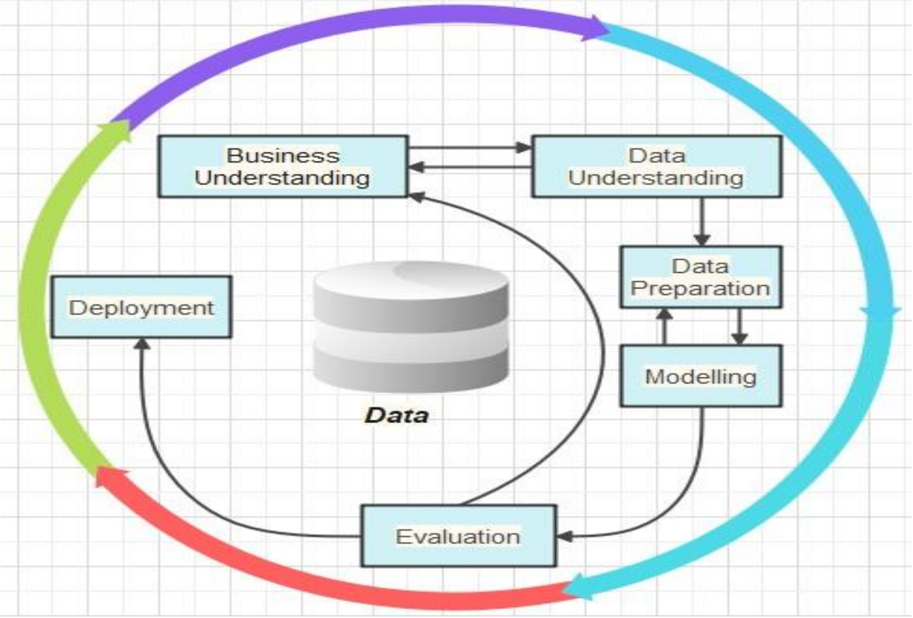


Figure 2. 3. CRISP-DM process diagram.

According to William Vorhies, one of the CRISP creators, DM's "CRISP-DM provides strong guidance for even the most advanced of today's data science activities" because all data science projects start with a business understanding, data that must be gathered and cleaned, and data science algorithms that must be applied (Vorhies, 2016). Believes that all data science projects are doomed to fail. "CRISP-DM gives good direction for even the most advanced of today's data science operations," says the author (Vorhies, 2016).

2.7.1 Business Understanding

This phase focuses on knowing the company and requirements from a business standpoint and then translating that information into a thorough grasp of business objectives, problem characterization, and a preliminary project plan to meet the goals.

When providing telecom services, the telecommunications industry generates and retains a massive amount of data. Fault detail data defines the severity of faults that traverse telecommunication networks; network data describes the status of hardware and software components in the network; and NAR describes telecom faults that affect the service, including user side

These massive amounts of data must be correctly managed for a variety of objectives, including fraud detection, network performance analysis, customer churn prediction, reporting to higher officials, network planning and optimization, and decision support [17].

From the above telecommunications data, network data is used for this research. The first and the major source of data was network alarm data's which are sent from each network element to the Fault Management System during failures. The high volume of data is generated from different network elements and stored in different databases for different purposes [17, 20].

2.7.2 Data Understanding

The data understanding phase begins with data collection and continues with actions to familiarize yourself with the data, uncover data quality issues, gain early insights into the data, or detect intriguing subsets to create hypotheses for hidden information. The two concepts of business and data understanding are inextricably linked. Require at least a basic comprehension of the situation.

2.7.3 Data Preparation

Data are necessary for the discovery of knowledge. Real-world business data, on the other hand, is frequently unavailable. The following step is to prepare data for modeling [21]. Data preparation techniques include data integration, data transformation, data cleansing, data reduction, and data discretization.

2.7.4 Modelling

There are various modeling techniques are chosen and employed during this phase, and their parameters are calibrated to ideal values. For the same problem type, there are usually many methods. Some approaches necessitate the use of specific data formats. Data preparation and modeling are inextricably linked. Often, during modeling, one discovers data issues or gets ideas for creating new data.

2.7.5 Evaluation

Data analysis standpoint, you've constructed one or more models that look to be of excellent quality at this point in the project. Before moving forward with the model's final deployment, it's critical to do a more thorough evaluation of the model and assess the processes taken to build it to ensure that it meets the business objectives. One of the main goals is to see to determine if there is some important business issue that has not been sufficiently considered.

2.7.6 Deployment

The model's creation isn't always the conclusion of the process. Typically, the information gathered must be structured and presented in a way that the client can understand. The deployment process might be as simple as generating a report, depending on the requirements. In many circumstances, the deployment procedures will be carried out by the user rather than the data analyst.

2.8 Overview of Node B

Node B is the third generation (3G) of mobile cellular systems that were developed by 3rd Generation

Partnership Project (3GPP). The rationale behind the development of UMTS is to give a higher data rate and good voice services than its predecessor, the second generation (2G) Global System for Mobile Communications (GSM). As a network architecture, which is an infrastructure that delivers services between endpoints, UMTS has three functional elements; namely, User Equipment (UE), Radio Access Network (RAN), and Core Network (CN) [22]

Three domains make up the UMTS network architecture.

– UMTS Terrestrial Radio Access Network (UTRAN): Provides the air interface access method for the user equipment.

– Core Network (CN): Provides switching, routing, and transit for user traffic.

- User Equipment (UE): Terminals serve as the base station's air interface equivalent. IMSI, IMEI, and other identifiers are among them.

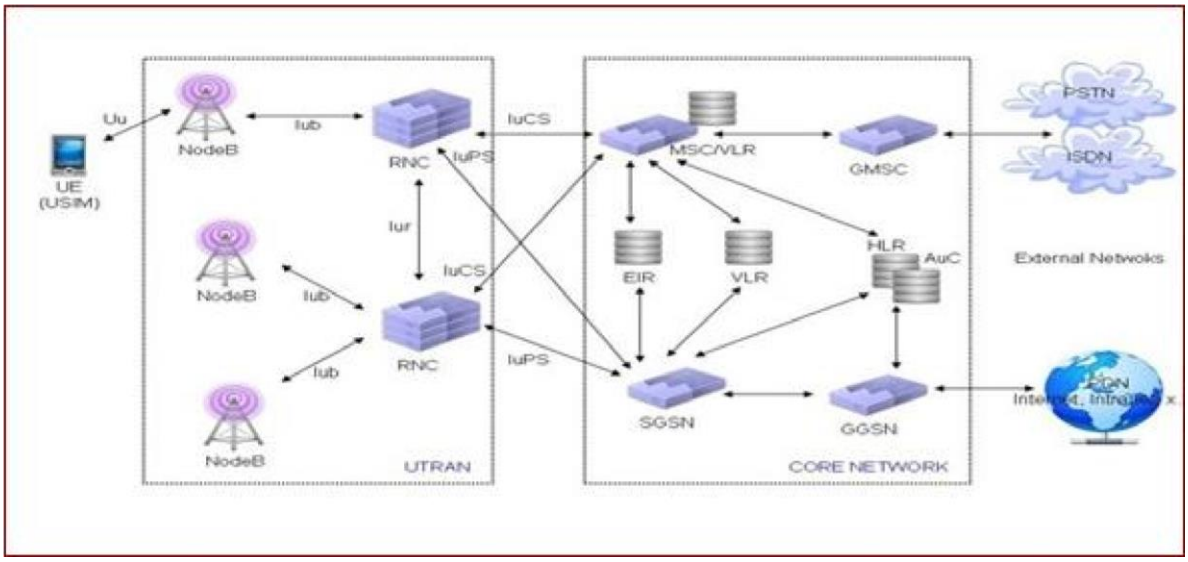


Figure 2. 4 Node B Architecture [23]

2.8.1 User Equipment

UE is a link between a user and radio air interface. It connects the user to Node B and has two parts. The first is the Mobile Equipment (ME) or phone, which is an interface for radio services. The second, on the other hand, the is UMTS Subscriber Identity Module (USIM), which is a smart card that stores essential subscriber data important to identify and authenticate the user who is given an actor of the network [23].

Technology lets a user access the network based on time. On the other hand, the UMTS system uses Wideband Code Division Multiple Access (WCDMA) technology to let users access the network based on code. WCDMA adds two access technologies; namely High-Speed Downlink Packet Access (HSDPA) and High-speed Uplink Packet Access (HSUPA), to increase the data rate to improve the Quality of Service (QoS) [19]. HSDPA supports a high-speed downlink data rate up to 10 Mbps (theoretically 14.4 Mbps) and HSUPA has a peak data rate increase up to 2 Mbps (theoretically 5.8 Mbps).

2.8.2 Radio access network (RAN)

A radio access network (RAN) is a type of network architecture that consists of radio base stations with big antennas and is widely used for mobile networks. A RAN links user equipment to a core network wirelessly. [23]

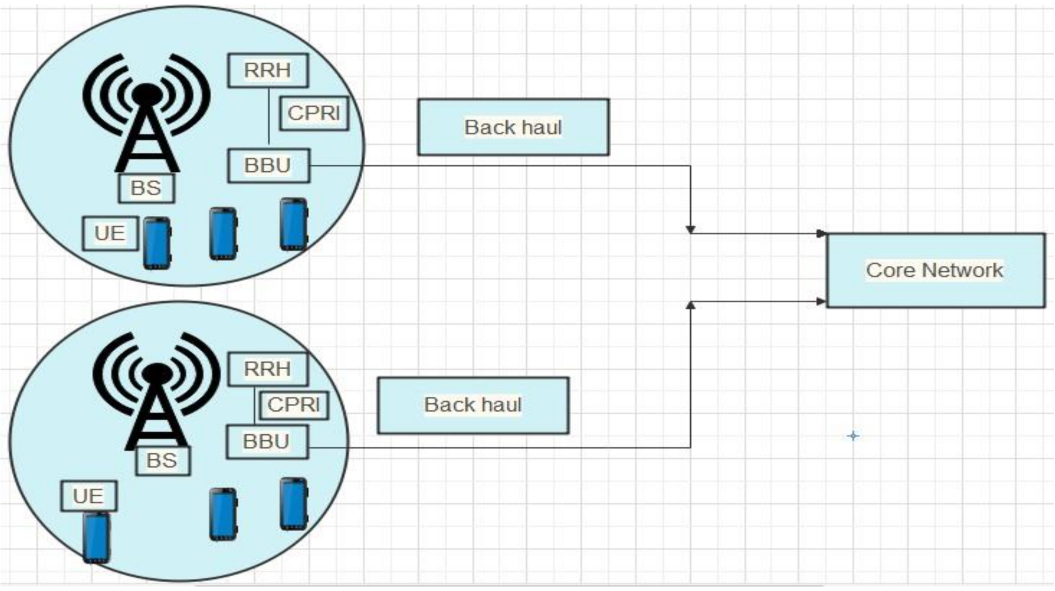


Figure 2. 5. RAN

A traditional RAN architecture where the radio unit, also called remote radio head (RRH), receives information from user equipment (UE) and sends it to the BBU via the CPRI for processing and transmission to the core network. Source: “Flying to the Clouds: The Evolution of the 5G Radio Access Networks” via Springer

2.8.3 Core Network

A mobile core network is an important component of a larger mobile network. It enables mobile users to use the services to which they are entitled.

2.8.4 Radio Network Controller

RNC is a network element that is responsible for controlling the radio access network or UMTS Terrestrial RAN (UTRAN). It controls, allocates, and checks radio resources like time, frequency and codes. RAN also controls parameters like code allocation, admission control, load control, and controls handover processes.

CHAPTER THREE

3. RELATED WORKS

3.1 Introduction

This chapter, discussed relevant research, ML, Cellular network related. Focus on RAN part and mobile network management. The network fault at locations or times from which no direct measurement data is available with the support of domain knowledge. Discussed and demonstrate how they can be solved using a real-world data set on measured mobile network assumed fault. Finally, by showing the BN and others in detail to achieve our research that is intended to be filled by the thesis work, conclude the chapter.

Neural network-based Mobile Fault Occurrence Prediction research work is conducted based on the Nonlinear Autoregressive (NAR) Neural Network time series prediction method to train the neural network actual fault occurrence time data are used of three months were used. MATLAB tool and neural network fault prediction were employed to build the models, from which accuracy of 90.71% was achieved. Recommended cell level, but didn’t take into account the No direct measurement data only on device Y. Wondie and A. Tesfay [1]

Machine learning approaches are being used to solve performance prediction in wireless networks that result in faults. When direct measurements are not available, these problems frequently require predicting network performance using current measurement data [8].

For automatic fault identification in cellular networks, Barco et al. [61] suggested a model based on discrete Bayesian networks, called smooth Bayesian networks. The major goal of this model is to improve the issue detection process by lowering the sensitivity of diagnosis accuracy, which is caused by the model's parameters being imprecise. In other words, the model was created to improve diagnosis accuracy by overcoming model parameter inaccuracy.

For fault detection, the suggested approach takes into account alerts and key performance indicators recorded daily by the network management system. In the situation of the inaccuracy of the model's parameters, experimental results on data from GSM/GPRS networks show that the suggested smooth Bayesian network outperforms standard Bayesian networks. Ruiz et al. [62] used a Bayesian Network to identify and assign probabilities to the reasons for network failures at the optical layer. The Bayesian network receives data in the form of two-time series describing bit error rates and received power as input. The data was discretized to train the Bayesian network. Then, it can predict two types of failures, namely inter-channel interference and tight filtering. The approach was found to provide high accuracy in a simulated environment.

Khanafer et al. [63] developed a Bayesian network-based fault isolation approach for Universal Mobile Telecommunications System (UMTS) networks. Given some symptoms (KPIs and alarms), the system can predict the cause. Because data about symptoms are continuous, Khanafer et al. first discretized the data (using two methods, namely percentile-based discretization, and entropy minimization discretization). This enabled the automatic detection of thresholds for symptoms that could suggest a problem. The conditional probabilities relating causes to symptoms were then learned from training data based on the thresholds using a Naive Bayes network. Experiments on data from a real UMTS network revealed that the proposed method correctly identified the cause of problems 88% of the time.

He et al. [13] proposed a Neural Network-based prediction model to handle the problem of customer turnover defect in a large Chinese telecom company with over 5.23 million consumers. The overall accuracy rate, which reached 91.1 percent, was used as the prediction accuracy standard.

To model the churn problem in telecoms that lead fault la t, Idris [14] presented a genetic programming solution with Ada Boost. The model was put to the test on two different data sets. One by Orange Telecom and the other by cell2cell, with the cell2cell dataset achieving 89 percent accuracy and the other 63 percent...

Huang et al. [15] looked into the issue of consumer complaints from a big data platform malfunction. The researchers wanted to show that, depending on the volume, diversity, and velocity of the data, big data can significantly improve the process of anticipating churn. To create the fissures in data from China's top telecoms' company's Operation Support and Business Support departments, a big data platform was required. The AUC was used to assess the Random Forest algorithm.

In recent years, a variety of technologies have been developed for more efficient radio access networks (RANs), resulting in ultra-dense heterogeneous infrastructure with high-complexity deployments. The suggested framework's main aim is to apply intelligent capabilities to the coverage planning problem in complicated multi-tier scenarios to improve network performance [16].

The author, "Detect fault occurrence and examine RC of BS outage by evaluating the TT, log file, and fault history data considering Critical Alarm, utilizing BN result demonstrates that Bayesian network performance with an accuracy of 94%." Produce RC probability for a selected BS, but note that it does not consider (cover) all aspects of the mobile network. Because the quality and faults of mobile network services are closely related to optimization, which is repaired by driving test analysis and optimization, Problems on the user's end [9]

The author notes that the cause of the fault in the mobile network system can be established based on a naive Bayesian classifier. The purpose of the paper is the radio access network (RAN) troubleshooting, which consists of detecting faults, finding the cause, and solving the issues. Use a list of faults, symptoms, and a correlate

between fault, and symptoms and use BN. The method is based on a naive Bayesian classifier, which calculates the probabilities of all the possible causes using the Bayes' rule to addresses the problem of the root cause. The naive Bayesian classifier can be studied using the theory of BNs, which allows efficient representation of a joint probability distribution over a set of random variables [10]. Parameter tuning improves the way.
For universal mobile telecommunication system (UMTS) networks, an automatic diagnosis in TS is available. The author of the paper suggests using BN techniques. Fault detection (FD) Include TS. The most complex and time-consuming one is the diagnosis of the cause of defects. A cause could be a hardware failure (like a broken baseband card in a node B). The diagnosis model represents the knowledge on how the identification of fault causes is carried out. The elements of the model are causes and symptoms. The inference method is the algorithm that identifies the cause of the problems based on the value of the symptoms. The chosen BN structure is a so-called Naive Bayes Model (NBM). The NBM consists of a parent node, whose states are the possible fault causes, and children nodes, which represent the symptoms and may have any discrete number of states. The author states specifying prior probabilities of the causes can be easily defined by diagnosis experts, taking into account their experience on the frequency of occurrence of each fault type [4] [9].

Mathematical and numerical modeling are the leading methods used for predictions of performance. A. Abu Samah et al. [14] proposed a probabilistic framework for the prediction of failure using the Bayesian Network approach, complemented by the extraction of rules for the r prediction of failure with the estimation of lead time and the index of predictability. The methodology proposed is tested, and promising results are found [23]

Statistical approaches, artificial intelligence approaches, and model-based approaches are the three primary categories of PdM approachable to monitor equipment conditions for diagnostic and prognostic reasons. Artificial intelligence approaches are increasingly used in PdM applications, as model-based approaches demand mechanical knowledge and theory of the equipment to be monitored, while statistical approaches require a mathematical background. Baptista et al. (2018), for example, compare several artificial intelligence systems to a statistical methodology (named the life utilization model) to estimate when a piece of equipment will break in the future; the results show.

Another researcher from an operational standpoint, this study analyzes the literature on the application of machine learning to fault management in cellular networks. As 5G networks evolve, summarize the main difficulties and their implications for fault management. Based on the building elements of a typical fault management system, define the essential machine learning approaches, from reinforcement learning to deep learning, and assess the progress that has been made in their implementation.[25]

Short-Time Cell Outages (STCO) is a phenomenon that affects Base Stations (BSs) in a mobile cellular operator network, according to another study. A short-term outage of all or partial BS cells (sectors) that lasts up to 30 minutes in a day is classified as an STCO, As a result, more than 98 percent of operations are guaranteed. It's a form of an outage that can't be noticed by a network monitoring system used by the operator. Although a complete characterization of STCOs has never been reported, such events are influencing the cellular network of every mobile operator. In particular, a statistical analysis of STCOs based on BSs measurements of a complete operator mobile network is performed. [22]

Due to the expansion of sensing technology, the amount of data retrieved from manufacturing processes has expanded tremendously. Data may extract useful information and knowledge from the manufacturing process, production system, and equipment when processed and examined. In industries, equipment maintenance is an important key and affects the operation time of equipment and its efficiency. Thus, equipment faults need to be identified and solved, avoiding shutdown in the production processes. Machine Learning (ML) technologies have emerged as a viable tool in Predictive Maintenance (PdM) applications for preventing breakdowns in the equipment. However, the performance of PdM applications depends on the appropriate choice of the ML method [21]

ML systems could benefit enormously from being constructed by a community of experts, resulting in more inclusive systems that allow for multiple points of view on the same knowledge. The study [59] also discusses the advantages over current approaches to building ML systems.

As indicated in the objective of this thesis, the ultimate purpose of the study is to fault collection of DT based on the user complaint and analyzed by applying machine learning and data are selected sample sites those are very critical area. Which in turn produces high customer experience and satisfaction. It will also contribute a lot to the company’s effectiveness and efficiency in its resource management.

Table 3.1 depicts fault detection and diagnosis methods. A lot of data-driven fault detection and diagnosis methods have been proposed [11]. Comparing the methods based on the four comparative criteria: - the consideration of uncertainties, the possibility of using both data and expert knowledge.

Table 3. 1 Algorithm

Method	Robust	Transparent	Mixed data	Large data	Probabilistic	Adaptive
DT	Yes	Yes	No	Yes	Yes	Yes
BN	Yes	Yes	Yes	Yes	Yes	Yes
SVM	Yes	No	No	Yes	Yes	Yes
K-NN	Yes	Yes	Yes	No	Yes	Yes

Based on the algorithms decision tree (DT), BN, support vector machine (SVM) and KNN are some of the methods. Compared to robust, transparent, mixed, and large data, probabilistic and adaptive criteria BN is selected as indicated in Table 5.2. BNs, also called belief probabilistic networks, have been proposed by many authors as the modeling technique for the development of diagnosis. This is because BNs have been used mostly to analyze data on previous fault Mixed data.

3.2 Why a BN/Naive Bayes?

To use machine learning techniques and obtain good performance, humans usually need to be involved in data collection, model, and algorithm selection. As such Naïve Bayes, Bayesian Networks provide a useful tool to visualize the probabilistic model for a domain, review all the relationships between the random variables, and reason about causal probabilities for scenarios given the available evidence. In our case, the Mobile Fault problem can't be detected by existing management especially focused on voice. Also, Bayesian networks are ideal when dealing with larger data sets, missing values, discrete variables, many variables, and where there exist dependencies between the random variables (Faltin & Kenett). All these qualities suggest that constructing a BN for the task at hand is the best option. (Explained in Chapter three), but there exists a dependency between some random variables (i.e., Interference and total Pilot Pollution). Moreover, there is a need to predict, given the observed values of multiple random variables, the probability that the total

Pilot Pollution in a specific area is greater than the value fetched from OSS's. It is also worth noting that interested in a probability value and not the most likely value. NB also considers the uncertainty and knowledge of the experts.

Bayesian Networks are a type of probabilistic expert system that uses probability as a measure of uncertainty to produce the optimum graphical structure for the data [55, 56]. Because NB makes use of all the variables in the model, it can be employed in circumstances where data is lacking [59]. By examining multiple causes, diagnostic reasoning in BN ensures that a judgment regarding the symptom and the cause is made [58]. Unlike rule-based ML approaches like NN, SVM, and BN is an inference and reasoning method. These features enable users to run queries that demonstrate cause-and-effect linkages between model variables [59]. With each new piece of information collected in BNs, the network's posterior probability values are updated. As a result, using BN in prediction issues yields more accurate outcomes. BN is superior to other ML approaches such as k-nn and NN because all relationships in the network structure are transparent. Furthermore, it can yield excellent results even when the data set is small and the number of variables is large. Because of these advantages, BN has become a popular strategy.

3.2.1 Bayesian Network

BNs are a type of probabilistic graphical model that can be used to build models from data and/or expert opinion. They can be used for a variety of activities including prediction anomaly detection, diagnostics, automated insight, reasoning, time series prediction, and decision-making in uncertain situations. They are also commonly referred to as Bayes nets, Belief networks, and sometimes-causal networks [22]. BNs are probabilistic because they are built from probability distributions and also use the laws of probability for prediction and anomaly detection , for reasoning indecision diagnostics, decision-making under uncertainty, and time series prediction. A simple BN in Figure 3.6 consists of only three nodes and two links. It represents the conditional probability distribution (CPD), Prior probability, and joint probability distribution.

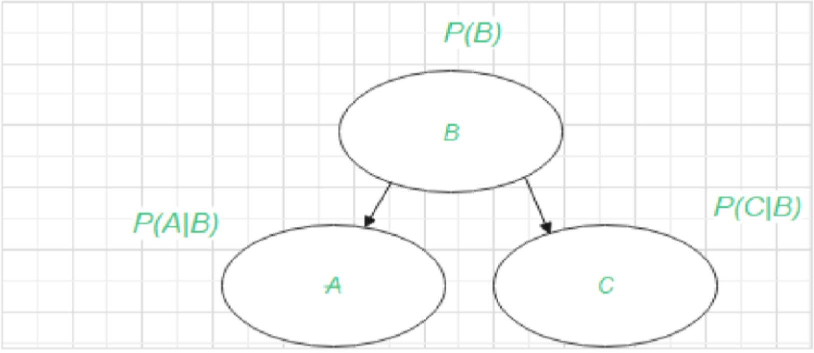


Figure 3. 1. A simple Bayesian network

CPD: is specified at each node that has parents. Example A and C have parents. CPDs of Variables A and C, are P (A | B) and P(C |B) respectively.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Prior probability: is specified at a node that has no parents (the root node). Example B has no parents the prior probability of B is P (B). The edges in the BN represent the joint probability distribution of the connected variables. For example, the joint probability distribution for the edge (B, A) is P (A, B) which represents the probability of joint event A and B. $P(A, B) = P(A|B) * P(B)$ The following features make BNs, in many cases, a suitable technique [29]: Knowledge representation: it is easy to maintain consistency and completeness in probabilistic knowledge bases. Conditional independence: can be calculated by an expert using graphical models. Flexibility: BN allows uses the same model to evaluate, predict, diagnose, and optimize decisions.

Additionally, BNs are:

▪ Graphical models that can simply and intuitively display relationships.

▪ Capable of representing cause-effect relationships.
▪ can handle uncertainty through the established theory of probability.

▪ Mathematical support: allows the analysis of the model given the knowledge of its performance and precision before implementation is carried out.

▪ Robustness: approximate answers can be obtained, even when the existing

Information is incomplete whenever new information becomes available.

▪ it is simple and effective. Types of BN connections are diverging, converging, and serial connections [25].

Diverging connection: Information can flow between the child of connections A unless A is known.

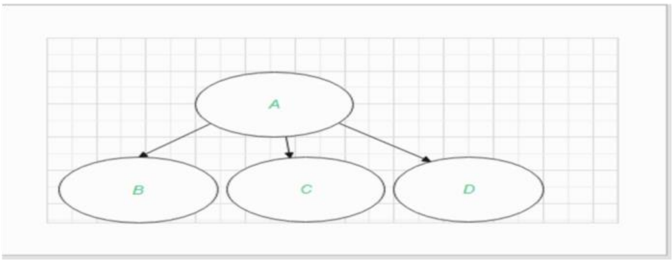


Figure 3. 2. Diverging connections

Converging connection: Information cannot flow between the parents unless A is known.

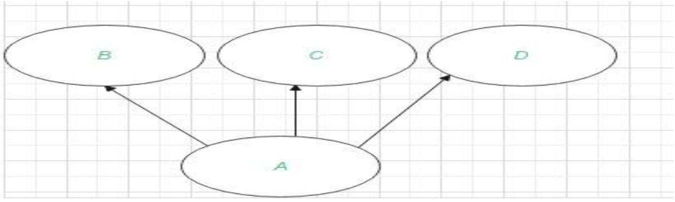


Figure 3. 3. Converging connections

Serial connection: Information can flow from A through B to C unless B is known.

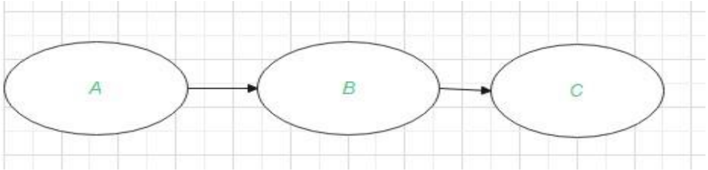


Figure 3. 4. Serial connection

3.3 What is DT

Every good RF design should be reviewed once it has been implemented. There are several methods for doing so, including KPI (Key Performance Indicator) analysis, prediction tools, and signal interference another very common and efficient way to find fault and evaluate the network is by conducting a Drive Test. [25]

The Drive is a Technology Neutral test that takes place in cellular networks (GSM, CDMA, UMTS, LTE, etc. ...). This is how data on vehicle movement is acquired. It offers a Walk Test version, which collects data by walking through regions of interest. Drive test analysis is essential for any expert in the sector of IT and Telecom, and it consists of two phases: data collecting and data analysis. Although it may find faults and problems such as dropped calls and call issues via KPI analysis, driving testing to allow for a more in-depth study in the field. Identifying areas of each coverage sector, interference, network modification evaluation, and several other parameters

3.4 The procedure to perform a test while driving.

The vehicle does not matter, you can do a driving test using a motorcycle or bicycle. What matters is the hardware and software used in the test. A notebook - or another similar device

(1), with collecting Software installed

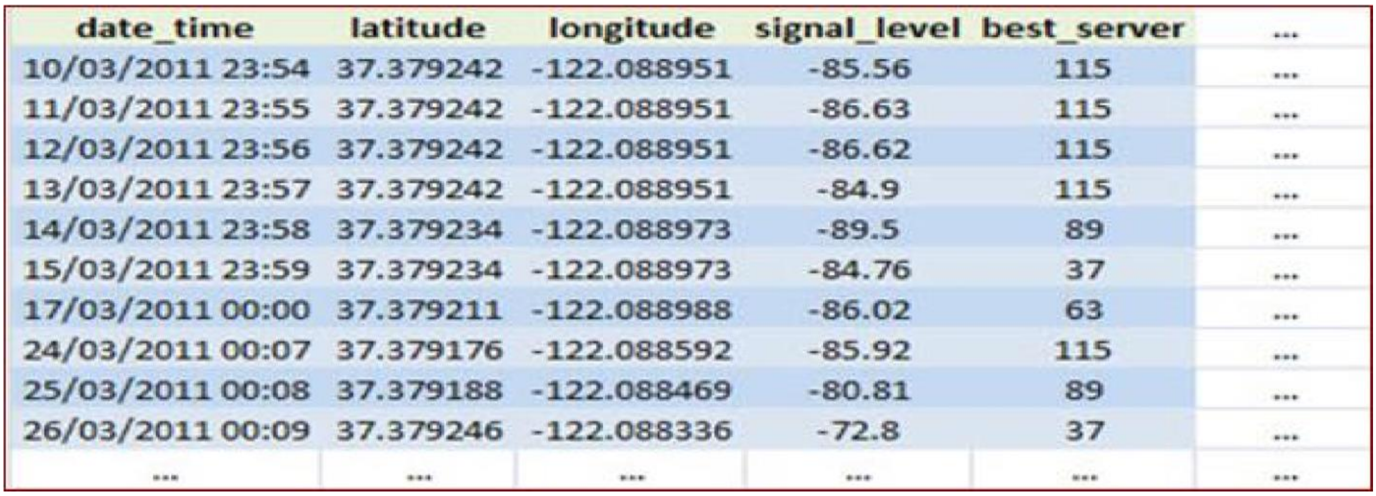
(2), Dongle - common to these types of software

(3), at least one Mobile Phone

(4), one GPS

3.4.1 Test Schedule depending on the purpose

The test can be done at any hour of the day or night. A midday Drive Test reveals the network's current state, particularly in terms of loading capacity. A nocturnal driving test allows you to test transmitters without putting the public at risk. A nightly Drive Test is frequently performed on tasks like the Design of the System. For instance, consider the integration of new websites. Daytime Drive Tests are used for Performance Analysis as well as Maintenance Faults. Important: Always verify with the responsible area, regardless of the hour, to see which sites have alarms or are out of service



date_time	latitude	longitude	signal_level	best_server	...
10/03/2011 23:54	37.379242	-122.088951	-85.56	115	...
11/03/2011 23:55	37.379242	-122.088951	-86.63	115	...
12/03/2011 23:56	37.379242	-122.088951	-86.62	115	...
13/03/2011 23:57	37.379242	-122.088951	-84.9	115	...
14/03/2011 23:58	37.379234	-122.088973	-89.5	89	...
15/03/2011 23:59	37.379234	-122.088973	-84.76	37	...
17/03/2011 00:00	37.379211	-122.088988	-86.02	63	...
24/03/2011 00:07	37.379176	-122.088592	-85.92	115	...
25/03/2011 00:08	37.379188	-122.088469	-80.81	89	...
26/03/2011 00:09	37.379246	-122.088336	-72.8	37	...
...	...	...	...	...	...

Table 3. 2. Sample DT collected sample data [27]

Daily basis - or at least once a week. Stop the ongoing calls once the collection is complete, and then stop collecting. Otherwise, these cries could be misinterpreted as falls. Discussed the data once it has been obtained. Also, depending on the equipment utilized and the objective of the Drive Test, it may differ. And the purpose of the Drive Test. In the case of mobiles, there are collected all the messages exchanged between the sites and, with all layers of information - Typical example of data output is shown below Equipment and Collection Software. [27]

3.5 Drive test Case Study

Automating the optimization and management of wireless mobile networks has the potential to significantly reduce the operational costs for network operators, as well as to improve the quality of user experience [8]. Currently, much of network management involves human interaction, ranging from conducting drive tests to evaluate the network coverage and performance to diagnosing and customer complaints. That machine learning techniques can assist in these tasks by reducing the need for drive tests, and helping to predict and diagnose network fault even before they noticeably degrade the quality of service of the network users.

This thesis focus on fault problems related to the mobile network fault UE side. Generally, involve predicting the network fault at locations or times from which no direct measurement data is available.

3.6 DT Analysis

Many reasons may lead to the call issue problem, and call problem is an expression of the deep network problems. This section focuses on the assumption of the call problem reasons, commonly-used call analysis methods assumed as a fault.

3.6.1 Call problem Caused by Poor Coverage

The prerequisites of effective coverage for a sampling point in the definition of the network coverage area that RSCP and Ec/Io should be better than the stated threshold. Bad coverage is represented in this area by a low RSCP score. It's worth noting that coverage at cell boundaries is an exception. Coverage at cell borders would have a low RSCP value but a high Ec/Io, However, the coverage in these cell boundaries is still considered poor. In a UMTS network, maintenance of various services would have varying coverage needs.

The power of the dedicated channels for the UL and DL before the call problem, which can be conducted using the following methods, can be used to estimate the coverage condition at the network's UL and DL.

If the UL TX power before the call problem has reached the maximum value and the UL bad, or it is found out through the single user tracing record at the RNC that the Node B has reported failure, then the call problem is caused by bad UL coverage. If the DL TX power before the call has reached the maximum value and the DL is bad, then the call problem is caused by bad DL coverage.

3.6.2 Call problem Caused by Neighbor Cells

3.6.2.1 Missed neighbor cell

The optimization of neighbor cells is a key part of radio network optimization. If specific cells are included in the neighbor cell list of one cell but not in the neighbor cell list of another cell, call problems may occur, network interference will grow, and system capacity will be compromised. As a result, neighbor cell optimization is a critical component of engineering optimization. It is possible to estimate whether the cell is configured with any neighbor cell, and you can playback the call collected data, perform Act ix analysis, and analyze the scanner data.

The UE would get the neighbor cell list from Node B, and the scanner would scan the 512 PSCs and record the Ec/Io. The missed neighbor cell problem occurs when one of the PSCS is not included in the neighbor cell list, its pilot strength is more than the threshold, and the occurrence lasts for a few seconds.

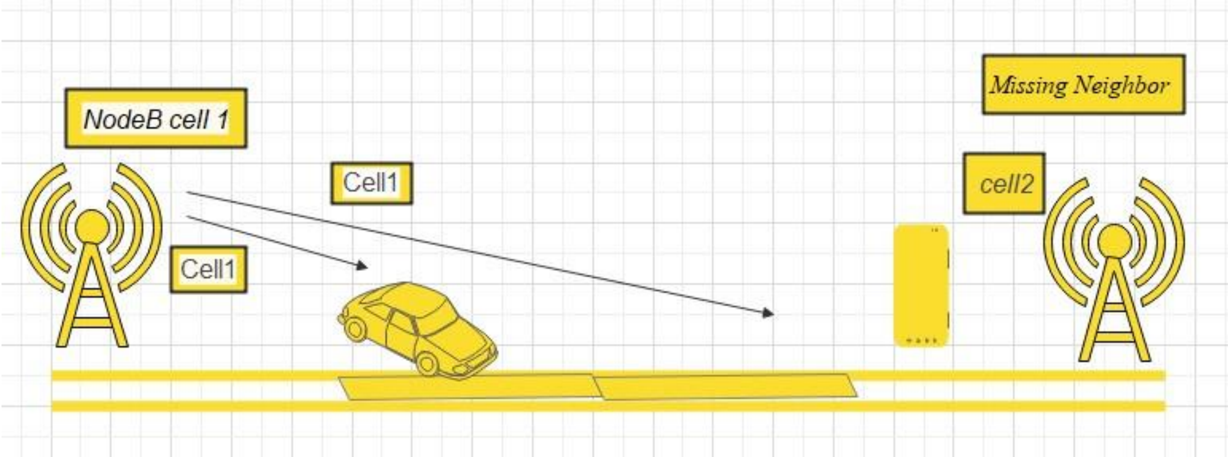


Figure 3. 5 .Missing Neighbor

3.6.3 Call problem Caused by Interference

Distinguish the UL and DL interferences. The interferences from the UL and DL would both lead to call problems. Problems when the active set's RSCP exceeds -85dBm and the comprehensive Ec/Io falls below 13dB, a call difficulty arises, and the call problem, such as drop, is caused by DL interference. The RSCP of the serving cell may be good, but the Ec/Io is bad when the handover is not prompt. The monitored set's RSCP and Ec/Io, on the other hand, are both outstanding in this situation. Pilot pollution and a missed neighbor cell are two possible causes of DL interferences with the same scrambling code. The UE would frequently reselect the neighbor cell or perform the handover in the pilot pollution area, and the incoming and outgoing calls could hardly reach the UE because the RSCP of these cells is good, but Ec/Io is bad, and the UE would frequently reselect the neighbor cell or perform the handover, and the incoming and outgoing calls could hardly reach the UE because the RSCP of these cells is good, but Ec/Io is bad. Generally, factors would lead to pilot pollution in the network.

Overshooting of sites at a high location z Node B in ring-shaped distribution z Wave-guide effect, large reflectors, and some other effects that may cause the distortion of signals that leads to calling problem.

The typical feature of the DL call problem is that the RNC sends the Active Set Update message, while the UE cannot receive it.

When DL interference exists, the UL TX power is very small. If UL interferences exist, the same problem would insist. Thus, in actual analysis, this method can be used to distinguish whether interference exists.

3.6.4 Call Failure Caused by Two Cells Using the Same PSC

3.6.4.1 Scenario one

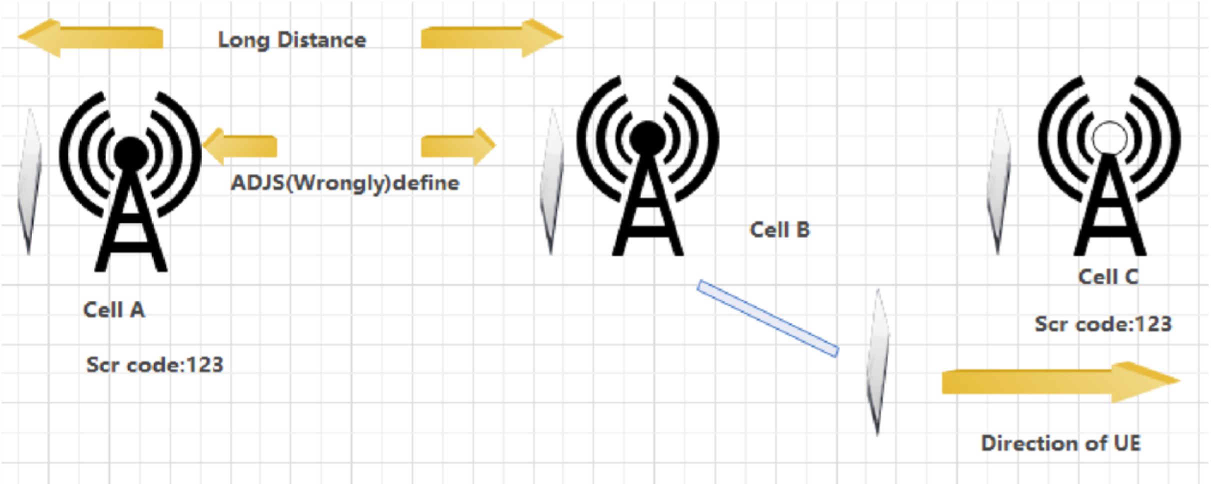


Figure 3. 6. Scenario one that may cause the same PSC problem

Cell A (source cell) and Cell B (neighbor cell) are set up as neighbors, although the geographical distance between them is enormous. Cell C and Cell B (source cell) have the same PSC, but Cell C and Cell B are not designated as neighbor cells for each other.

When the UE detects signals from Cell C, it sends an Event 1A request to Cell C to be soft-handed over. In the Event 1A request, the PSC is 123. After receiving the Event 1A request, the RNC searches Cell B's (source cell) neighbor cell list for cells having a PSC of 123, then locates Cell A. The RNC then attempts to establish a radio link with Cell A. The RNC then attempts to establish a radio link with Cell A. Cell A is added to the UE's active set by the RNC. Then, if the cell measured by the UE differs from the cell where the radio link is formed, the update of the active set times out.

3.6.5 Call problem Caused by Engineering Causes

3.6.5.1 Reversely-connected antenna

The antenna would only generate power when UEs try to access the network, and the measured value of the power equals the demodulation power. Check the ratio of two antennas, if the power of one antenna is lower than the other one in a long period, then the diversity must be reversely connected.

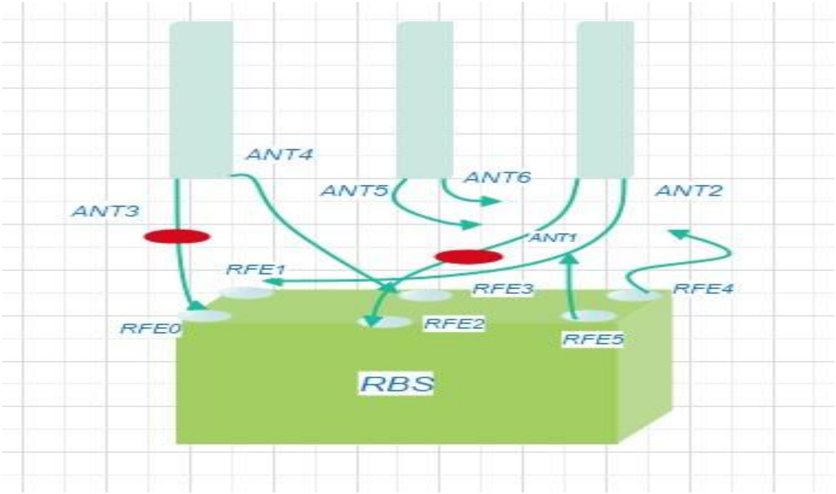


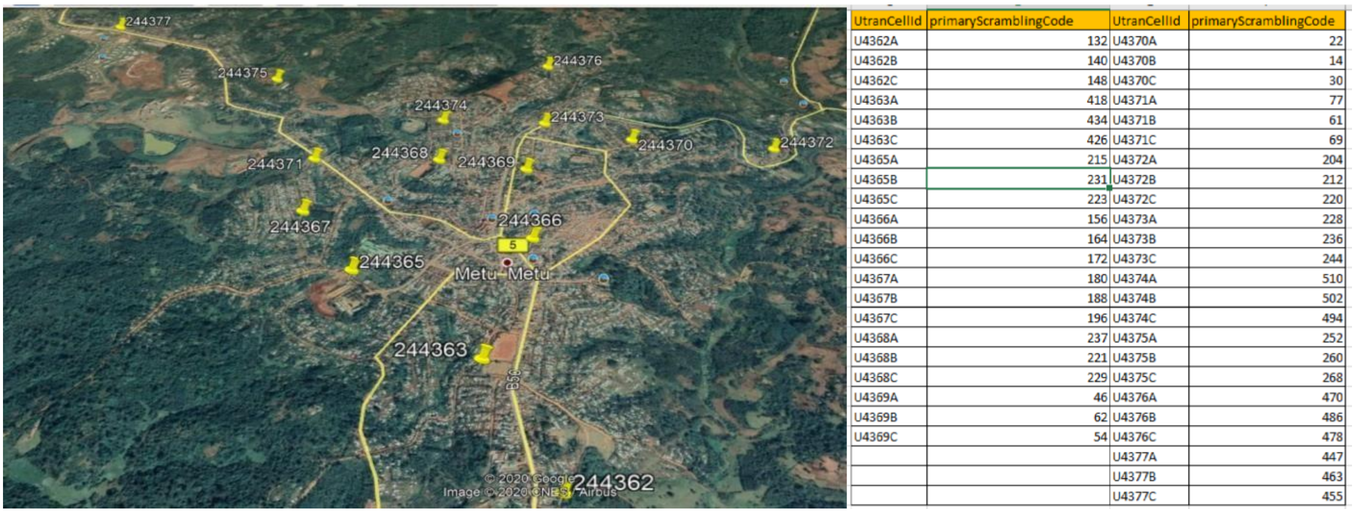
Figure 3. 7. Crossed Antenna Connection

3.6.5.2 Multi-band antenna problem

Multi-band antennas can be found in the networks of various cities. For fear of disrupting subscribers on the existing 2G network, the operator frequently refuses to change the specifications of the multi-band antenna. Then there's the possibility of pilot contamination or overshooting. The operator to alter the antenna so that the 2G and 3G networks can have separate antennas to fix this problem. If these antennas cannot be modified, the individual environment must be thoroughly examined before any action is taken. To avoid the problem, you can optimize the neighboring cells [38].

Consider the following Sample

Node B

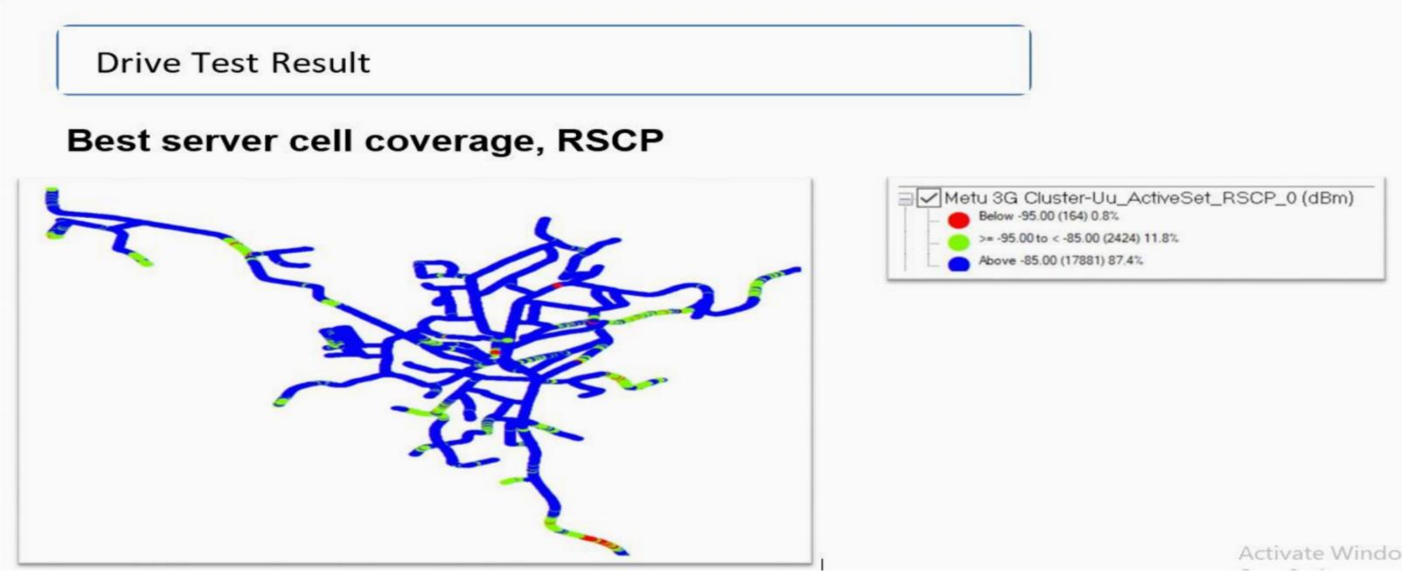


UtranCellId	primaryScramblingCode	UtranCellId	primaryScramblingCode
U4362A	132	U4370A	22
U4362B	140	U4370B	14
U4362C	148	U4370C	30
U4363A	418	U4371A	77
U4363B	434	U4371B	61
U4363C	426	U4371C	69
U4365A	215	U4372A	204
U4365B	231	U4372B	212
U4365C	223	U4372C	220
U4366A	156	U4373A	228
U4366B	164	U4373B	236
U4366C	172	U4373C	244
U4367A	180	U4374A	510
U4367B	188	U4374B	502
U4367C	196	U4374C	494
U4368A	237	U4375A	252
U4368B	221	U4375B	260
U4368C	229	U4375C	268
U4369A	46	U4376A	470
U4369B	62	U4376B	486
U4369C	54	U4376C	478
		U4377A	447
		U4377B	463
		U4377C	455

Figure 3. 8 . Node B [Google Map]

From Major cities, DT collected data to sample problematic areas with

reason.



Drive Test Result

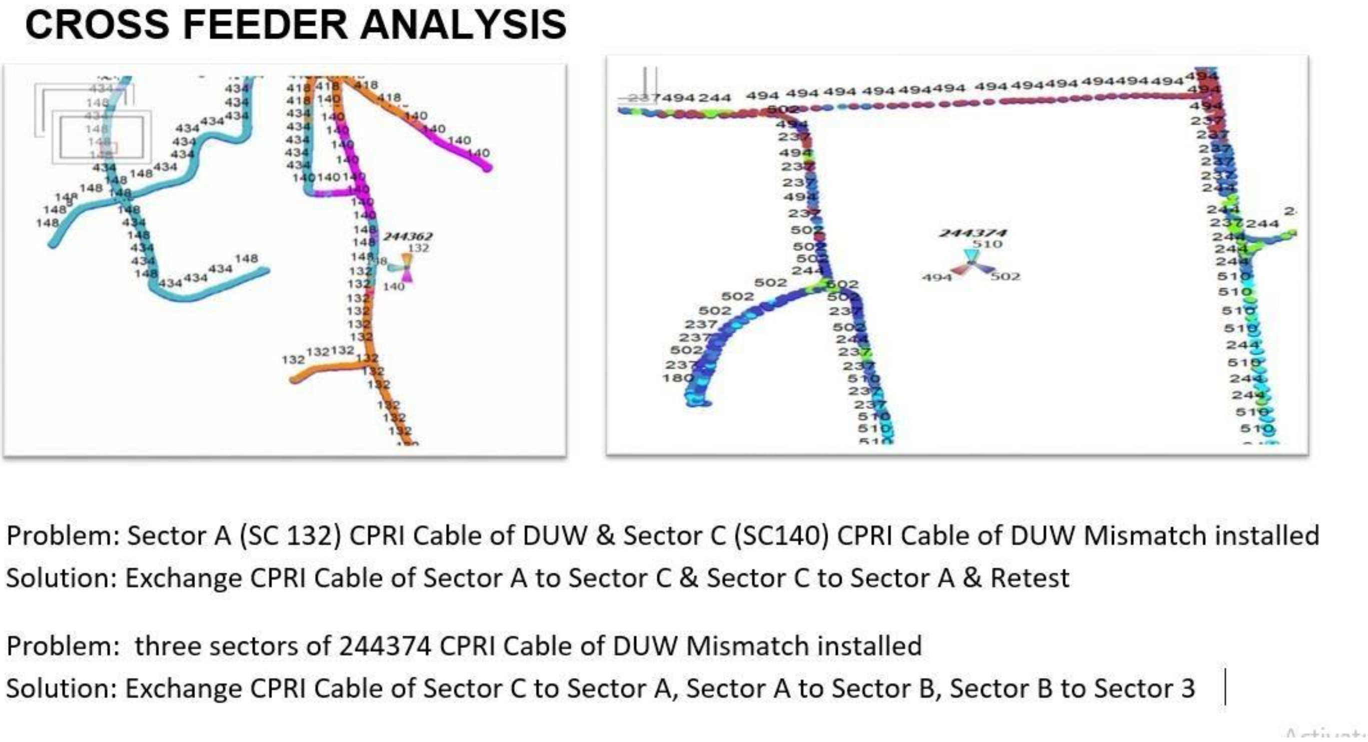
Best server cell coverage, RSCP

☑ Metu 3G Cluster-Uu_ActiveSet_RSCP_0 (dBm)
● Below -95.00 (164) 0.8%
● >= -95.00 to < -85.00 (2424) 11.8%
● Above -85.00 (17881) 87.4%
Activate Windo

Figure 3. 9. RSCP [Act ix Analyzer]

Received Signal Code Power (RSCP) is a term used in the UMTS cellular communication system to describe the power measured by a receiver on a specific physical communication channel. It's utilized as a signal strength indicator, a handover criterion, a downlink power control criterion, and to quantify route loss. A physical channel in CDMA systems corresponds to a specific spreading code, hence the name (received signal code power). Receiver Side Call Power is another name for RSCP [54].

While RSCP can be applied to any CDMA system, it is most commonly associated with UMTS. Furthermore, while RSCP can theoretically be monitored on both the downlink and the uplink, it is only defined for the downlink and is thus assumed to be measured by the UE (User Equipment) and reported to Node B [54] in our instance, the RSCP is in good shape and is on the rise

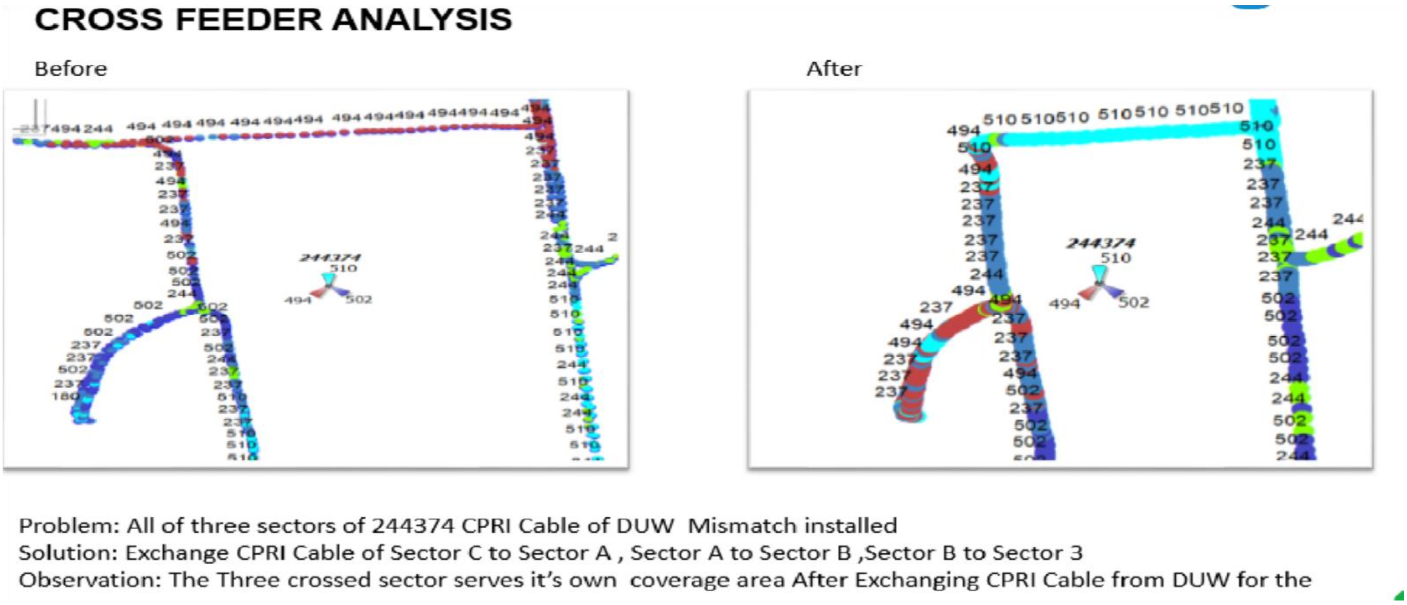


CROSS FEEDER ANALYSIS

Problem: Sector A (SC 132) CPRI Cable of DUW & Sector C (SC140) CPRI Cable of DUW Mismatch installed
Solution: Exchange CPRI Cable of Sector A to Sector C & Sector C to Sector A & Retest

Problem: three sectors of 244374 CPRI Cable of DUW Mismatch installed
Solution: Exchange CPRI Cable of Sector C to Sector A, Sector A to Sector B, Sector B to Sector 3

Figure 3. 10.Cross feeder   Analysis [Act ix software]



CROSS FEEDER ANALYSIS

Before

After

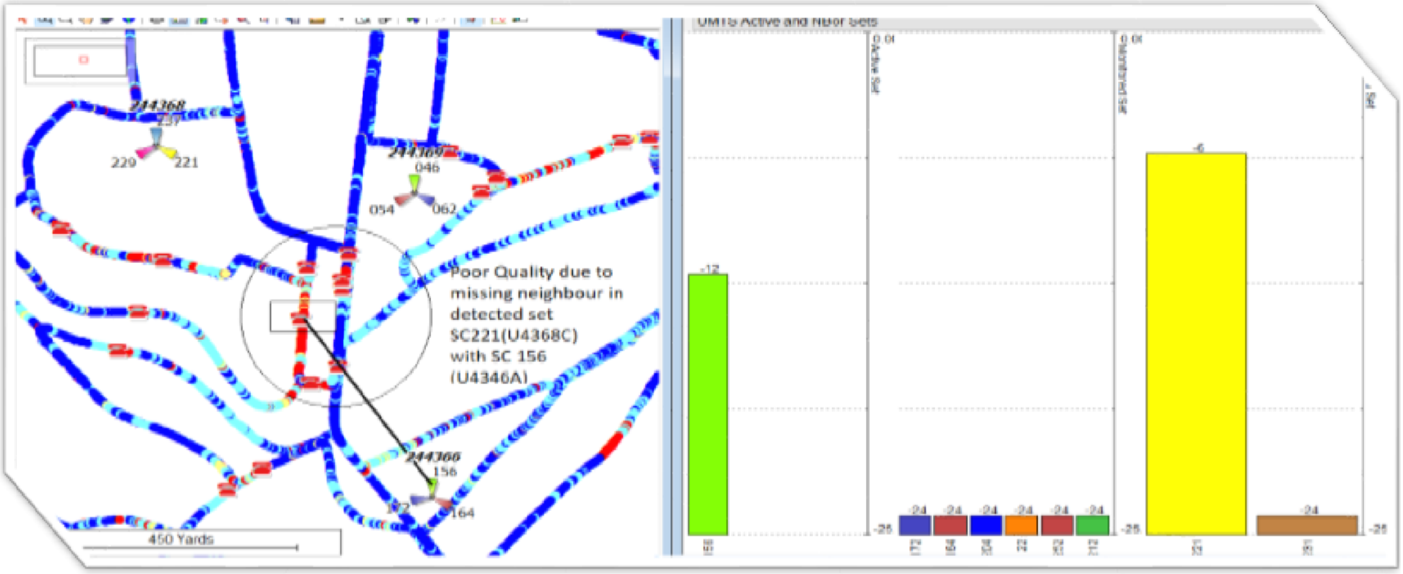
Problem: All of three sectors of 244374 CPRI Cable of DUW Mismatch installed
Solution: Exchange CPRI Cable of Sector C to Sector A , Sector A to Sector B ,Sector B to Sector 3
Observation: The Three crossed sector serves it's own coverage area After Exchanging CPRI Cable from DUW for the

Figure 3. 11. Cross feeder   Analysis [Act ix software]

When the feeders for two or more sectors in a site are linked incorrectly, this error occurs. This issue is most likely to arise during the initial setup of a site or maintenance. Consider the following scenario for a new cell site with three sectors -Sec A, B, and C:

If the feeder for sector A is connected to sector C and vice versa, the crossed feeder problem occurs. When this occurs, the sectors are usually able to continue to offer adequate coverage. The network settings for the two sectors, on the other hand, are reversed. For example, in our situation, the parameters containing the sectors' BCCH (in a GSM network), SC (UMTS), or PN (CDMA), and the researcher focus on SC.

Call problem analysis due to missing Neighbor



Poor Quality due to missing neighbour in detected set SC221(U4368C) with SC 156 (U4346A)

Problem: Poor quality(EcNo)on circled area due to missing neighbor ,EcNo in dected set SC 221(U4368B) is not defined with Serving SC 156(U4366A)

Solution: ADD Source cell ID (U4366A) with Target cell ID (U4368B) & Retest

Figure 3. 12. Call Problem due to Missed Neighbor [Act ix Analyzer]

Network quality issues can be exacerbated by an insufficient neighbor cell plan. The impact of a missing neighbor cell on network performance is dependent on the coverage provided by other cells in the same area. A missing neighbor definition might have a major performance impact in certain situations, as per the above-shown SC221 (U4368B), especially if few other cells give coverage and the lost adjacency would have been the main HO target. These facts point to the requirement for a proactive mechanism that examines existing neighbor cells and flags any that are missing for every operating cell. A new approach for finding these missing adjacencies during the operational phase is presented in this study.

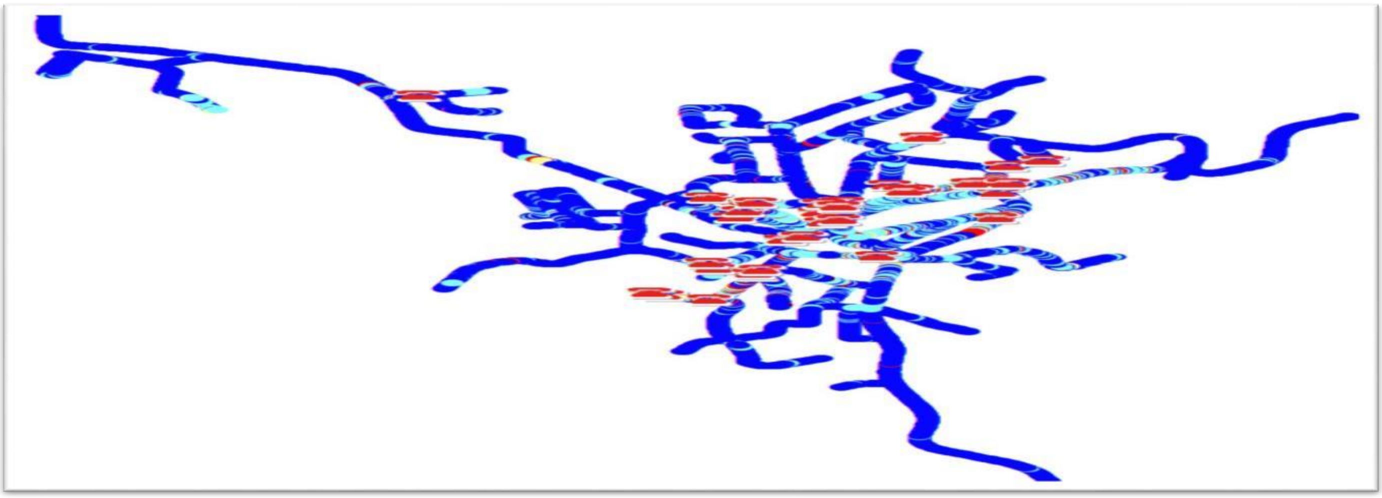


Figure 3. 13. Call Problem Area [Act ix software]

Based on the assessment of the existing issues Ethio telecom using DT, the main fault happens reasons EE, Pilot Pollution missed Neighbor and so on.

Rx Level and Signal quality- For the proper function of UMTS equipment, it is not only sufficient to provide enough field strength. One other critical parameter is the E_C/I_0 that can be compared to the signal-noise ratio in other communication systems. Experience shows, however, that, at the coverage borders, measuring a parameter comparable to the field strength (in UMTS expressed as RSCP) is sufficient to evaluate call issues that lead to fault. Quality of voice (Rx Qual):- Quality of voice at receiving level measure bit Error rate (BER).

Hand over failure- in a cellular network, handover is a very common phenomenon to degrade the network performance that leads to a fault. Handover means handover to and fro between a cell pair frequently. Occurs high signal fluctuation at the common boundary of the cell pair. Since the Ping-Pong handover increases the times of handover and thus the loading of the network, network operators must reduce this undesirable effect. However, the conventional technology does not provide a systematic and objective solution for the operators to find the cell pair suffering from the Ping-Pong effect,

Fault_ reason- identifies how much the different device communication happened in fault. It can be highly time series fault data.

Technical Failures – this is beyond anyone's control and operators generally monitor downtimes through well-equipped network operation centers, which can be due to transmission path or power outages. Some reasons for technical problems include initial configuration conditions. Broadly speaking, there are different reasons for the fault of RBS. So, this effect is one of the reasons for service degradation. Based on the assessment of existing issues in Ethio telecom mobile, the main cause of fault reasons User side is, EE, pilot pollution Ping pong and missed neighbor, sector cross, Handover issues, the neighbor list defines missing and so on.

Pilot Pollution: The pilot channel pollution represents an issue for operators that work on Node B. Basis The pollution is a type of interference, this occurs when there is no pilot dominance this means that there are several pilots with high power levels, but the mobile terminal cannot choose between them, and so these lead to call fault for UE. Generally, that usually results in a higher rate of calls issue, leads to a fault, and it's not always clear why this is happening. The operations and support systems (OSS) of network operators, for example, do not mostly identify or distinguish various sorts of faults that why our researchers focus on ML.

Table 3. 3 Summary of Related work

Study/Years	Category	Data / Evaluation	Objective	Strength	Observation
[1]2018	Neural Network	EMS, using NAR (Autoregressive Neural Network)	To predict fault using existing on EMS data	Predict the TS of dispatched from UNMS, manage the site pre-site of time	Use only DS, there is FDS because when BTS down also May Management Synchronization problem and RBS side and site Level not cell level, fault available No direct Measurements
[13]2018	Neural Network	Large number of customer 5.23 million consumers	prediction model to handle customer turnover problem	Predict the customer turnover	One input, They Didn't consider data that are not direct from Server.

[22]2018	Statistical analysis of STCOs based on BSs measurements	Existing Cell outage from Historical data of BS's	A Measurement Study of Short-time Cell Outages in Mobile Cellular Networks	Detect the outage of Cell	They didn't consider fault that not direct from Server.
[8]2020	Bayesian Network	5-month History alarm TT, Log-file from OSS	RCA of BS outage using BN by TT, log file, and model-based	Model-based BS Outage analysis	In experiments, encompasses and consider OSS causes analysis and Not consider data which is not directly available, not use all Input for BN, no Knowledge expert
[3]2018	Bayesian Network	4-month data from UMTS Historical data is available	Diagnosis the fault using Naïve Bayes	the existing data from the BS side	Not consider no direct measurement data available BS side, Input Level, one Input
[61]2017	Bayesian Network	3-month data from GSM/GPRS networks	Causes and symptoms (KPIs, alarms)	Causes and symptoms	In the experiments, the diagnostic model does not take into consideration the alarm
[62]2016	Bayesian Network	Same Time-series	Time series about received power and error rate	Used time series data	The model has good accuracy but only 2 causes considered
[28]2018	Decision Tree	Collected sample data from NNOC	Reason Call Problem fault analysis	identifying the reason	Use Traditional way, not Machine Learning

				for mobile call drops	
[27]2019	Neural network	GSM KPI Alarm	Prediction of Call Drops	Prediction of Call Drops	Not consider other KPI

The key difference between this thesis and these prior studies is that some researches deal with the analysis of wireless fault networks by considering faults like time series analysis using existing Management Server or EMS, but no one involves using existing measurement data to predict network cause of fault where direct measurements are not available. In the proposed thesis focuses on relevant issues while considering the Node B network Cell Level.

In summary, during the review of related works, most of the researches are focused on TS analysis Site Level, covering all elements of cellular network Cell Level. Hence, there is a gap to study the data generated from telecommunications, mobile networks to discover patterns that determine the fault of Node B. Leads to customer satisfaction and revenue maximization. This research also fills the gap by creating a using machine learning algorithms the reason for the fault. Therefore, the main objective is to use Machine Learning to Improved predict network fault Cell Level

CHAPTER FOUR

4. PROPOSED SOLUTION

4.1 Overview of the reason for the fault and DT analysis

For mobile operators, there is a level of challenge in identifying the reason for fault due to investment and operational competence. Most Operators stated that their company had an excellent ability to understand the reason for fault and service degradation. But unfortunately failed so proactive measure should be needed. Network difficulties, physical connection failures, and network congestion/overloads were the top reasons for network defects highlighted by mobile carriers, according to the chapter, among the faults, are Board faults in network equipment, Call problems, power outages, and transmission equipment not working. But in our, the proposed solution selected only fault related to call issue and Assumption is made based on DT analyzed as fault Node B Side due to most of us experienced in other side NE is running but User complaint high. To solve this proposed system, so selected eight classes of UE-related fault and one Specific fault. Finally, after the model is trained it detect and classify nine types of UE-related fault and one Specific fault namely Pilot pollution, Ping pong, Undefined Neighbor cell, Handover Failure, Engineering Error, Overshooting, Blockage, Same scrambling code, Interference, based on Signal Level and Signal Quality.

4.2 System Architecture

This section describes the proposed architecture for increasing proactive fault intervention before recurring faults occur. It is made up of various components, as shown in the diagram below.

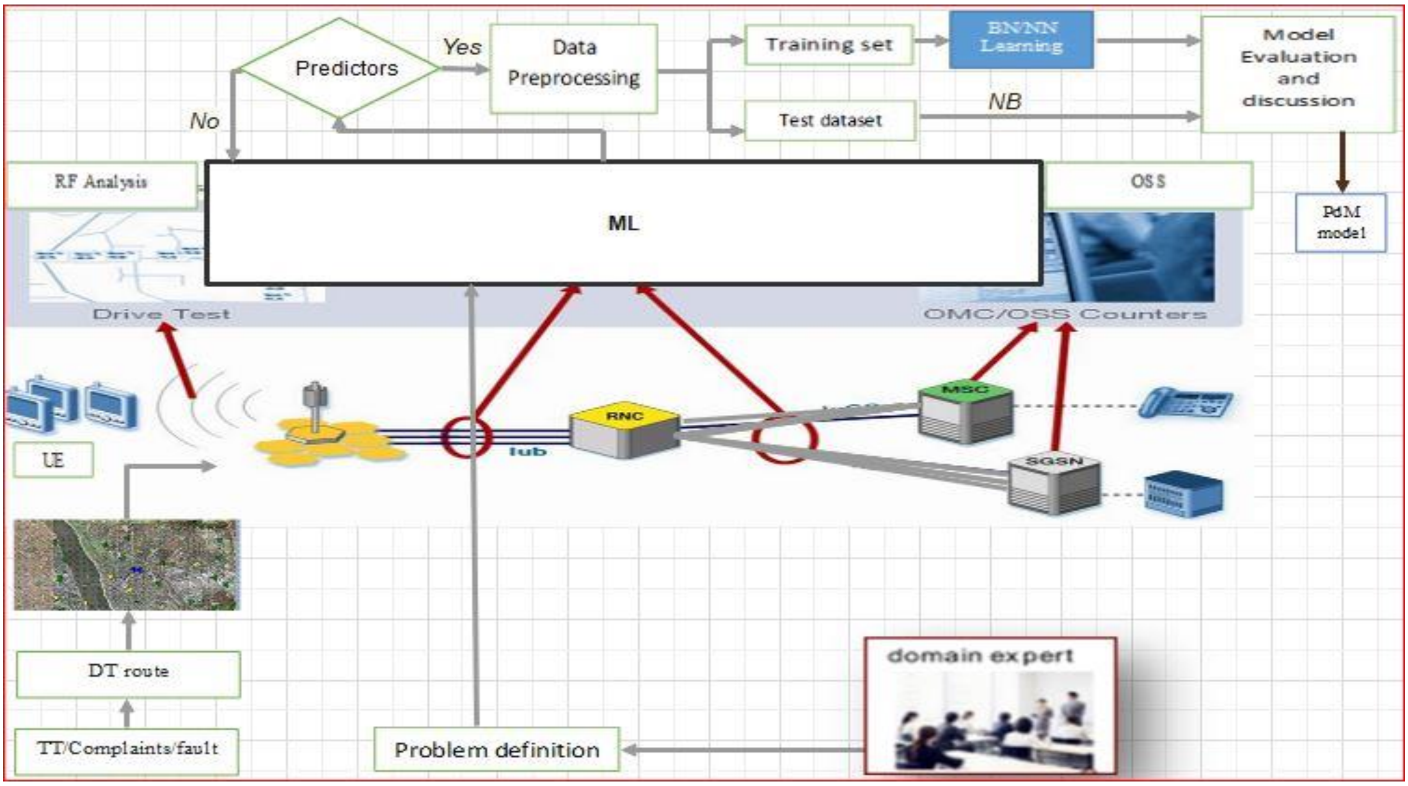


Figure 4. 1. Summary of proposed Solution

Based on the Scenario, the Domain expert analyzed the fault as per above Fig.4.1

4.2.1 Components of the proposed work

4.2.1.1 Problem Definition

Inspired by machine learning, here, define the problem, and based on the problem definition, proceed. Machine learning techniques have deeply rooted in our everyday life, humans are heavily involved in every aspect of machine learning. To make machine learning techniques easier to apply and reduce the demand for experienced human experts

4.2.1.2 Tuning Process

As shown in Figure 4.2 Once a learning issue is defined, need to find some learning tools to solve it, can target different parts of the problematic area, i.e., feature extraction. To obtain a good learning performance, try to set good decisions using our personal experience about the underneath Scenario Fig4.3. Then, based on the feedback about how the learning tools perform, adjust the Prior Knowledge wishing the performance can be improved.

In machine learning applications, domain knowledge is frequently used (sometimes without knowing that you are doing it). Feature extraction is a nice example. How to know these characteristics are crucial? Cellular coverage, for example, full in determining Rx Quality but not in detecting outages? Then there are the parameters that can choose from, such as PP, Pip classifier, and so on. How can to know a particular detect is X times worse than the other when the Level is not good due to BL and OVS, and so on that leads to a fault, when doing level classification? Some notion of prior or domain knowledge was incorporated. Bayesian models make extensive use of domain information in the form of priors for various parameters.

With previously acquired knowledge information Examine how past knowledge aids learning, as well as constructivism [65]. Updated: 09/14/2021.

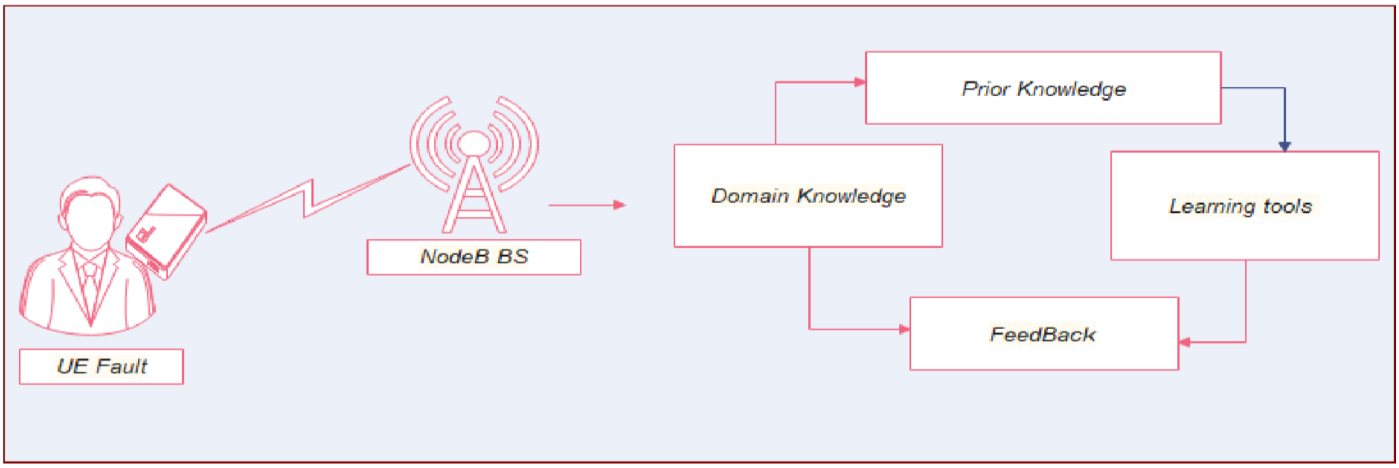


Figure 4. 2. The process of case study tuned.

4.2.1.3 UE Log.

When using KPI to analyze faults, DT is a useful tool. After integrating the UE logs, several problems, such as signaling to trace on the network side and tracing of difficult-to-locate problems, can be finally discovered.

Act ix is the most widely used DT software.

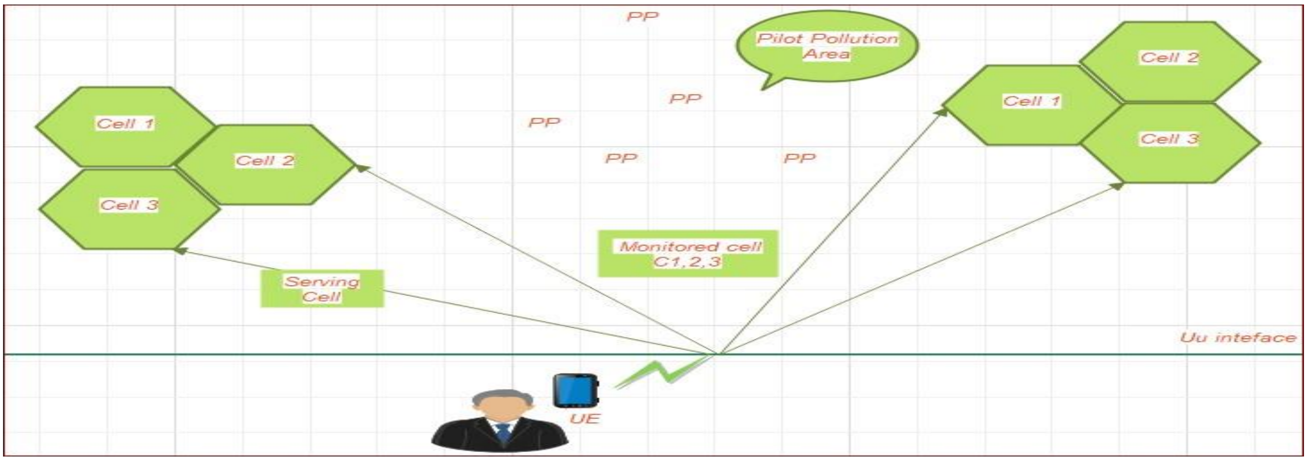


Figure 4. 3. Pilot Pollution

The above describes a method to measure the UMTS call issue. Because measuring along every potential route across the country would take an inordinate amount of time and effort, it is sufficient to measure across boundary lines dividing troublesome areas. It is not enough to give enough field strength for UMTS devices to function properly. The EC/I0, which can be related to the signal-to-noise ratio in other communication systems, is another important statistic. However, experience suggests that at coverage edges, to evaluate call issues monitoring a parameter similar to field strength (in UMTS defined as RSCP) is sufficient that result in impaired Quality of Service and are directly related to faults reports from consumers, so the researcher assumes that lead to a fault in our thesis

4.2.1.4 Data collection

Collected the desired data from DT and identified and analyzed using Act ix tools for Proposed System especially call-related Problems, Including the Specified route from that discussed 20 datasets and the most impacted area in the proposed work nine datasets as described previously.

4.2.1.5 Data Preprocessing

Redundant data from the dataset and empty records are removed. The parts of the dataset presented in different .csv files are combined into single dataset files

4.2.1.6 Model’s Training

The preprocessed data is fed into the machine learning algorithm to feature extraction is done. Which are expected to be explanatory and non-duplicate, enhance learning and generalization. Feature extraction is related to minimizing dimensions. It is suspected to contain redundant entries, minimized type of features When the input data to an algorithm is too large to be processed and analyzed, and. A smaller representation of the initial features and feature selection is the process of finding. The minimized representation of selected features incorporates critical and essential information from the input data, using this reduced representation of the data rather than the original data allowing the task to be completed. Finally, the Node B Fault prediction model is created using the dataset with extracted features.

4.2.1.7. Data Reduction

Due to noisy data instances or repeating data points, data reduction operations can lower the number of features and also delete some rows (i.e., data points). Some of the features that are strongly associated or are not very relevant may be deleted from the dataset to create models faster. This work is mostly used as a supplement to other machine learning tasks for classification assumed those are Predictors are independent each of or equal.

4.2.1.7 Classification

Are often interested in a predictive modeling problem where researchers want to predict a class label for a given observation. For example, classifying the Fault -based on measurements of the Node B. The classification module accepts input from the Naïve Bayes classifier and is responsible for classifying fault based on categorizing different types of faults are classified. Those faults are PP, Pip, Quality, OVS, Interference, SS, and HoF and are stated under Fig5.4.

CHAPTER FIVE

5. IMPLEMENTATION AND EVALUATION

5.1 Overview

In this chapter, the assumptions that the researcher considers when considering implementation using NN, Naïve Bayes- based Case Study, and the tools that use during the implementation of the system are presented. And then, the implementation and the results are discussed.

The proposed fault reason deployment approach enables Proactive action. The implementation was done in the MATLAB environment (particularly Mat lab 2021a).

5.2 Bayesian Network Model

BN are graphical models, which means that they contain a part that can be depicted as a graph. The reasons for the choice are multiple. Above all, to demonstrate and apply the ideas, and on which to evaluate the resulting algorithms, and use probability theory. Finally, construct BN structure from data and evaluate the result and which is a high frequency for fault-based on Case Study.

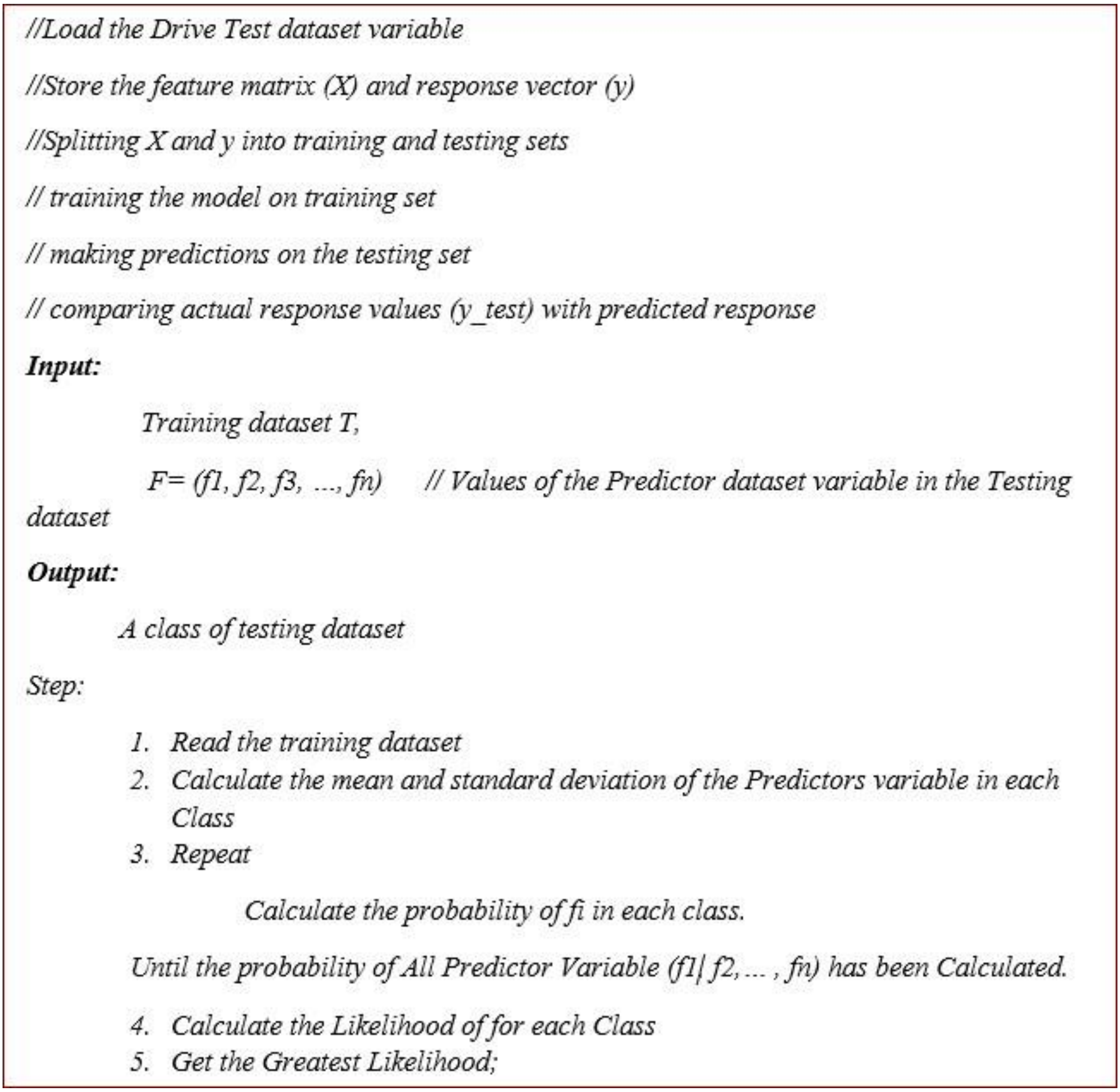
5.2.1 Construction Structure

1. The random variable should be defined as around 20. The variables such as interference, Engineering error, Missed neighbor, Ping pong, cell cross, Signal Level, Signal Quality, and so on

2. The relation between variables is defined. Analysis variables define the Predictors This consists of estimating the probabilities from the subjective judgment of an expert

3. The conditional probabilities values are stored in their corresponding tables. These conditional probabilities for each random variable should be computed.

As follow;



```
//Load the Drive Test dataset variable
//Store the feature matrix (X) and response vector (y)
//Splitting X and y into training and testing sets
// training the model on training set
// making predictions on the testing set
// comparing actual response values (y_test) with predicted response
Input:
          Training dataset T,
          F= (f1, f2, f3, ..., fn)     // Values of the Predictor dataset variable in the Testing
dataset
Output:
        A class of testing dataset
Step:
        1. Read the training dataset
        2. Calculate the mean and standard deviation of the Predictors variable in each
           Class
        3. Repeat
                   Calculate the probability of fi in each class.
        Until the probability of All Predictor Variable (f1| f2, ... , fn) has been Calculated.
        4. Calculate the Likelihood of for each Class
        5. Get the Greatest Likelihood;
```

5.3 Case study based Explanation of Bayesian Network

Bayesian belief networks or BNs are probabilistic graphical models represented as DAGs. These are applied in many fields where reasoning under uncertainty is required. The networks are composed of nodes, representing variables of interest (e.g. the occurrence of an event or a component of a system), and links joining the nodes, representing causal relations among the variables. Nodes and links constitute the qualitative part of the network, i.e. its structure, while the quantitative part is represented by the probability associated with the variables. Each node has a finite number of exhaustive and mutually exclusive states that it can assume. Every node with direct predecessors (parent) is associated with a CPT that contains the probability of each state of the node for any possible combination of the states of the parents

A graph is defined as a set V of vertices or nodes together with a set E of edges or links connecting some vertices in pairs.

Figure 5.1 shows an example of a directed graph with V = {OVs, PP, Qual, Level, Cov, Cell,} and E = {(OVs, level), (Qual, PP), (Cell, Cov),}. A path in a graph is a sequence of edges such that each of them starts with the vertex ending in the previous edge, e.g. {(PP, Qual), (Cov, Cell), (Level, cov)} Figure 5.1, where the notation (OVs, level) represents the edge connecting the vertex Inter to vertex Cov. Two vertices in a graph are connected if there exists a path between them i.e. an edge cannot connect the same vertex (e.g. (Level, level)), a directed graph may have directed cycles, that is a path starting and ending with the same vertex. A graph is called acyclic if it contains no such directed cycles. When a graph is both directed and acyclic, then it is called a DAG.

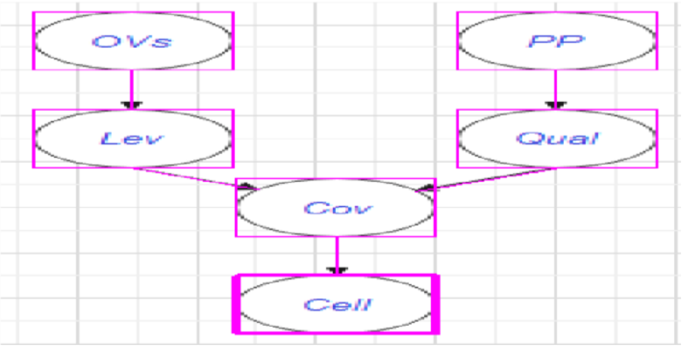


Figure 5. 1. Directed Graph

Example of BN with five variables as in Figure 5.2. The joint distribution is given by:

P (bl, OVs, level, cov, umts_cell) = p (bl), p (OVs) p (level|bl, OVs), p (cove|level), p (cell|cove). The prior probability has to be specified for OVS, bl that is the only node without parents. For all other variables, a CPT has to be provided.

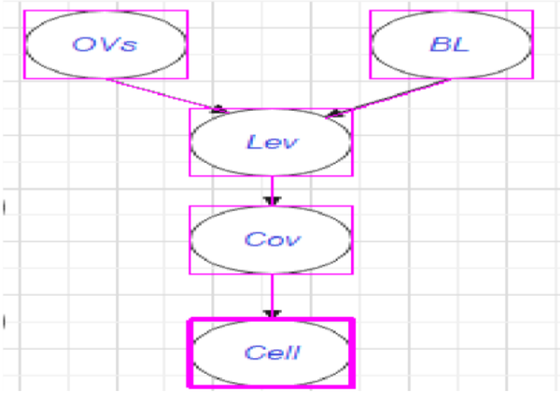


Figure 5. 2. BN

Exemplary: RAN for excessive call fault-related system

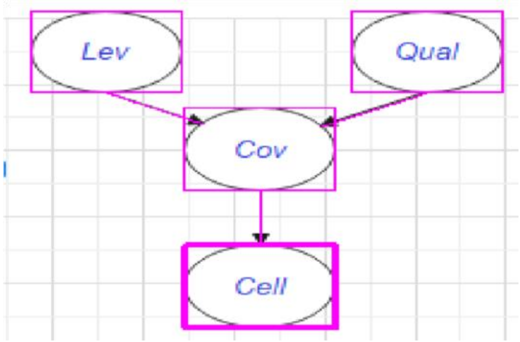


Figure 5. 3. BN for Call fault

Let us take Figure 5.3 the exemplary BN for the call fault system. The call causes by a Qual or by a Level. There is a certain probability of the limited cover a failing Qual and Level cases. When the limited coverage, UMTS is failing to serve. The variables of a BN that can model this system are Qual, Level, cover, and UMTS_Cell level.

True (t) and false (f). To specify the joint distribution of the prior probability for the variables without parents, Quality, and Signal Level, are needed. To compute the probability for the given exemplary lets us

Assume: P $(Qual = t) = 0.002$, P $(Qual = f) = 0.998$

P $(Level= t) = 0.1$, P $(level=f) = 0.9$

Moreover, the CPTs for the variables Coverage and UMTS_Cell level are shown in Table 5.4 and Table 5.5.

Table 5. 1 CPT for the node Coverage

Parent nodes		Coverage	
Quality	Level	T	F
T	T	0.9	0.1
	F	0.7	0.3
F	T	0.7	0.3
	F	0.002	0.998

Table 5. 2   CPT for the node Cell

Parent node	Cell

Coverage	T	F
T	0.9	0.1
F	0.002	0.998

The probability of UMTS_cell can be calculated using the formula

P (*UMTS_Cell)* = □$_{\text{Qual}}$□$_{\text{coverage}}$□ $_{\text{level}}$ *P* (*UMTS_cell*, *Coverage*, *Qual*, *Level*),

Figure 5.5 the cause of UMTS_cell heartbeat is due to call failure and its reason is Qual and Level.

Writing explicitly the sums for the variable Level, Qual and Coverage becomes

P (UMTS_Cell, Coverage = t, Qual = t, Level = t + P (UMTS, Coverage= t, Qual = t, Level = f) + P (UMTS_Cell, Coverage = t, Qual = f, Level = t+ P (UMTS_Cell, Coverage = t, Qual = f, Level= f) + P (UMTS_Cell, Coverage = f, Qual= t, Level= t+ P (UMTS_Cell, Coverage = f, Qual =t, Level = f) + P (UMTS_Cell, Coverage= f, Qual= f, Level = t +P (UMTS_Cell, Coverage =f, Qual =f, Level = f)

Each of these can be obtained in terms of the conditional probabilities and the prior probabilities, therefore the probability of UMTS_Cell Level becomes:

p (cell) = p (cell |coverage= t) p (coverage =t | qual = t □ level = t) p (qual = t) p (level= t) + p (cell |coverage= t) p (coverage = t | qual = t□ level= f) p (qual = t) p (level= f) + p (cell |coverage = t) p (coverage= t | qual = f □ level= t) p (qual = f) p (level = t)+ p (cell |coverage = t) p (coverage= t | qual = f □ level = f) p (qual= f) p (level= f)+ p (cell |coverage =f) p (coverage = f | qual = t □ level = t) p (qual = t) p (level= t) + p (cell |coverage = f) p (coverage= f | qual = t □ level= f) p (qual = t) p (level= f) + p (cell |coverage= f) p (coverage = f | qual = f □ level =t) p (qual = f) p (level = t) + p (cell |coverage = f) p (coverage = f | qual = f □ level= f) p (qual = f)p (level= f)

For the state true of UMTS_Cell level, from the CPTs, it follows that:

$P\ (Cell = t) = 0.9*0.9*0.002*0.1 + 0.9*0.7*0.002*0.9 + 0.9*0.7*0.998*0.1 + 0.9*0.01*0.998*0.9 + 0.001*0.1*0.001*0.1 + 0.002*0.3*0.002*0.9 + 0.002*0.3*0.998*0.1 + 0.002*0.99*0.998*0.9 = 0.0741$

When evidence is given on particular variables in the network, the posterior probability can be determined using the **Bayes rule**.

Scenario 1

ET from July to August received Complaints after swapping the network ZTE to Eriksson Network, an operator's office received subscriber complaints at fault saying it's difficult to make calls in and signals were not stable during calls.

, So based on Assumptions use as input Customer Complaints and DT results as a fault.

The following steps describe the flow using DT to test the call problem [9].

1. Call issue data

The call issue data refers to the act ix test data.

2. Call problem spots

By Using Act ix to analyze the call case to acquire the location of whether the call issue happened. Then acquire the following data: pilot data collected before and after call issue, active set and monitoring set information collected by the cell phone.

3. Stability of the primary serving cell

The stability of the primary serving cell refers to its changes. If the primary serving cell is stable, then analyze RSCP and Ec/Io. If the primary serving cell changes frequently, then the handover parameters should be changed to avoid the Ping-Pong effect.

4. RSCP and Ec/Io of the primary serving cell check the RSCP and Ec/Io of the optimal cell, and then
5. When the RSCP is bad, the coverage is poor.

- When the RSCP is normal, while the Ec/Io is bad, pilot pollution or DL interference exists.
- When RSCP and Ec/Io are both normal if cells in the active set of the UE are not the optimal cells (which can be checked through playback of data), then the call disturbed may be caused by missed

neighbor cell or untimely handovers; if cells in the active set of the UE are the optimal cells, then call issue may be caused by UL interferences.

6. Reproducing of problems with DT

Collect all necessary information by DT, then it's shall be performed to collect sufficient data. In addition, can also help whether the call issue is random or always happens at the same spot. Generally, if a call issue always happens at the same spot, this problem must be solved, or if call problems happen randomly, multiple DTs must be performed to find inner reasons the researcher gathers above 5 Month data but for ML those data are redundant reduced to 2 months.

5.4 Result and discussion

5.4.1 Result

The fault reasons are established for the BN and the interaction between variables is modeled.

This model is shown in Figure 5.4 using BN.

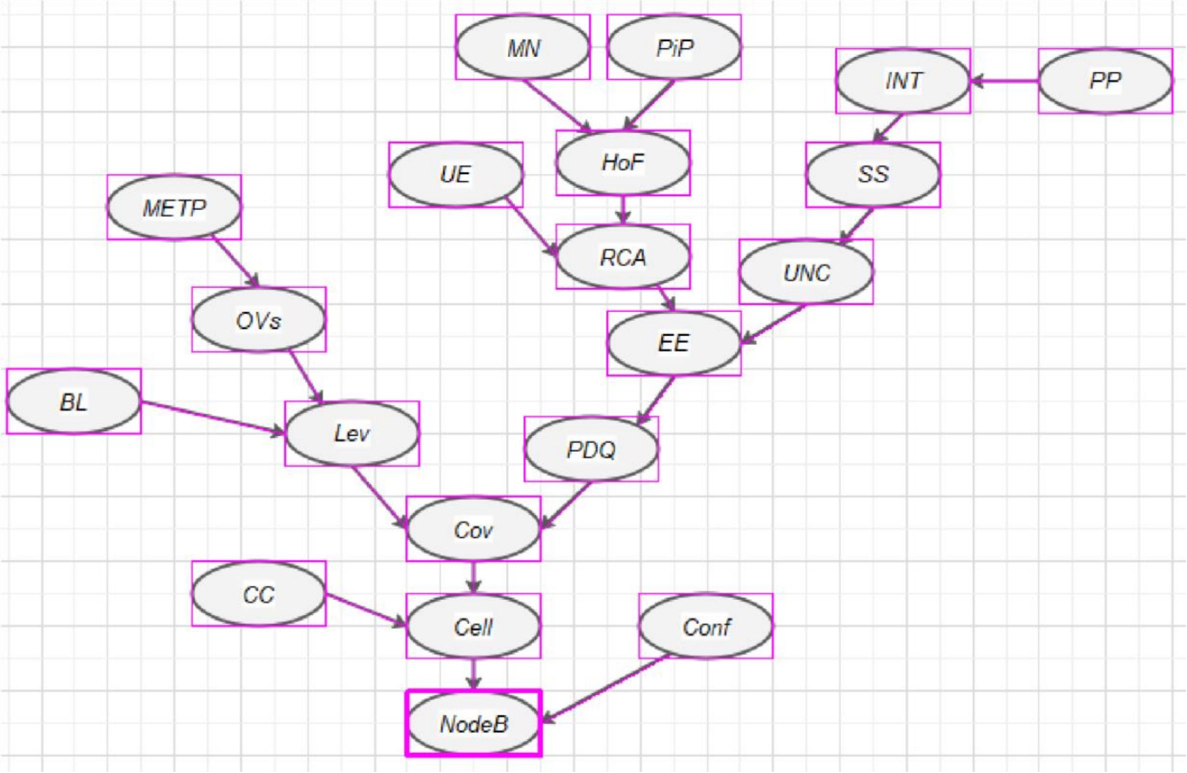


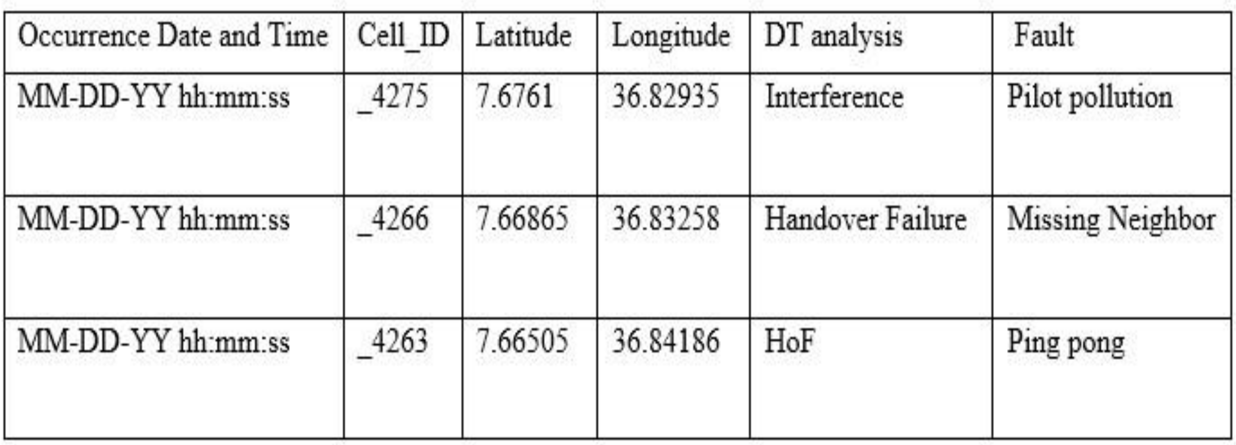
Figure 5. 4. The decision for Bayesian network relationship between case occurrences

Node B-UMTS Standard, **Cell**, **UE**- User Equipment, **Conf**-Configuration, **CC**-Cell Cross, Sec-sector, LC-Limited coverage, **HoF**- Handover failure, **EE**- Engineering Error (missing Neighbor), **SS**-the same Scrambling code, problem, **PDQ**-Poor Downlink Quality, **RCA**- Reverse connected antenna P**IP**-Ping-Pong, **MN**-Missed Neighbor, **PP**-Pilot _ Pollution, **UNC**- Undefined -Neighbor Cell, **OVS-** Overshooting

Based on customer complaints, drive test data was collected, including a log file for a highly sensitive area estimated to be about two months from July to August 2020. During the data collection, assumptions were made regarding parameters like signal level and efficiency that were not present in the OSS, resulting in a fault with low coverage area customer experience.

The proposed approach was tested. The following statistical 7800 data from the DT and log file was used to produce the results. Given an input or proof for the BNs structure, the conditional probability between the input and output nodes can be calculated, allowing for a more reliable probability inference. The conditional probability would provide useful information for determining the decision of the DT issue. The conditional probability for each component is now determined using the Bayes net tool, which is installed on MATLAB, given the evidence that the network fault occurs.

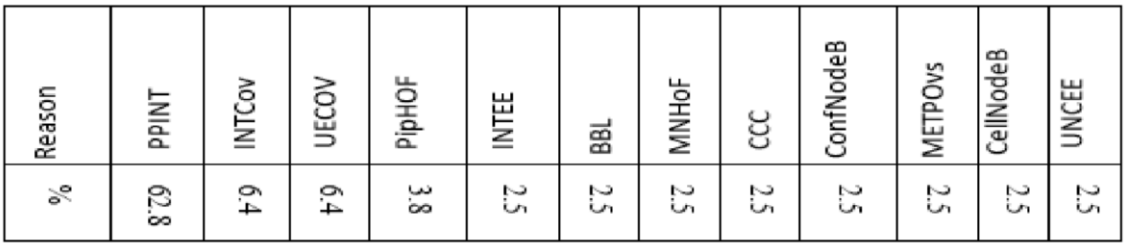
Table 5. 3 Sample data input format



Occurrence Date and Time	Cell_ID	Latitude	Longitude	DT analysis	Fault
MM-DD-YY hh:mm:ss	_4275	7.6761	36.82935	Interference	Pilot pollution
MM-DD-YY hh:mm:ss	_4266	7.66865	36.83258	Handover Failure	Missing Neighbor
MM-DD-YY hh:mm:ss	_4263	7.66505	36.84186	HoF	Ping pong

An input of 7800 data for the BN model generates an output of the Probability of an occurrence, prior probability, according to Table 5.3, sample data input format.

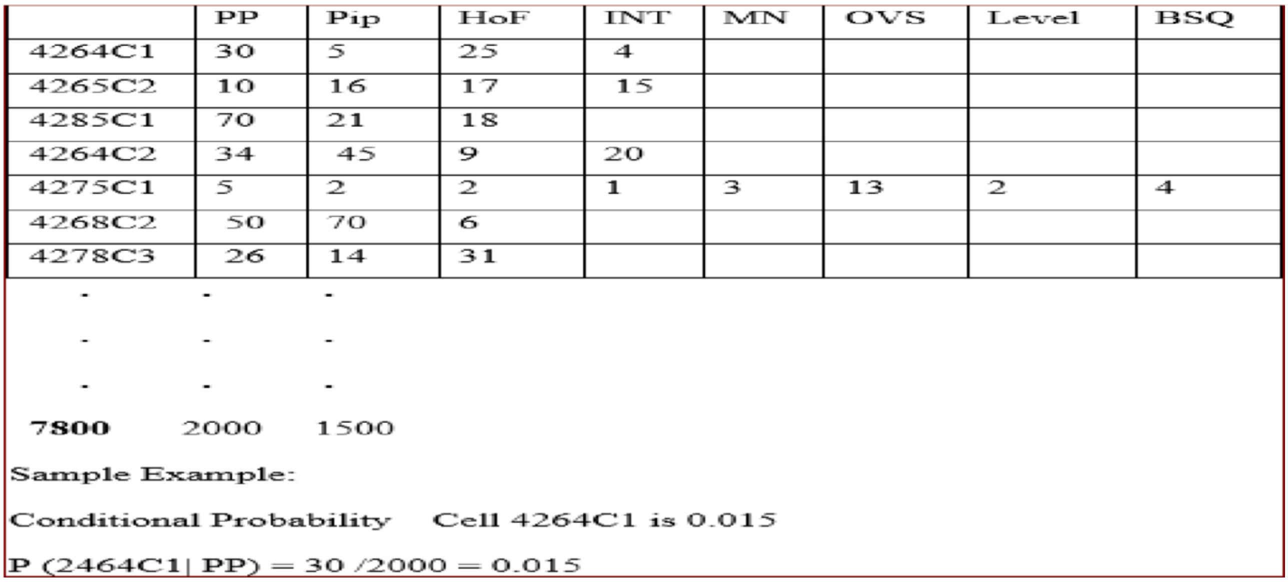
Table 5. 4. The Prior probability for the input data



Reason	PPINT	INTCov	UECOV	PipHOF	INTEE	BBL	MNHoF	CCC	ConfNodeB	METPOvs	CellNodeB	UNCEE
%	62.8	6.4	6.4	3.8	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5

Using Bayes theorem, Table 5.4 helps to find the CPD for all possible reasons. For instance, consider the UMTS network call reason of problem during initialization and after connection due to the cause of UE during mobility.

Table 5. 5 Conditional Probability Sample



	PP	Pip	HoF	INT	MN	OVS	Level	BSQ
4264C1	30	5	25	4				
4265C2	10	16	17	15				
4285C1	70	21	18					
4264C2	34	45	9	20				
4275C1	5	2	2	1	3	13	2	4
4268C2	50	70	6					
4278C3	26	14	31					
.	.	.						
.	.	.						
.	.	.						
7800	2000	1500						

Sample Example:

Conditional Probability Cell 4264C1 is 0.015

P (2464C1| PP) = 30 /2000 = 0.015

The results of the Cause UMTS call problem using the Drive test, which was taken over two months, are shown in Figure 5.5

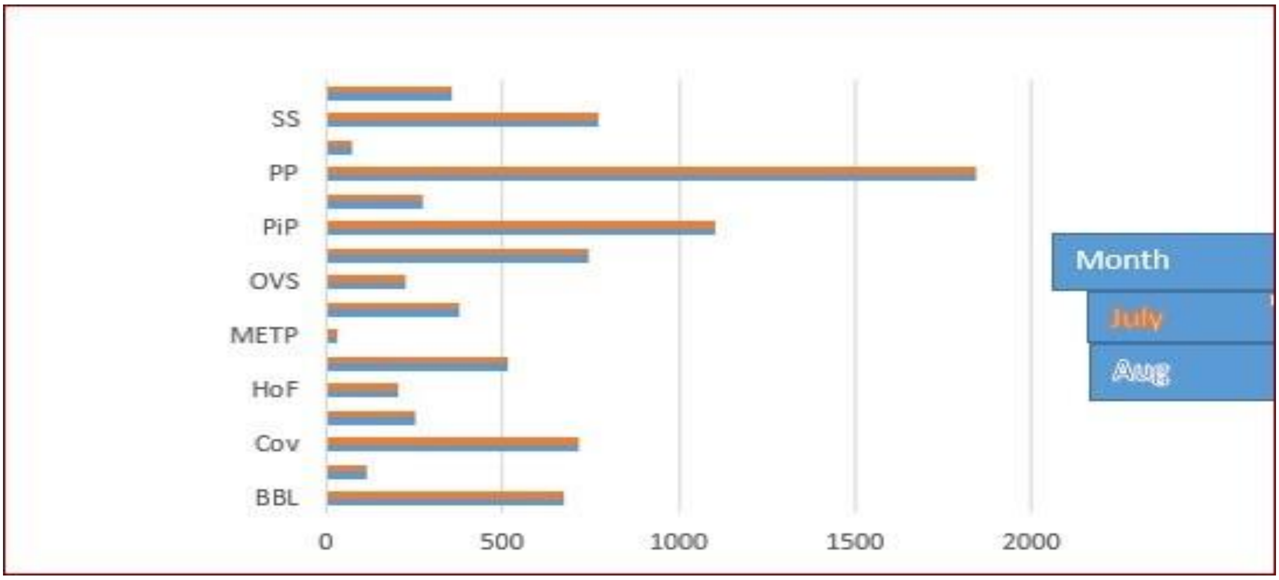


Figure 5. 5. Distribution of UMTS call issue

As per above from the collected data concerning Assumed Fault, so from above the number of PP and Pip are high during Collection of DT from the existing data distribution of UMTS call issue observed and used case

Pilot pollution and the remaining There are call fault problems collected from DT and due to of them are due to system release are due to Handover problem, interference, Missing NBR & Pilot pollution

Solution:

- Create a strong signal for pilot pollution & interference
- Add Neighbor & Retest

Generally, Excellent Coverage and Very Good Quality factory. In addition to this, Call Problem & Call setup success are not meet the scope of Cluster due to Missing Neighbor, Denied Admission due to DL

Power, interference, pilot pollution, EE & Handover problem, and so on … The key reason for the probability of a UMTS call fault issue is a downlink handover problem, according to the analysis results (pilot pollution and Ping-Pong) is high contribution.

Other factors, in this case, include a missing neighbor, and overshooting due to an engineering error, among others. When in the case of Ping-Pong, the UMTS call fault issue occurs more frequently, so to pre-engagement to use DT than before Customer Complaints.

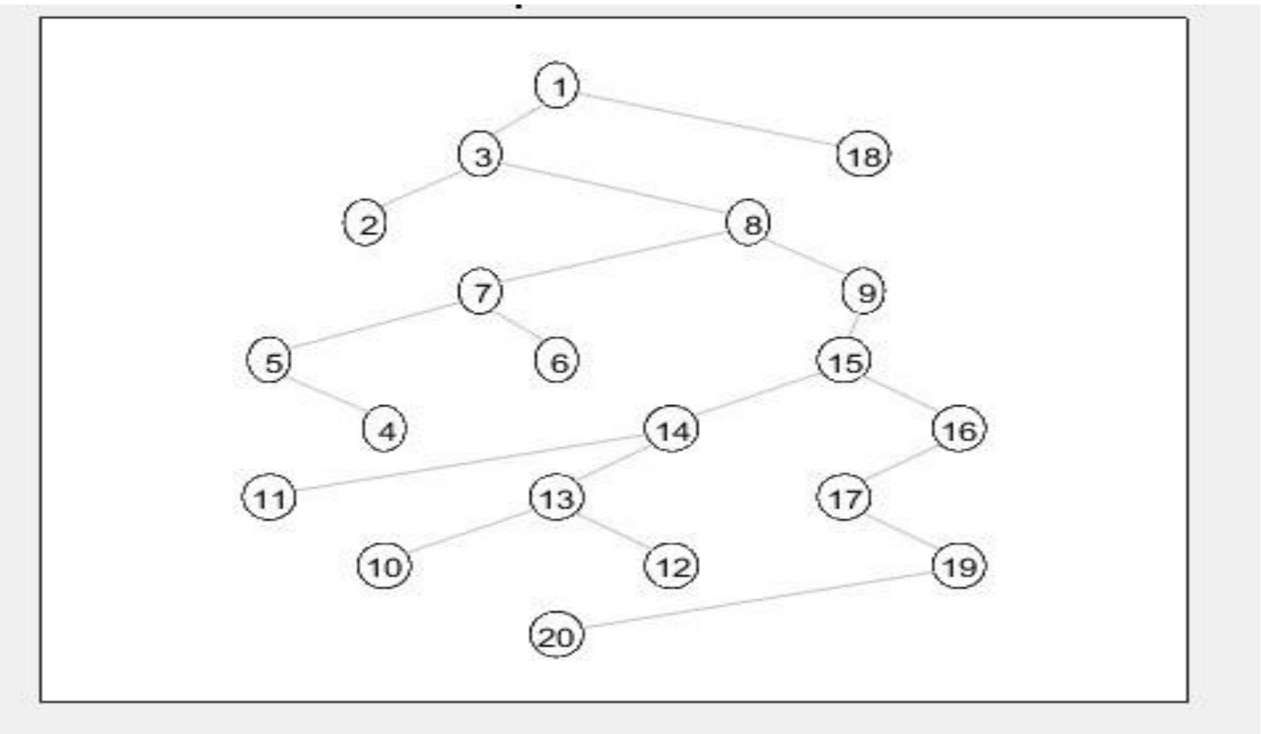
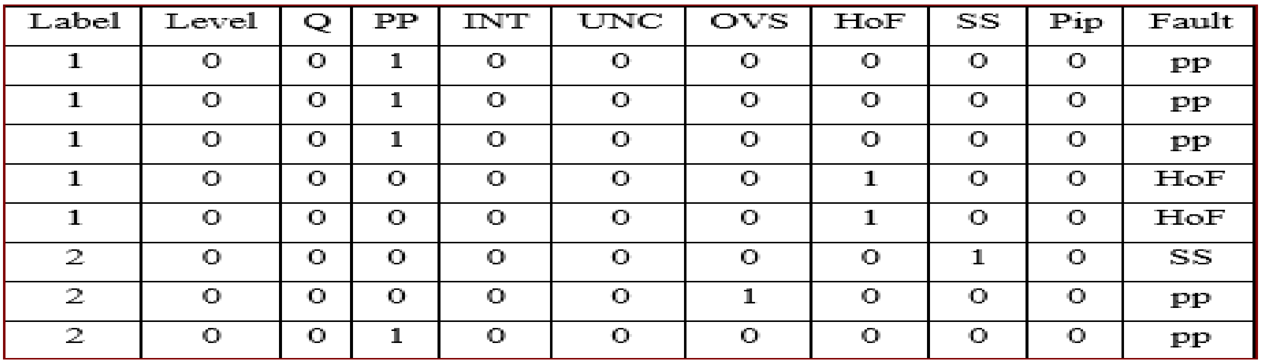


Figure 5. 6. Learned structure of the Bayesian network model using MATLAB

The model reflects the case analysis level.

Performance evaluation: BNT is a MATLAB package for guided graphical models that is open-source. Probability distributions, exact and approximate inference, parameter and structure learning, static and dynamic models are all supported by BNT. 30% sample for testing and the remaining 70% for training using the Naive classifier. There are 2340 records for research and 5460 records for training out of the total of 7800 records.

Table 5. 6 . Labelling of UE fault



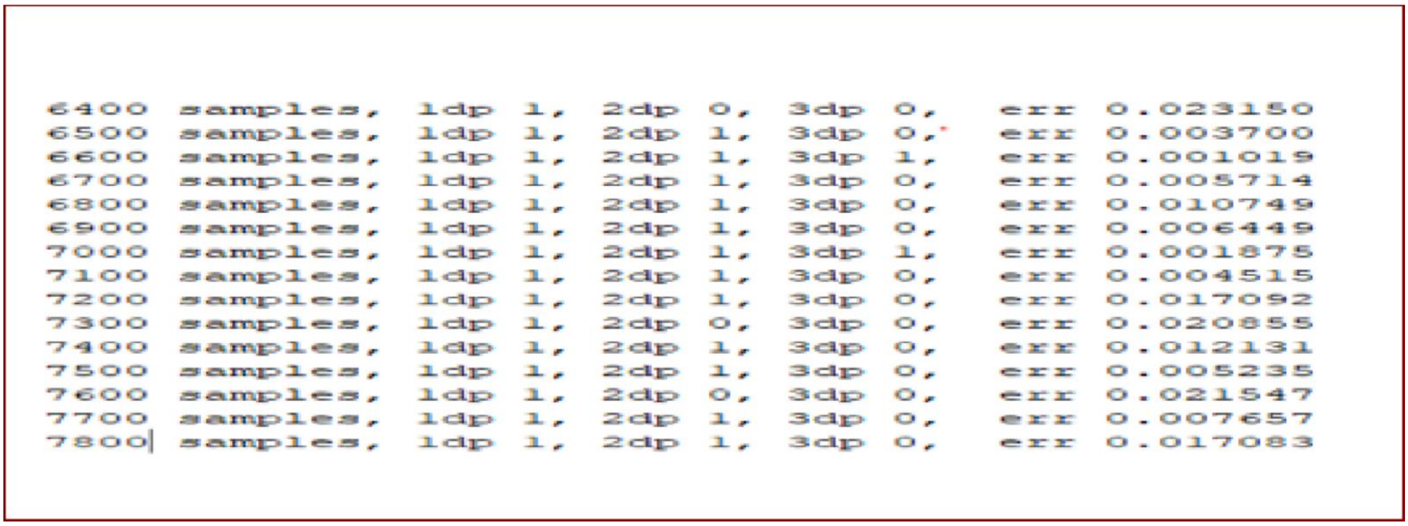
Label	Level	Q	PP	INT	UNC	OVS	HoF	SS	Pip	Fault
1	0	0	1	0	0	0	0	0	0	pp
1	0	0	1	0	0	0	0	0	0	pp
1	0	0	1	0	0	0	0	0	0	pp
1	0	0	0	0	0	0	1	0	0	HoF
1	0	0	0	0	0	0	1	0	0	HoF
2	0	0	0	0	0	0	0	1	0	SS
2	0	0	0	0	0	1	0	0	0	pp
2	0	0	1	0	0	0	0	0	0	pp

As shown in the sample data in Table 5.9, the first row indicates the possibility of the DT analyses for the UE fault happening. The last column shows the occurrence of the exact DT cause among the listed DT call-related problem. For the first event, the cause is PP; the second event is also PP, proceeds the event

Table 5. 7. The Result after Training

True Labels	Predicted Labels
‘PP’	‘PP’
‘PP’	‘PP’
‘PP’	‘PP’
‘HoF’	‘HoF’
‘HoF’	‘HoF’
‘SS’	‘SS’
‘OVS’	‘PP’
‘PP’	‘PP’

After doing DT analyses and data preparation, can calculate the probability distribution of the assumption under DT fault using naïve Bayes classifier and can determine probability. The higher probability gives priority checking. In addition, train the data and sample results taken as depicted in Table 5.7. After training, the first column indicates the true label, and the second column is the predicted label. Comparing the two columns, the true and the predicted labels are the same except for column Seven.



```
6400 samples, 1dp 1, 2dp 0, 3dp 0,  err 0.023150
6500 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.003700
6600 samples, 1dp 1, 2dp 1, 3dp 1,  err 0.001019
6700 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.005714
6800 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.010749
6900 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.006449
7000 samples, 1dp 1, 2dp 1, 3dp 1,  err 0.001875
7100 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.004515
7200 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.017092
7300 samples, 1dp 1, 2dp 0, 3dp 0,  err 0.020855
7400 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.012131
7500 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.005235
7600 samples, 1dp 1, 2dp 0, 3dp 0,  err 0.021547
7700 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.007657
7800 samples, 1dp 1, 2dp 1, 3dp 0,  err 0.017083
```

Figure 5. 7. Model Error

The Bayes net classifier algorithm generalizes by estimating the test sample, and classification error. For the duration of testing, from the total 7800 collected data, the algorithm provides a loss of 0.017083 (Loss the percentage of bad predictions. If the model's prediction is perfect, the loss is zero; otherwise, the loss is greater) hence the accuracy is 0.9821 (Accuracy is a metric for how much of the predictions the model makes are true. The higher the accuracy is, the better. However, it is not the only important metric when you estimate the performance).

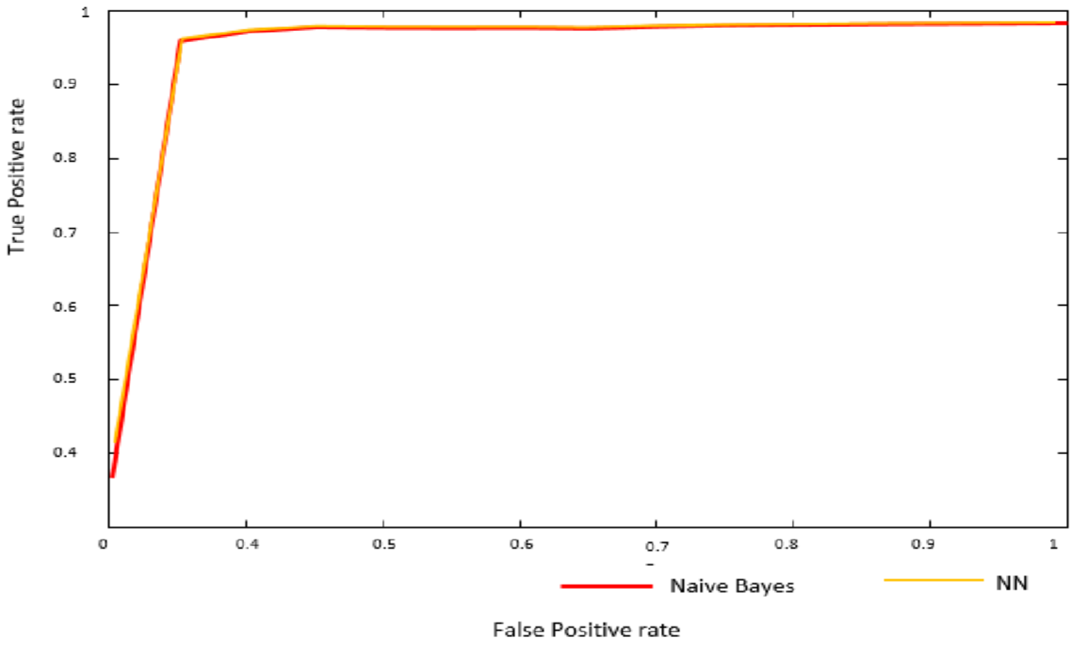


Figure 5. 8. The Area under the curve

AUCnb = 0.9821, AUCnn = 0.9801

The result shows that Naïve Bayes has slightly better sample average performance for this sample data. The area under the curve is 0.9821. Maximum AUC by 1, corresponding to the ideal classifier. The higher the AUC Value, the better the classifier's performance. Performance of the classifier the performance of a Naive Bayes classifier and its structure with data learned. The AUC (area under the ROC curve) is shown against the number of test instances selected at random.

CHAPTER SIX

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

Machine learning (ML) approaches are increasingly being used in decision-making processes. Based on differences in the level of Impact in this thesis Assumed Call-related faults, RSCP, Signal Quality, Coverage, PP, and a total number of fault causes, etc… Naïve Bayes (NB)—one of the most useful ML methods. The results compared to well-known machine learning algorithms such as Neural Network (NN). To our knowledge, this is the first performance comparison thesis of Bayesian network through NB models and machine learning models for Improved Prediction of Cellular Network Fault using these characteristics.

The purpose of this research work using Machine learning by Naïve Bayes used for Proactive measures during customer complaints and after Project Implementation is done. Fault-related cases like calls need the proactive measure. For Operator challenging to identify because no direct measurement is available to give the solution and must be done drive test and analyses before Customer complaints and its resource consuming way, so ML was used.

So to give the solution there are fault prediction, detection, and diagnosis methods. Among the methods, BN through NB is selected. BN’s is a probabilistic graphical model, when given any evidence; it can calculate the conditional probability by analyzing collected Drive Test data (DT), User Equipment (UE). The conditional probability can be calculated using the BN tool, which is installed on MATLAB.

According to the findings, when information such as the type of involvement of KPIs such as Power level, Quality, and the number of different faults is included, the algorithms utilized and the Naïve Bayes can predict the proper classes with high accuracy rates. Although other machine learning (ML) algorithms generate high-quality results. In this regard, it outperforms other machine learning techniques such as artificial neural networks (NN).

Excellent Coverage and Very Good Quality were noticed in the Case Study. In addition to this Call issue success are not meet the scope due to Missing Neighbor, interference, pilot pollution, EE & Handover issues. The key reason for the probability of a UMTS call fault issue is a downlink handover problem, according to the analysis results (pilot pollution and Ping-Pong) is high contribution. Finally, that machine learning (ML) techniques can assist in these tasks by reducing the need for drive tests (DT), and helping to predict network fault even before they noticeably degrade the quality of service of the network users.

The present research has both practical and strategic implications for future research, through the machine learning tools, and the collaboration of experts. This will allow the company to make better decisions in the management of faults data-related, especially impacted faults. At the same time, a company can be more competitive with vendors. However, the results of some research show that not all have a positive effect on the results related [27] [67]. Besides the above, there is the limitation that the sample comes from a single data observed and No direct measurement data available, which limits the generality of the results. In this context, found through the collaboration of experts and the applied machine learning methodology proposed that integrate the key factors.

6.2 Future Work

Future work should focus on other NE like GSM KPI, and other techniques, especially using Specific KPI and applying different Network Elements (NE’s). It's better to apply with unequal Assumption of other features.

Bibliography

[1] A.Tesfay and Y. Wondie, Neural Network-based 3G Mobile Sites Fault Prediction, Addis Ababa, 2018.

[2] N. Kolokas, T. Vafeiadis, D. Ioannidis and D. Tzovaras, "forecasting faults of industrial equipment using machine learning classifiers," in 2018 Innovations in Intelligent Systems and Applications (ISTA), 35 July 2018.

[3] E. T. a. M. Geremew, "“Root cause analysis of mobile site outage using Bayesian network: The case of Ethio telecom,”," November 2018

[4] G. Sati and Sonika Singh, "A review on drive test and site selection for Mobile Radio Communication," 2014.

[5] E. Š. G. B. D. H. &. J. G. Taras Maks, "Intelligent framework for radio access network design," 2020.

[6] E. Boz, Benjamin Finley, Antti Oulasvirta and Kalevi Kilkki, "Mobile QoE prediction in the field," June2019.

[7] J. R. a. P. Mähönen, "Machine Learning for Performance prediction in Mobile Cellular Network," 10 January 2018.

[8] Tesfaye.T and Beneyam.T, "Root Cause Analysis of Base Station Outage using Bayesian Network," February, 2020.

[9] A.Othman3G Optimization & Drive Test Analysis," 2018.

[10] D. Mak, “Data-based fault diagnosis model using a Bayesian causal analysis framework,” Int. J. Inf. Technology, vol. 17, no. 2, pp. 583–620, 2018.

[11] P. Panigrahi, "3G Tutorials: Introduction to 3G Foundation of Cellular Concept," [Online]. Available: https://www.3glteinfo.com/3g-tutorials-introduction-to-3g/.

[12] SDx Central, "What Is the Radio Access Network," January 17, 2018.

[13] O. P. K. a. J. I. Agbinya, "Prediction of faults in cellular networks using Bayesian network".

[14] O. Kogeda and P. Agbinya, "Prediction of Faults in Cellular Networks Using Bayesian Network Model," 23 May 2014.

[15] O. P. Kogeda and J. I. A. a. C. W., Sydney, Australia, 11 - 13 July.

[16] J. I. A. a. C. W. O.kuthe P. Kogeda, "“Impacts and Cost of faults on Services in Cellular Networks”, Proc. IEEE International Conference Mobile Business,” Sydney, Australia, 11 - 13 July 2015, pp. 551 - 555.

[17] O. P. Kogeda, "Automation of Cellular Network Faults," April 2011, DOI: 10.5772/16211.

[18] https://freedomhouse.org/print/47663, 2014.

[20] F. A. A. M. N. S. Thyago P. Carvalho systematic literature review of machine learning methods applied to predictive maintenance," https://www.sciencedirect.com/science/journal/03608352, Volume 137, November 2019, 106024.

[21] L. C. osip Josipncz, "A Measurement Study of Short-time Cell Outages in Mobile Cellular Networks," no. https://www.researchgate.net/publicataction933226, 14 June 2018.

[22] A. A. Samah, “Bayesian-based methodology for the Extraction and Validation of Time Bound Failure Signatures for online failure prediction," no. DOI:10.1016/j.ress.2017.04.016, May 2017.

[23] N. Guo, "Appling an Improved Method Based on ARIMA Model to Predict the Short-Term Electricity Consumption Transmitted by the Internet of Things (IoT)," Volume 2021 |Article ID 6610273 | https://doi.org/10.1155/2021/661027.

[24] D. MULVEY, "Cell Fault Management using Machine Learning Techniques," August 2019IEEE Access PP (99):1-1.

[25] C. Yuan and T. Lu, "“Most relevant explanation in Bayesian networks,”," vol. 42, pp. 309–352, 2011.

[26] a. M. X. B. Cai L. Huang, "Bayesian networks in fault diagnosis,”," IEEE Trans. Ind. Informatics,,, vol. 13, no. 5, pp. 2227–2240, 2017.

[27] F. Erunkulu, "Prediction of Call Drops in GSM Network using Artificial Neural Network," DOI:10.14710/jtsiskom.7.1.2019.38-46, January 2019.

[28] T. B. Getahun Semeon, ""DENTIFYING THE REASON FOR MOBILE CALL DROPS USING DATA MINING TECHNOLOGY"," DROPS USING DATA MINING TECHNOLOGY, 2018.

[29] P. F.-V. G. H. F. N. Mourad Nouioua, "A Survey of Machine Learning for Network Fault Management 10 May 2021.

[30] P. F.-V. G. H. F. N. Mourad Nouioua, "A Survey of Machine Learning for Network Fault Management," https://www.researchgate.net/publication/350576388, 10 May 2021.

[31] A. Terán-Bustamante, A. Martínez-Velasco and G. Dávila-Aragón, "Knowledge Management for Open Innovation: Bayesian Networks through Machine Learning," J. Open Innov. Technol. Mark. Complex, p. https://doi.org/10.3390/joitmc7010040. 2021, 7, 40.

[32] J. Senders, P. Staples, A. Karhade, M. Zaki, W. Gormley, M. Broekman, T. Smith and O. Arnaout, "Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World Neurosurgery," 2018, 109, 476–486. [CrossRef].

[33] K.-H. Yu, A. Beam an,d I. Kohane, "Artificial Intelligence in Healthcare. Nat. Biomed. Eng.," 2018, 2, 719–731. [CrossRef].

[34] S. Raschka, “Naive Bayes and Text Classification I - Introduction and Theory," ArXiv Preprint ArXiv:. Retrieved from http://arxiv.org/abs/1410.5329, (2014).

[35] S. D. &. G. S. Sarkar, “Empirical study on filter-based feature selection methods," International Journal of Computer Applications, 81(6), 38-43., (2013).

[36] A. G. M. A. S. &. J. M. A. Karegowda, "Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management (2), 271-277.," (2010).

[37] P. F.-V. G. H. F. N. Mourad Nouioua, "A Survey of Machine Learning for Network Fault Management," All content following this page was uploaded by Philippe Fournier Viger, 10 May 2021.

[38] A. J. a. K. A. Abdelrahim Kasem Ahmad, "Customer churn prediction in telecom using machine learning in big data platform," Ahmad et al. J Big Data (2019) 6:28.

[39] David Soldani, “Means and Methods for Collecting and Analyzing QoE Measurements in Wireless Networks," Proceedings of the 2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM'06).

[40] O. F. 1. A. 2. X. 3. L. 3. A. B. Eduardo Baena 1, "Cellular Network Radio Monitoring and Management through Virtual UE Probes: A Study Case Based on Crowded Events,” https://doi.org/10.3390/s21103404, 1 April 2021 / Revised: 26 April 2021 / Accepted: 8 May 2021 / Published: 13 May 2021.

[41] F. van Lingen, M. Yannuzzi, A. Jain, R. Irons-Mclean, O. Lluch, D. Carrera, J. Perez, A. Gutierrez, D. Montero, J. Marti and e. al., "The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge.," IEEE Commun. Mag., 2017, 55, 28–35. [CrossRef].

[42] J. Ramiro and K. Hamied, "Self-Organizing Networks (SON): Self-Planning, Self-Optimization a,nd Self-Healing for GSM, UMTS and LTE," Wiley Publishing: New York City, NY, USA, ,012.

[43] "3GPP. Telecommunication Management; Subscriber and Equipment Trace; Trace Control and Configuration Management; Version 17.1.0.; Technical Specification (ts); 3rd Generation Partnership Project (3GPP): Nice, France, 2020,"

[44] E. Alpaydin, “Introduction to machine learning. Methods in Molecular Biology (Vol.1107, pp. 105– 128). https://doi.org/10.1007/978-1-62703-748-8-7," 2014.

[45] ,. A. M.-V. A. G. D.-A. Antonia Terán-Bustamante, "Knowledge Management for Open Innovation: Bayesian Networks through Machine Learning," Journal of open innovation, 2021.

[46] R. Duda and P. J. W. a. S. 2. Hart, "DG Stork Pattern Classification;" Inc.: Hoboken, NJ, USA, 2001.

[47] M. E. P. d. l. R. B. e. l. T. d. D. 2. Rivera, Available online: http://www.urosario.edu.com/Administracion/documentos/investigacion/laboratorio/miller_2_3.pdf (accessed on 16 January 2021).

[48] G. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers.," In Proceedings of the Eleventh Conference Uncertainty Artificial Intelligence; Morgan Kaufmann: San Mateo, CA, USA, 1995; pp. 338–345..

[49] Rokach, L.; Maimon, O., “Decision Trees. In Data Mining and Knowledge Discovery Handbook; Maimon, O., Rokach, L., Eds.; Springer," US: Boston, MA, USA, 2005; pp. 165–192. ISBN 978-0-38725465-4.

[50] Wei, W.; Viswes waran, S.; Cooper, G.F., “The Application of Naive Bayes Model Averaging to Predict Alzheimer’s disease from," Genome-Wide Data. J. Am. Med. Inform. Assoc. 2011, 18, 370–375. [CrossRef] [PubMed].

[51] Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J, “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, 4th ed.; Morgan Kaufmann," Publishers: Burlington, MA, USA, 2017; ISBN 1-55860-552-5.

[52] Zhang, S.; Cheng, D.; Deng, Z.; Zong, M.; Deng, X., “A Novel KNN Algorithm with Data-Driven k Parameter Computation. Pattern Recognition." 2018, 109, 44–54. [CrossRef].

[53] A.Abu-Samah; M.KShahzadd, "Failure Prediction methodology for Improved Pro-active Maintenance Bayesian Approach," Available online at www.sciencedirect.com, IFAC-Papers Online 48-21 (2015) 844–851.

[54] "https://wiki.teltonika-networks.com/view/Mobile_Signal_Strength_Recommendations," This page was last edited on 27 July 2020, at 10:04.

[55] Spiegel halter, D.J.; Dawid, A.P.; Lauritzen, S.L.; Cowell, R.G., "Bayesian Analysis in Expert Systems. Stat. Sci. 1993, 8, 219–247.," 1993, 8, 219–247.

[56] Jensen, F.V.; Nielsen, T.D., "Bayesian Networks and Decision Graphs, 2nd ed.; Information Science and Statistics; Springer: New York," NY, USA, 2007; ISBN 978-0-387-68281-5.

[57] Pearl, J., "Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference; Revised Second Printing; Morgan Kaufmann:" San Francisco, CA, USA, 2014; ISBN 978-0-080514895.

[58] Korb, K.B.; Nicholson, A.E., "Bayesian Artificial Intelligence; CRC Press: Boca Raton, FL, USA, 2011; p. 479."

[59] Nojavan, A.F.; Qian, S.S.; Stow, C.A., "Comparative Analysis of Discretization Methods in Bayesian Networks. Environ. Model." Softw. 2017, 87, 64–71. [CrossRef].

[60] M. Roemer, C. Byington, G. Kacprzynski, and G. Vachtsevanos.

[61] C. Pinhanez, "Machine Teaching by Domain Experts: Towards More Humane, Inclusive, and Intelligent Machine Learning Systems," [v1] Mon, 19 Aug 2019 20:47:18 UTC (540 KB).

[62] Ruiz, M., Fresi, F., Vela, A.P., Meloni, G., Sambo, N., Cugini, F., Poti, L., Velasco, L., "Service triggered failure identification/localization through monitoring of multiple parameters. In: Proc. 42nd European Conference on Optical Communication," pp. 1–3. VDE (2016).

[63] Barco, R., D´ıez, L., Wille, V., L´azaro, P., "Automatic diagnosis of mobile communication networks under imprecise parameters. Expert systems with Applications," 489–500 2009.

[64] Khanafer, R.M., Solana, B., Triola, J., Barco, R., Moltsen, L., Altman, Z., Lazaro, "Automated diagnosis for umts networks using Bayesian network approach. IEEE Transactions on vehicular technology 57(4), 2451–2461 (2008)".

[65] "https://study.com/academy/lesson/prior-knowledge-definition-theory-quiz.html".

[66] Ziqiu Kang, Cagatay Catal b, "Machine learning applications in production lines: A systematic literature review," journal homepage: www.elsevier.com/locate/caie, 2020.

[67] "The effects of open innovation activity on performance of SMEs: the case of Korea," International Journal of Technology Management (IJTM), Online publication date: Sat, 06-Apr-2013.