

Real-Time Telecommunication Network Management using Data Mining and Machine Learning Techniques



By

Syeda Hajra Farhat Hashmi

MSIT-18 205855

Supervisor

Dr.Saad Qaiser

Department of Computing

In

School of Electrical Engineering and Computer Sciences(SEECS)

National University of Sciences and Technology(NUST)

Islamabad,Pakistan

(Aug 2022)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Real-Time Telecommunication Network Management using Data Mining and Machine Learning Techniques" written by SYEDA HAJRA HASHMI, (Registration No 00000205855), of SEECs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____  _____

Name of Advisor: Dr. Saad Qaisar Senior
Member IEEE

Date: 15-Jul-2022

HoD/Associate Dean: _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled "Real-Time Telecommunication Network Management using Data Mining and Machine Learning Techniques" submitted by SYEDA HAJRA HASHMI have been found satisfactory for the requirement of the degree

Advisor : Dr. Saad Qaisar Senior
Member IEEE

Signature:  _____

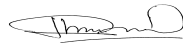
Date: 15-Jul-2022

Committee Member 1:Dr. Syed Taha Ali

Signature:  _____

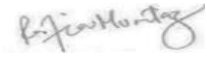
Date: 15-Jul-2022

Committee Member 2:Dr. Arsalan Ahmad

Signature:  _____

Date: 15-Jul-2022

Committee Member 3:Dr. Rafia Mumtaz

Signature:  _____

Date: 15-Jul-2022

Dedication:

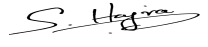
This thesis is dedicated to my family, friends, their continuous support plays a vital role in thesis completion. And also my respectful supervisor Dr. Saad Qaiser, and thesis committee members Dr. Rafia

Mumtaz, Dr. Talha, Dr. Arsalan, without their guidance this could not have come to an end.

Certificate of Originality

I hereby declare that this submission titled "Real-Time Telecommunication Network Management using Data Mining and Machine Learning Techniques" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEecs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEecs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name: SYEDA HAJRA HASHMI

Student Signature:  _____

Acknowledgement:

First of all I would like to thank Allah Almighty for completion of this important journey of my life.

Secondly, I would like to thank my department, my supervisor Dr. Saad Qaiser and respected committee members, who have guided me and supported me throughout the process.

And lastly, I would like to thank my parents, spouse for their emotional support.

Table of Contents

1. Introduction:.....	1
2. Literature Review:.....	9
3. Methodology:.....	21
3.1. Dataset:.....	21
3.2. Decision Tree Classifier:.....	30
4. Results:.....	37
4.1. 10-fold Cross Validation:.....	38
4.2. Comparison of Decision Tree Algorithm with other Algorithms:	49
5. Discussion:.....	51
6. Conclusion:.....	54
7. Future Work:.....	55

List of figures

Figure 1:A generic network fault management process[1].....	3
Figure 2: Key functions included in fault management system [4]....	5
Figure 3 Online illustrating data mining process steps[5].....	7
Figure 4Transformation from pattern to graph [8].....	10
Figure 5Comparison of K-meanand K-methodoid [9].....	11
Figure 6 Dirty data base algorithm procedure[19].....	13
Figure 7 Deep Learning methods taxonomy[22].....	16
Figure 8 Proposed Alarm Management System[23].....	17
Figure 9 Shows an MSAG present on the locality [23].....	20
Figure 10 Showing Huawei Network Management system[24].....	22
Figure 11 Visual representation of a network on NMS[25].....	23

Figure 12 Temp Client preview showing different alarms categorise by the color of alarm[26].....24

Figure 13 Python code showing ten rows of the Data set.....27

Figure 14 Python code showing description of the data set.....28

Figure 15 A generic Decision tree work flow diagram[27].....30

Figure 16 Flow diagram of decision tree working for this example.34

Figure 17 K times cross validation flow diagram[28].....31

Figure 18 Showing Accuracy trend for 10-fold data analysis.....43

Figure 19 Showing trend between accuracy and different proportions of training & testing data sets.....47

Figure 20 A comparison of different algorithms with Decision tree algorithm[29] [30].....49

Abstract:

There are currently many type of industries that requires 24/7 monitoring on different levels for smooth operations. Telecommunication is one of the industry where millions of alarms trigger on daily basis from communication equipment's and needs to be handled within time limit. These monitoring operation is normally handled by human in loop which means huge amount of time wastage. In order to minimize downtime, limit human control over this monitoring, companies have implemented data mining and machine learning techniques that helps in not only proactive monitoring of alarms along with suitable actions and also there is a huge time saved. In this paper we have experimented with some real time telecommunication alarms that are gathered from different telecommunication devices and occurred at different times. We have created a system that can predict future occurrence of an alarm on the specified machine using machine learning technologies. In this paper we have used decision tree classifier in order to classify huge number of data received from devices. We are using it to predict alarms that are to be appeared on a specific device/machine at an specific time stamp.

Keyword- telecommunication network, decision tree classifier

Introduction:

There are a huge number of alerts generate in the network operation process in telecommunication industry. Daily these alarms needs to be monitored and processed by operations team in a telecommunication company. There are multiple devices that contributes in a communication system and each of the device sends in alarms which are in millions in number. These alarms that are generating from devices are of different categories for example major, minor and alarms with other categories. One device failure may cause alarms from other devices as well. That is why the quantity of alarms per day are in millions. These telecommunication companies have different methods of dealings with the faults occuring in the network system. These methods or processes are collectively known as fault management processes that helps in catering a huge number of network faults occuring in the network. Fault management is required to ensure security of network as well as availability, reliability and optimisation if there is any. Fault management is the key element for quality service providance in a telecommunication industry. These faults that are of any category or any level will eventually contribute in degradation of the services which is the key element of the industry. Therefore, the perfect system for fault management is required in these companies. There is a separate department allocated to for detection and rectification of these faults in the network which is called network operations center

(NOC). The responsibility of this department includes detection of alarms using different softwares and then rectifying the problem been caused by creating tickets to the field team.

Fault managemnet is responsible for four main tasks that are[1]:

1, Detection of alarm: There are numerous amount of alarms coming in from different communication devices which are been monitored on netwrok mangemnet systems. Fault management system first categorize these alarms as major, minor etc. It helps in identifying the root cause alarm among huge number of alarms.

2, Diagonosis of an alarm: This task involves identifying root cause alarm from all the other alarms.

3, Isolating the root cause alarm: This process isolates the main problem causing alarms from the rest of other alarms.

4, lastly, fixation of the issue: After isolating the main alarm, the issue gets fixed. [1]

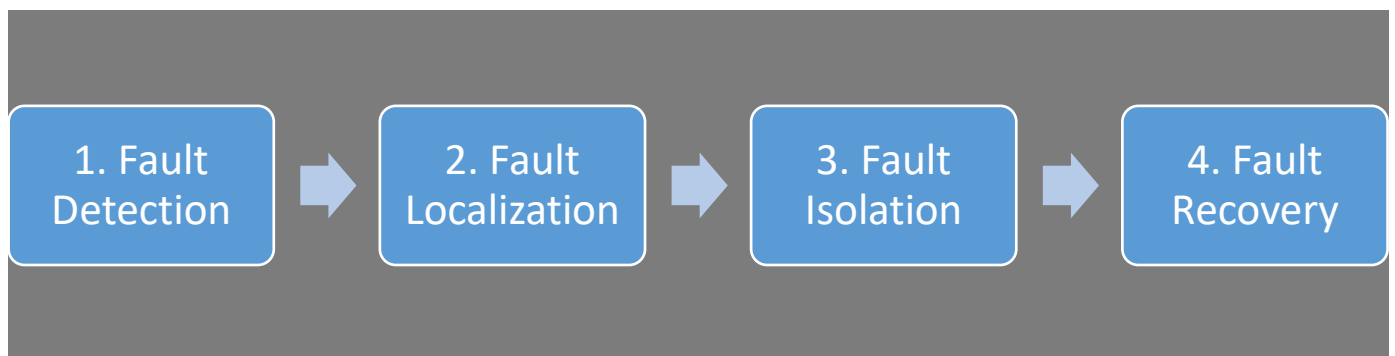


Figure 1:A generic network fault management process [1]

As shown in above figure, fault management system involves the whole process from alarm detection that is coming from the network devices involved in the communication to the full recovery of any problem that has occurred in the network for communication.

Previously, these fault management processes were handled manually using human force. But as the increase in the need of communication systems and network systems, these processes are now handled using machines. Data mining and machine learning techniques are used nowadays in order to manage faults in the telecommunication industries. Hence minimizing human control and human error in these crucial systems.

There is a department of network operation center in each telecommunication industry, whose sole purpose is to implement fault management system. They have installed different softwares that helps them to ensure this process. Network management systems and TEMIP ensures the detection, categorisation of alarms whereas other softwares like Service Manager helps them in creating tickets to the concerned teams which is responsible for the faulty devices. These teams rectify the problem and process tickets on Service Manager.

Data mining and machine learning involves number of steps that helps in processing a huge amount of data into a format that can be used further for decision making [2]. First step in the data mining is known as data cleansing in which irrelevant data is removed so that the decision making process gives a more

accurate results. This step also involves combining data from multiple source if any. Second step includes further examination of data and transforming it into a suitable format for the data mining algorithm. After this various data mining techniques are applied on the data and pattern evaluation takes place where unique and repetitive patterns are monitored, which helps further in addressing the issue [3].

Now machine learning / data mining not only helps in detection of various alarms, it also helps in alarming the teams beforehand. These tools are used to predict the pattern of alarms occurring on the devices using different or suitable algorithm and then on these predictions, pre-actions can be taken by the teams.

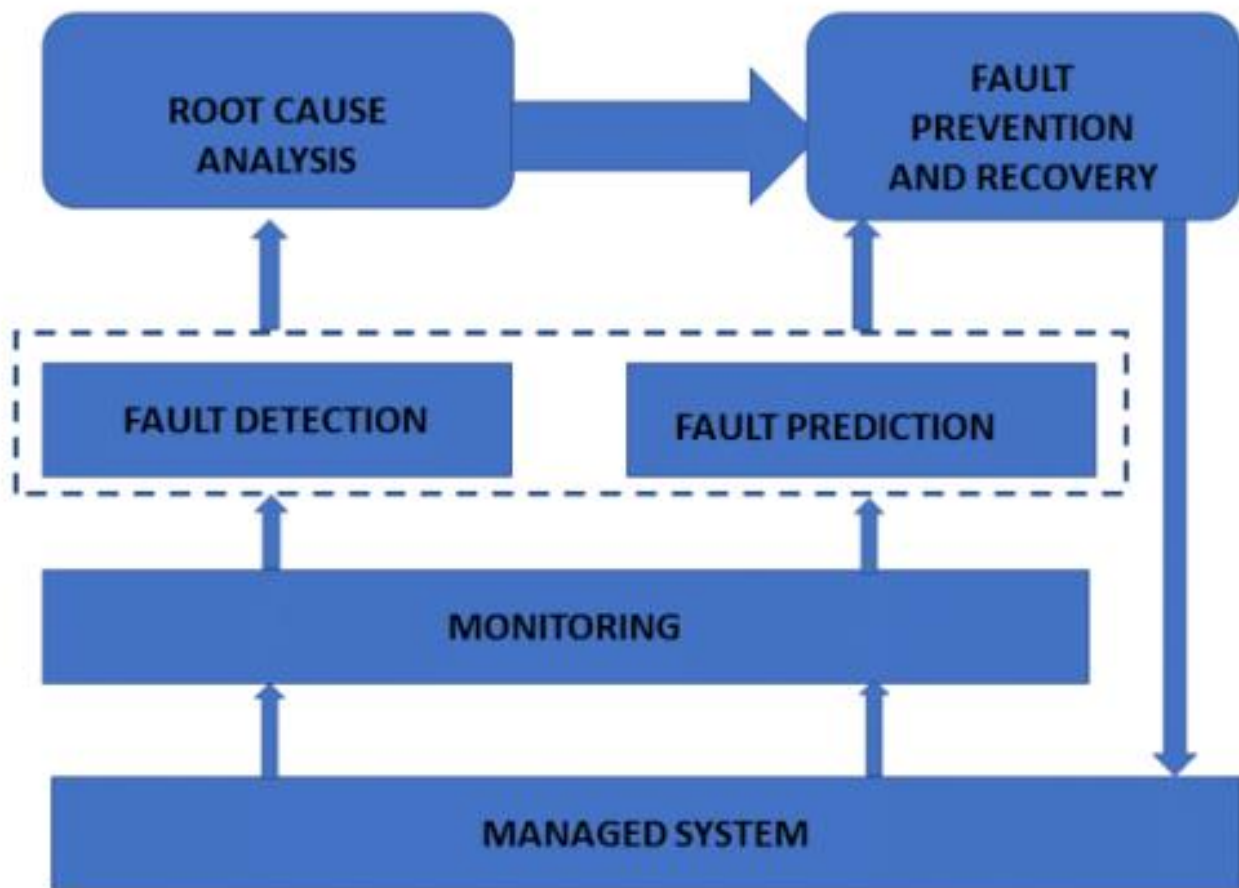


Figure 2: Key functions included in fault management system [4]

As shown in the figure above, the systems that are used in the fault management system, first there are managed systems like MSAG, MSAN etc the elements that are involved in communication. From the managed systems the alarms are pushed towards the monitoring softwares. These monitoring softwares categorise alarms into minor, major and critical alarm respectively, that is identifying the root cause alarm. From here fault prediction is also done which identifies if the alarm leads to an outage or a non-

outage alarm. After this the actions are taken accordingly that is for outage alarm teams must visit the site area and rectify the problem where as, for non-outage alarms early warnings are generated to the teams so that preventive measures will be taken for future fault occurrence.

Early alerts/warnings can be assigned to the various teams via SMS or Email to their incharge will help in rectifying the problem beforehand. Like for example, if the alarm of battery level on the threshold value is appearing then a alert or warning is to be generated to the field team telling them to be ready for the charging of battery as the device will go down when the battery charge reaches below the threshold value.

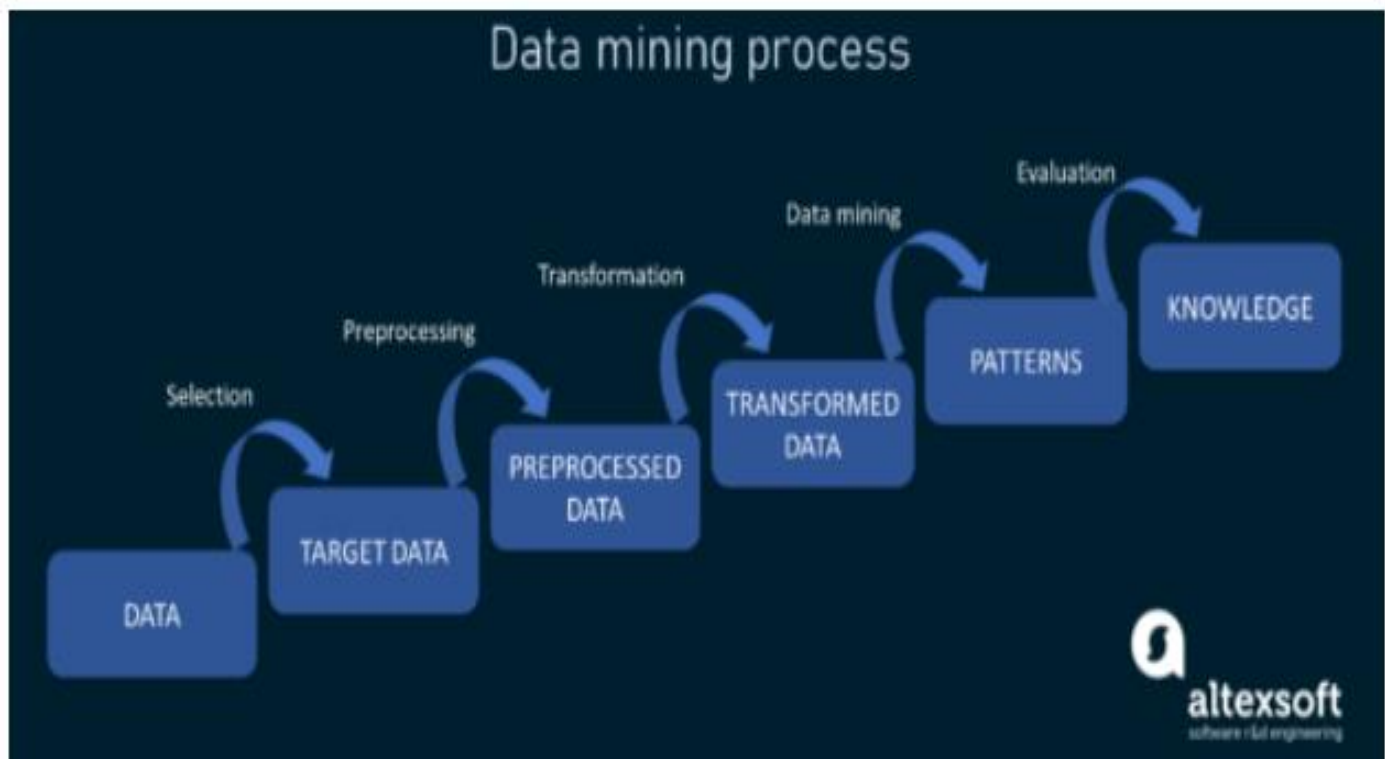


Figure 3 Online illustrating data mining process steps [5]

As shown in above figure, data mining process involves selection of data, on which data mining needs to be performed. Then the next step involves the processing of the data, that is including the specific attributes that can help in prediction process. Processing of the data also involves removing any recurring data or empty attributes that may cause error in future prediction. Now this transformed data is used to find pattern if any visible in the data itself. That will eventually help in future prediction by data mining algorithm. The more your data is refined the more it will help in algorithms working and the more accurate results will be achieved for the problem statement. After this step algorithm is made to learn on the data. Then the working of the algorithm is monitored. With any problem in data, above steps are repeated again.

Data mining is a very important part of machine learning. It is used to find the pattern in the dataset so that anomalies in data can be known which will eventually help machines to predict any anomaly occurrence in future. Data mining involves human intervention for completion of the task whereas machine learning has algorithm that learn periodically with time working as a substitute of human beings.

The comparison between machine learning and data mining is given below. This comparison shows that the data mining involves the techniques and methods to sort the whole data in order to find pattern or for prediction of any new occurrences. Whereas machine learning involves the methods or techniques that not

only requires any human intervention but also need to get machine powerful in order to make decisions on human behalf. These machines are capable of learning from their experiences.

Data Mining	Machine Learning
Data mining molds the data into a standardised format by processing the data using different methods.	In this mechanism, machines learn themselves from their past experiences with the data and improve themselves just like human beings.
Human intervention is required in data mining to sort the data into useful format for algorithm to work properly.	Machines learn themselves therefore no human intervention is required.
It serves the purpose of turning raw data into something useful in order to predict and forecast.	This serves the purpose of providing a complete autonomous system.
There are patterns that are present in the data. Data mining helps in order to find it.	It uses the sorted data and train the algorithm to predict the future occurrences of problem.
Transportation, retail, finance are some of the examples that use data mining.	Image recognition etc are the examples

Table 1 Comparison chart of Machine learning and Data Mining [6]

As per the above figure, data mining involves series of steps that help in finding a pattern in the data either repetitive or anything. These steps includes the collection of data, processing the data (in order to remove any repetitive items etc) . In other words data is transformed into a standardised form that helps in showin any patterns in the collected data. These sorting and cleaning of data needs human intervension. Mahine learning involves methods that helps machines to find anomaly in the data. And these machines learn with their experiences t

Literature Review:

There are pattern mining algorithms that have been used to analyze telecommunication alarms data for improving fault management system. A survey on fault management by Mourad Nouioua [7]describes association rules mining based approach in which analyzer defines association between multiple alarm sequesnces. This system however, is not suitable for long alarm sequences. Fault management system involves detection, isolation of alarms and then to find the root cause alarms and then to find the proper closure to the problem. This paper provides a detailed survey of different data mining and machine learning techniques that are used for proper fault management system.Lozonavu [8] proposed a method to discover alarm pattern using sequential pattern mining algorithm. In this algorithm a relationship graph is constructed between alarms and each relationship is given a weightage which is calculated by

confidence measure. These graphs represents a better view of network behavior and discovering of new network elements.

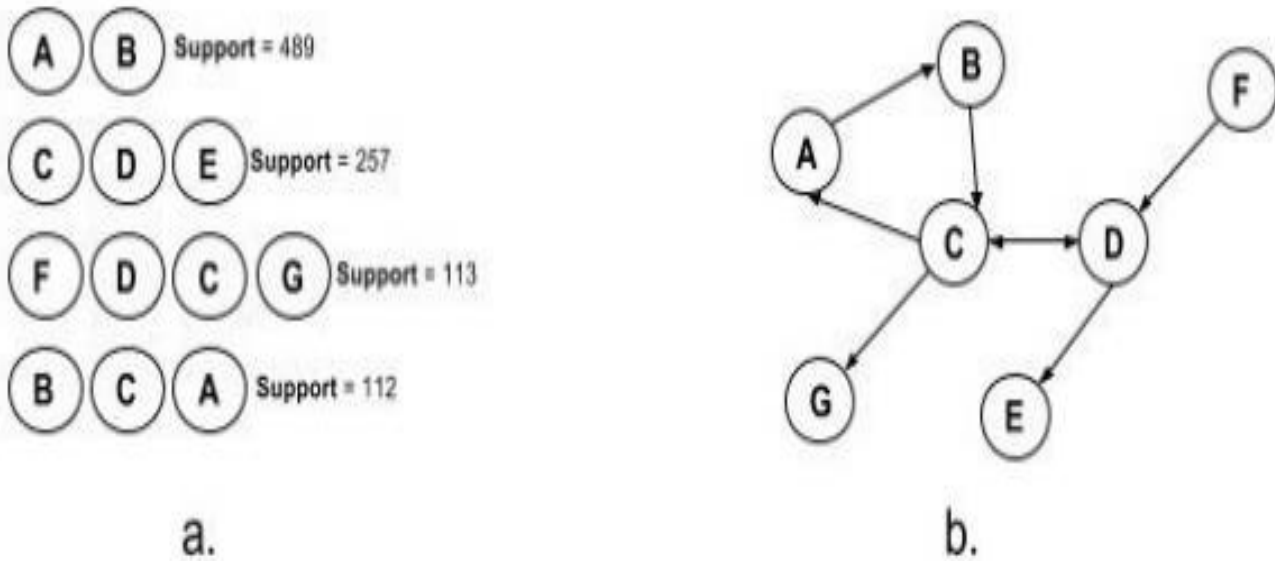


Figure 4 Transformation from pattern to graph [8]

As, shown in the above diagram, the transformation of nodes from pattern to graph helped the pattern mining better. Despite the order of the events these sequences are made by the occurrences of the events. Clustering algorithms are also used to group relevant alarms from a huge set of alarms and then finding the root cause alarm. K-means and K-medoids are two most widely used methods for clustering algorithms. In K-means, centroids are initially assigned a value and then all the remaining data is assigned to the nearest centroid. After that recalculation of centroid occurs until centroid does not move. K-means creates different clusters and a center point is measured in each cluster where all the points form a cluster around that center point.

The connection oriented data from broadband devices are used in order to find the quality of these classification techniques. It helps in the prediction of the connection oriented data. In this paper execution time for each algorithm is considered as then compared. [9] [10].

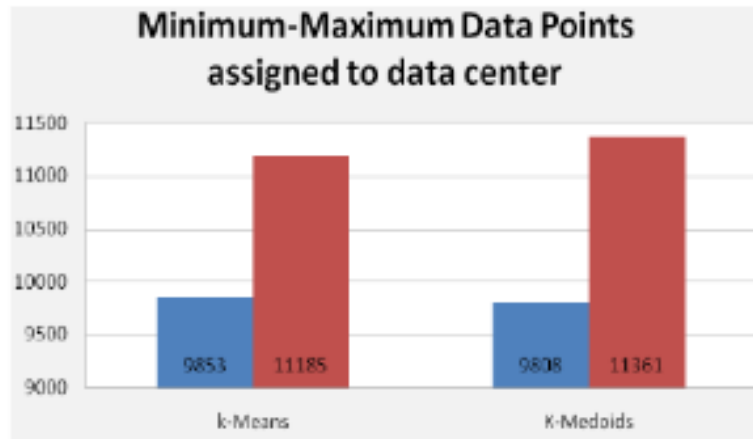


Figure 1. Min & max data access points.

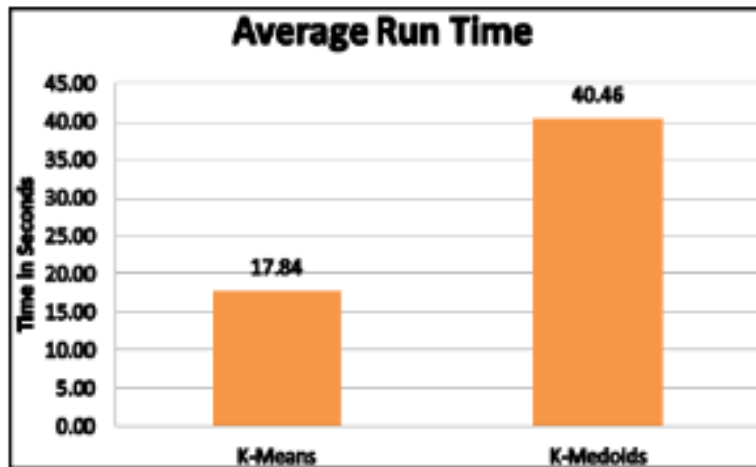


Figure 5 Comparison of K-mean and K-methodoid [9]

In the above figure as you can see the comparison of both k-means and k-medoids is shown. When maximum data points are assigned to the data center then k-medoids output is higher as compared to the k-mean. The second graph shown the average run time, which apparently k-medoids run time is greater than the k-mean algorithm's runtime [9].

This method may not necessarily find the optimal configuration as compared to global objective function, but it helps in making the correct choice for the telecommunication oriented data [11]. k-mean is used to check for the site team and assign the particular ticket to them. Since k-means is sensitive to initialization therefore a method of initialization is proposed method.

Whereas, in K-medoids algorithm, this algorithm initially selects random K-medoids for the data but on repeat tries to select a better choice. That is why this method is also known as representative object based algorithm. K-mean and k-medoids apparently work as same classification tools, both try to find the centroids and clusters of the data present [9].

Machine learning approaches are widely used for improving fault management system in the telecommunication industries. Artificial neural network based approaches include training of algorithm on supervised data so that prediction of desired outputs can take place [12]. Neural network also produces accurate results as compared to other algorithms. Telecommunication data is taken by the neural network and produces an output that helps in future forecasting of the system. The best example of neural networks is the fault management system, where the process of fault management system is completed

and future prediction of the events can also take place. Human brain is the inspiration behind these networks.

Hybrid approaches are also been used for prediction of anomalies in telecommunication industries and future forecasting. For example Mirjana Pejic [13] worked on the prediction of churn management by using three stage hybrid approach. Chi square is used to segregate different groups in accordance with the churn ration of the clients. Decision tree algorithm is also used to classify churn specific data in the clusters and produce outliers that will show the maximum churn ratios in the group. This paper have used k-means cluster analyses for identifying market segment in the customer churn dataset and then decision tree algorithm detected the churn ratio. Now the churn management happens in the way that these churn predictions are sent back to the custoer data base in order to improve the efficcieny of the system.

Another paper written by Yue Cheng in 2013 [14] on the use of advance version of Smith-Waterman algorithm in order to find patterns in alarm flood sequences which will eventually helps in determining the root cause in the alarm floods. Similarity index is calculated by a revised algorithm of Smith-Waterman. This paper helps in deternining the similarity index between alarms that are appearing in the flood. the sequences of the alarms are not takes into account but the similarity index is calculated by taking the alarm flood and compared with the neighboring alarms.

Saad Gadal, Rania Mokhtar [15], in their paper have used the hybrid version of k-mean array and sequential minimal optimisation(SMO) rating in order to increase accuracy of the anomaly detection

rate. This hybrid also helped them in reducing false positives hence accuracy is increased by classifying intrusion.

Yue Li [16] writes in his paper of using K-NN and decision tree algorithms for the accuracy increase and reduction in the rate of false alarms. Positive detection of 99.7%, false alarm rate of 0.2% and the accuracy of 99.6% is achieved when they have extracted optimised information through K-NN and DT hybrid algorithm. The false positive ratio is however decreased and the accuracy of this algorithm is increased. Hybrid algorithm of K-NN and DT appears to be the best choice for telecommunication data.

This paper [17] by Ying Zhou, the method used in this paper provides high precision knowledge of network security thus helped in the security and stabilisation of cloud based networks. the false positive rate and the accuracy are somewhat better when these algorithms are used in the telecommunication data set.

Paper [18] by Boyuan demonstrates a method called Dirty data base alarm prediction for future forecasting of alarms. In this paper self optimisation networks's data is used for this algorithm with 274 nodes and 487 links. And at the end of the paper describes that the accuracy is somewhat higher for different type of alarms. This data is collected from the old fashioned way optical equipment. And it is said that even after processing of data and balancing the data. This paper demonstrates the DAP technique for these networks and it turned out that these technique are very useful and provide results with great accuracy.

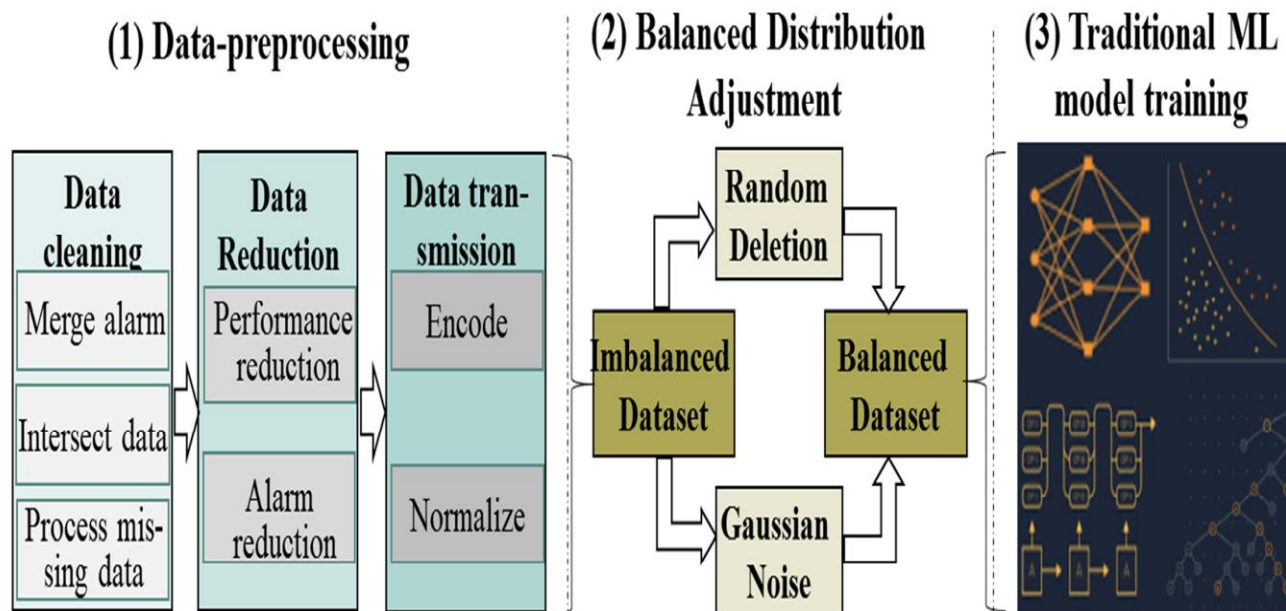


Figure 6 Dirty data base algorithm procedure[19]

In the figure above it can be seen that the DAP procedure involves data processing, balanced distribution and machine learning training of models. First data is processed that is all the incomplete or duplicate data are handled that the balanced distribution of data is done then the model is trained on this processed data and then the outcomes are out.

This survey [19] presented by provided a detailed review of noth data mining and machine learning algorithms in cyber netwroks. For determining the effectiveness of each methods presented in this survey for both misuse and anomaly detection in cyber networks, there is not one criteria but multiple criterias for each methods. In this paper the methods of machine learning and data mining are summarised based on their accuracy results for anomaly detection. Citation based these methods are

organised and also on the significance of evolving method. Cyber analytics is considered to be very important for smooth network connectivity therefore this survey holds a good position in the cyber security methods.

In this paper [20] different network anomaly tools and methods are discussed. For the safety of the network there are multiple tools, softwares that are used to protect this network form any intrusion. Intrusion detection involves the outlier or the alarm in a sequential alarms. In order to detect that it is required for different tools and softwares to work together. In this paper cartegorisation of existing anomaly based detection methods and system based detction methods are explained in detail. The tools that are reuired for anomaly detection and datasets that ae used by major reserchers for anomaly detections are described in detail in this paper.

This paper [21] emphasise on the use long short term memory (LSTM) which detects the relationships between incoming traffic packets in communication networks, this criteria is used for decreasing any false positives thus increasing the efficiancy of the anomaly detection algorithm in the network. This method also shows that the detection rate, false alarm rate and accuracy could be increased as compare to the other machine learning tools like artificial neural networks. This method has provided with detection rate 20%, 60% false alarm rate, 15% accurate results and it required the training time to be 70%. These techniques have also increased the speed and efficiency of the system. The data used was from the simulated data but the complex big data was handled with accuray by these methods. This

paper by Padmasiri [22] is a survey that provides detail review of the intrusion detection in the networks based on the deep learning algorithms. Detection using deep learning techniques is very fast, efficient and accurate as compare to other detection algorithms. Each approach is discussed thoroughly, their challenges, their weak points and also the advantages of using them as compare to others. CNN and LTSM are the successful algorithms for intrusion detection because of the scalability nature of the communication networks. This survey provided a detailed review of RNN, DBM methods for supervised learning where as GAN, RBN and Autoencoder is used for unsupervised learning. The branches of RNN includes Bi-RNN, LSTM and GRU as shown in the taxonomy given below. These all deep learning methods are reviwed in this paper and at the end CNN and LTSM are said to be the accurate ones and most reasonable ones for the scalable network.

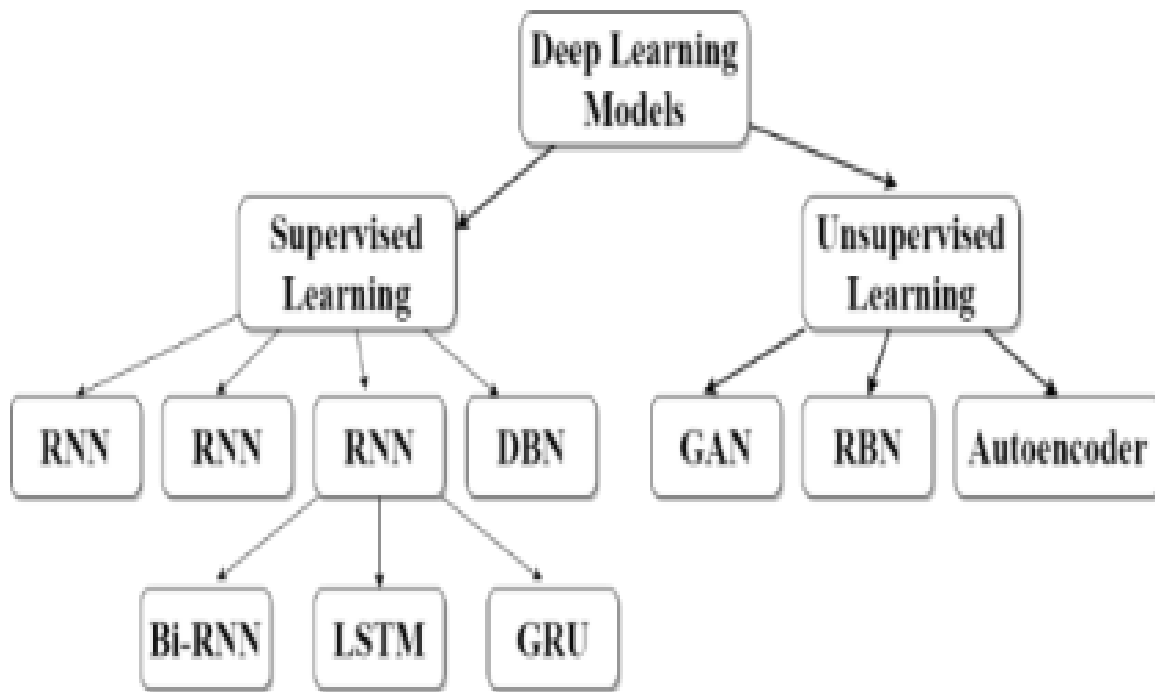


Figure 7 Deep Learning methods taxonomy[22]

This paper [23] perform the alarms correlation and root cause analysis on the alarms of telecommunication company. It uses rules management system and reinforcement learning in order to find the rules and also selection of rules in improved. They have emphasised on the rule generation part and also have performed testing on a huge data set of communication network. This correlection method appeared to provide accutrate results. The figure shows below is the proproposed rules management system that they have proposed.the alarms tickets or data entry by the network adminintrator wullmove to the assurance warehouse in rules management system. Form where using text mining rules are

generated and are then stored in the database inside the system. alarm is also prioritised accordance with their category and also sent to the management.

This rules management system is quite helpful in generating rules for the telecommunication alarms and the tickets(processed alarms). This also helps in network administrator to not involve to much in monotoring of the alarms and also applying ticketing criteria of the different alarms present. It takes the duty of network administrator and work on it autonomously and then the work is done that is helping the telecommunication company to get successful

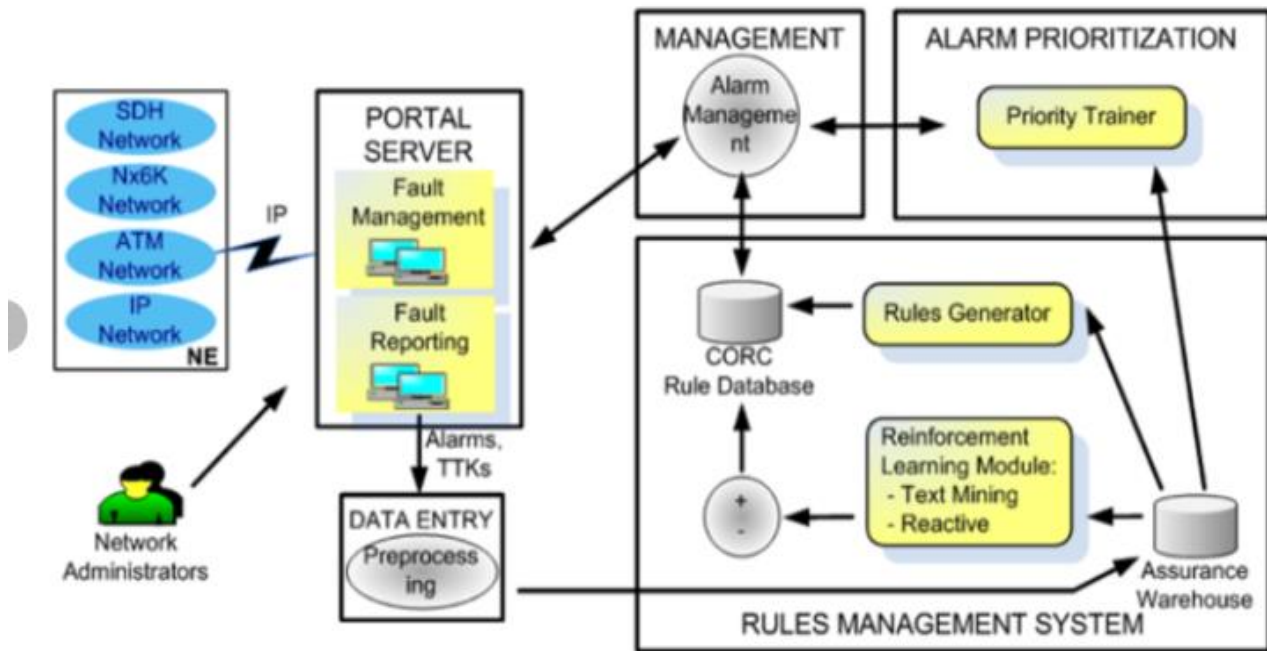


Figure 8 Proposed Alarm Management System[23]

Hence, the accuracy of the system proposed turned out to be very accurate for rules generation and eventually helping in the default management system. This paper [24] shows the LSTM technique as the revised version of RNN. It shows the ability of long term dependencies as compared to other machine learning algorithms. LSTM has performed better than Support vector machine (SVM) and Naïve Bayes. 0.8713 is the achieved training accuracy and 0.8483 is the achieved testing accuracy of the LSTM algorithm. This model consists of 10 neurons corresponding to the 10 features, there is a hidden layer with 10 neurons and also the output layer consisted of 5 neurons. The optimiser named “rmsprop” is used which is suitable for this type of dataset and long term dependencies. Also the accuracy, precision are all accurate for this model as shown in this paper. In this paper [25] Hsu et al, proposes two deep learning methods used for intrusion detection. First model is LSTM model while the second model is CNN-LSTM. The second model has achieved greater accuracy as compared to the LSTM model alone therefore the decision to use both the models for intrusion detection is a very good decision. CNN-LSTM model achieved 94.12% accuracy for binary classification while it has achieved 88.95% accuracy for multi-class classification. CNN-LSTM has been used for passing the vectors and features to the LSTM model as input. This paper included CNN to have learnt the spatial data while the LSTM have learnt the temporal features.

Methodology:

Dataset:

The data that we have used for this problem is gathered in a month long span. Probes were installed at the telecommunication devices ends and continuous monitoring on the capturing of the alarms against the device names were held so that all of the data could be collected properly, thus helping in the correct predictions of the attributes.

There are millions of alarms that are generated from multiple network devices daily. These alarms includes card related alarms, power related alarms, hardware related alarms and software related alarms.

These alarms indicates fault in different parts of communication network.

For example:

1. Device may go offline is an alarm that is related to the power of the device/network element.
2. Card is offline is an alarm that indicates the fault in the specific card of the network element.

We have included in our dataset only hardware related alarms.



Figure 9 Shows an MSAG present on the locality [26]

As shown in the figure above, these MSAGs produce different alarms that are being captured on different softwares depending on the vendor (i.e Huawei,Nokia,ZTE etc). These alarms are then beautified through a beautifier. Beautification of an alarm involves separation of alarm name, machine name, category of the alarm (that is critical alarm, major alarm, minor alarm etc). Through the process of beautification an alarm is readable by the human as well as actions can be easily taken on these

alarms. Each device is connected to its network management system (separately designed based on the compatible protocols of communication).

Network management system is the user interface that provides visual representation of the network present on any locality. When the network devices are in healthy state that shows the connectivity in healthy “ green “ color. But when the network element goes offline due to some fault then the connectivity turned “ grey “ in color.

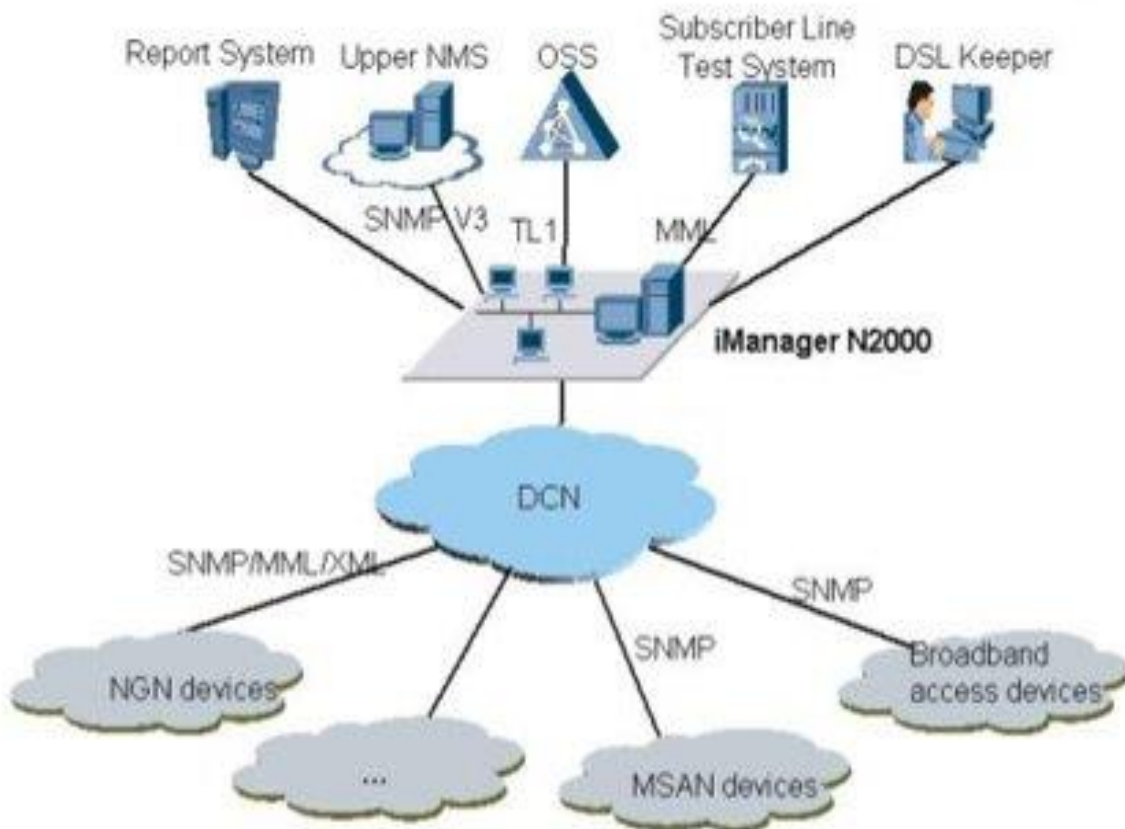


Figure 10 Showing Huawei Network Management system [27]

As shown in the figure above, all the telecommunication devices (NGN, MSAN, broadband access devices) are connected via DCN network with the network management system (iManager N2000). The protocol for communication is SNMP or for the NGN devices which are now obsolete SNMP/MML/XML.

Now after the alarms land on the network management system, it gives a visual representation of the network and where the fault occurred.

The visual representation of the network on NMS is given below:

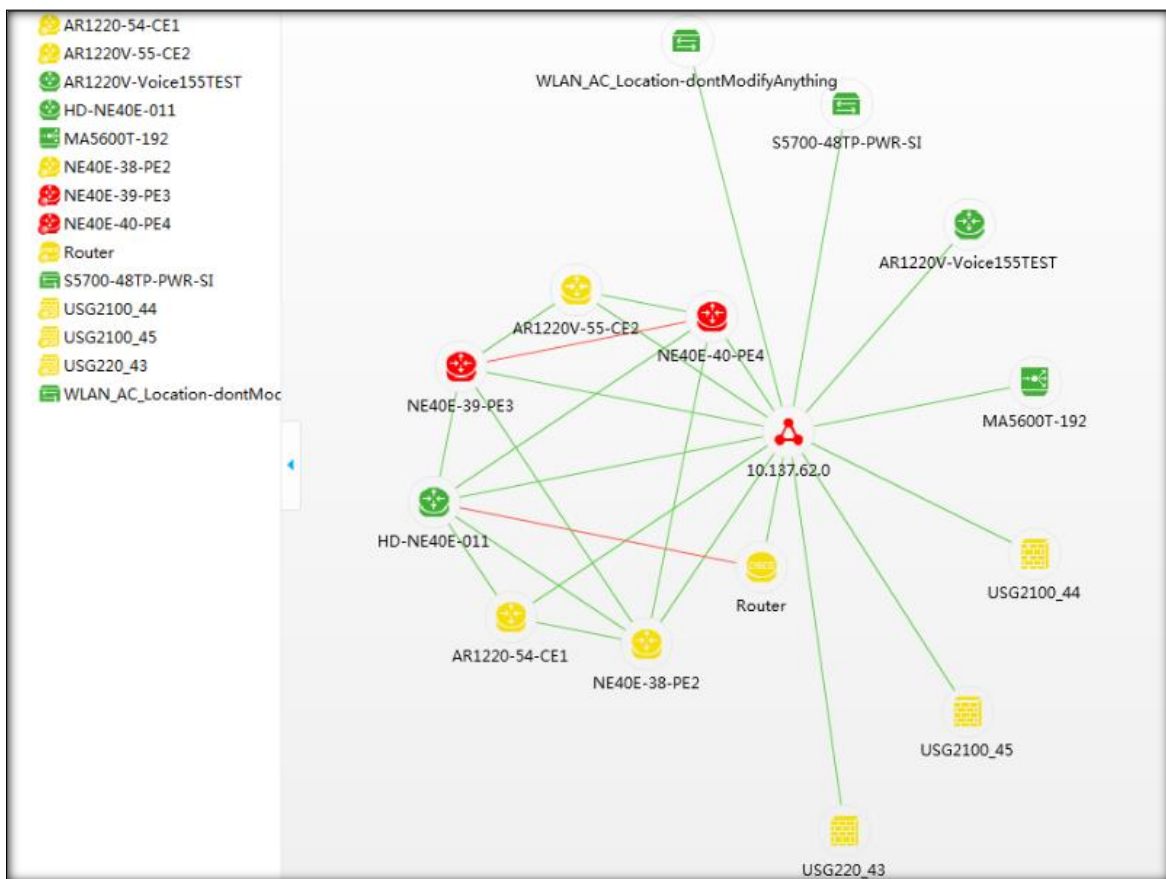


Figure 11 Visual representation of a network on NMS [28]

In this visual representation, once the alarm occurs on any node, it will turn red (showing fault) or any device that goes offline will show as a grey in color (showing a device offline). These NMS will push these alarms to a software named TEMIP (company Huawei) where each alarm is shown in tabular form for teams to understand alarm properly and take action accordingly.

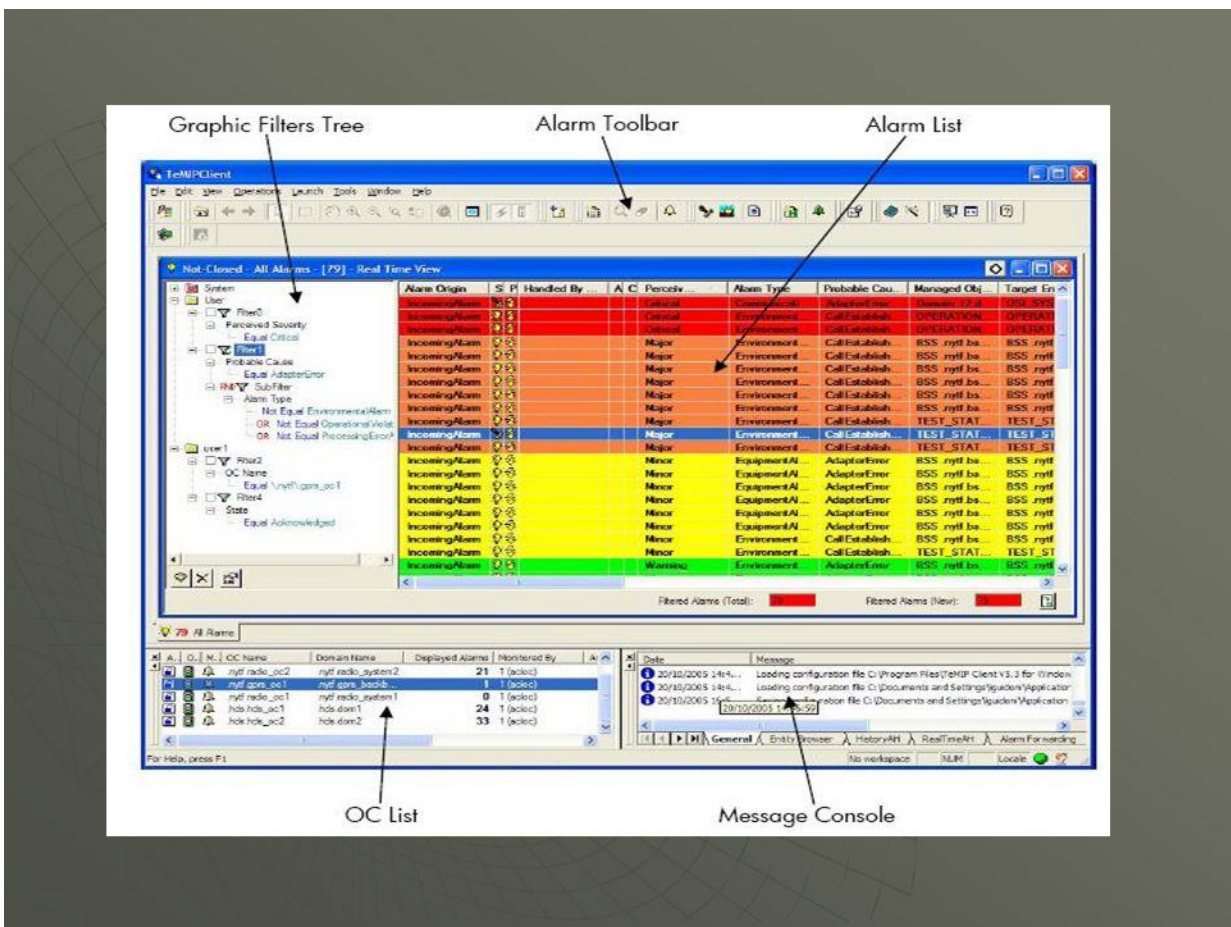


Figure 12 Temip Client preview showing different alarms categorise by the color of alarm [29]

Network management systems also push data towards the Data bases for storage of these alarms. You may retrieve this data using SQL code. Or reporting or predictions for future can be easily made on these data stored in the database.

My data set represents different type of alarms on different telecommunication devices that may or may not lead to disconnection of the device from the network. Few alarms that tells the condition of battery or related to card etc in the device are not that major alarm, but the alarms that tells the disconnection of device or service downgrade are the alarms that needs to be sent to the field teams for rectification.

Because these alarms effects the communication services been provided to the customers.

This data set is captured in the duration of 1 month. Total records of the data set contains approximately 20,000 rows. This data set contains alarms from different devices and it mainly include hardware alarms that is if the communication between the devices goes offline. Or the card present in the devices for telecommunication goes disconnected. These cards may cause degradation in the services been provide to the customer where as the offline of the device may cause delay in telecommunication services overall. Thus affecting the customers all together. This affecting of customers may cause customer churn in the telecommunication company.

Alarm name is also provided in the dataset in order to know about the nature of an alarm. Second attribute is NE name that means Network element name (or the device name) on which an alarm has occurred. Third attribute is the creation time or the occurrence time of the alarm on the device.

Repetition of alarm can tell the faulty device in the network. Device names are also defined in the data set that shows uniqueness to the alarm occurring in a particular device. Last column tells the nature of alarm whether it is an “Outage” causing alarm or not.

Following are the attributes of the dataset:

1. **NE_Name:** This tells the network element name in the telecommunication network
2. **Alarm_Name:** This column shows all the alarms appearing on any particular network device.
3. **CREATIONTIMESTAMP:** The appearance time including date etc of a particular alarm on the device.
4. **Outage/Non-Outage:** This column tells if an alarm is an outage leading alarm or not leading to an outage

First Ten Rows

```

                                NE_Name \
0  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
1  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
2  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
3  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
4  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
5  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
6  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
7  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
8  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...
9  S13KHICTXM009//10.138.1.26-S2-KHI-Chamber OfCo...

                                Alarm_Name CREATIONTIMESTAMP  CLEARTIMESTAMP \
0      Communication with the device failed    9/1/2020 10:03    9/1/2020 10:06
1  Port link status changed from up to down    9/1/2020 10:05    9/2/2020 7:55
2  Port link status changed from up to down    9/1/2020 10:05    9/2/2020 7:55
3  Port link status changed from up to down    9/1/2020 10:05    9/2/2020 7:55
4  Port link status changed from up to down    9/1/2020 10:05    9/1/2020 12:25
5  Port link status changed from up to down    9/2/2020 6:34    9/2/2020 9:41
6  Port link status changed from up to down    9/2/2020 6:34    9/2/2020 9:41
7  Port link status changed from up to down    9/2/2020 6:34    9/2/2020 9:41
8  Port link status changed from up to down    9/2/2020 6:34    9/2/2020 8:28
9  Port link status changed from up to down    9/2/2020 6:34    9/2/2020 8:46

Outage/Non-Outage
0      Non-Outage
1      Non-Outage
2      Non-Outage
3      Non-Outage
4      Non-Outage
```

Figure 13 Python code showing ten rows of the Data set.

Describe the Dataset

```
NE_Name \
count    20133
unique    9
top      P07LRESNDM073//10.139.146.110-LTN-SND-C08HafzR...
freq     10960

Alarm_Name CREATIONTIMESTAMP \
count    20133    20133
unique    72      6264
top      The Ethernet port link status changes from up ... 9/2/2020 13:14
freq     4520    1078

CLEARTIMESTAMP Outage/Non-Outage
count    20133    20133
unique    4974    2
top      (null)    Non-Outage
freq     6570    19598
```

Figure 14 Python code showing description of the data set.

First we need to import the data to the program, then we need to prepare the data set for the algorithm.

Preparation of data set involves removing null values and also removing any duplicate entries in the data.

These errors may affect the outcome of the algorithm. As have seen previously there are multiple

techniques in data mining and machine learning that can be used for prediction purposes. Here we are

using decision tree classifier for classification and prediction of the alarms on the machines.

Decision Tree Classifier:

Now let's start to talk about the algorithm that was used in this program. Decision tree classification is a

major example of supervised learning. It creates a tree like decision structure of all the possible tests on

the attributes. Each branch represents the outcome of the test (which is 'True' or 'False') and each leaf

node represents a class label. For this dataset, since we need to predict the occurrence of alarm on the

specific node at the specific time that's why decision tree was the suitable choice for such problem.

Decision node is the root node of the decision tree diagram from where the decision making starts, once

the first decision which is the crux started then onwards there are multiple decision nodes throughout the

decision tree diagram. The results of the decision nodes are the leaf nodes that are either yes or no

depending on the decisions that the decision tree is making. One decision node and two leaf nodes forms

the "sub tree" in the diagram. There will be multiple numbers of decision nodes and leaf nodes through

out the tree but only one root node.

This scenerio is better understood by he diagram shown below:

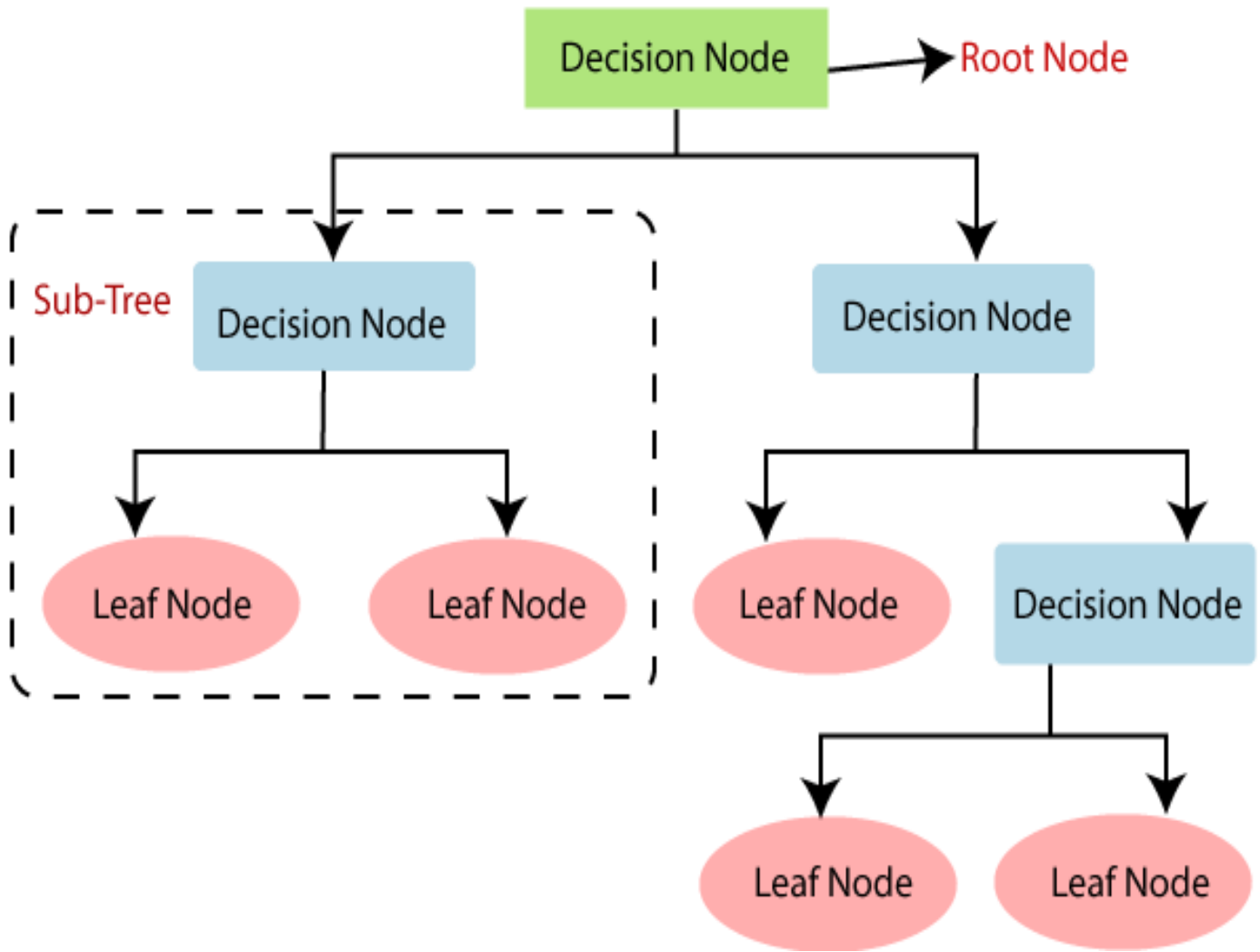


Figure 15 A generic Decision tree work flow diagram [30]

Decision tree rules have following form:

if (first condition) and (second condition) and (third condition) then outcom . [27]

In the following example the data variables are creation timestamp and network element name and the target variable is alarm name. First we need to divide the data set into two data sets. In the first data set I have the dependent variable (Alarm Name & Nature of alarm) and in the second data set I have the independent variable (Device name), as shown in below code:

```
#creating two separate datasets for inputs and outputs

X= alarms_data['NE_Name']

Y= alarms_data.drop(columns=['NE_Name'])
```

Now, we need to train the classifier with some data and then test the classifier with the required inputs for predictions. We will split the data by using `train_test_split` method.

```
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,train_size=0.8)
```

This method returns a tuple therefore we are splitting the data into training and testing datasets separately. `Size=0.8` means 80% of the data is to be used to train the algorithm and 20% of the dataset is then used to test the classifier.

```
model=DecisionTreeClassifier()
```

```
model.fit(X_train,Y_train)
```

Above code represents the fitting of the classifier on the training datasets, both data and target variables.

Now we will predict on the testing data set:

```
prediction=model.predict(X_test)
```

After this accuracy of the classifier is also calculated by comparing this output to the Y_test separated earlier. If it's 100% then it means that classifier is working properly. The split of the training and testing datasets also effects the accuracy of the classifier. With decreasing the split percentage (i.e 20%) will decrease accuracy as well.

```
score= model.score(Y_test,prediction)
```

```
print(score)
```

For accuracy measurement, both the input arrays need to be 1d arrays, therefore we will transform the 2d array to 1d array by using command:

```
Y_test.values.ravel()
```

Then the score between Y_test and prediction variables are calculated.

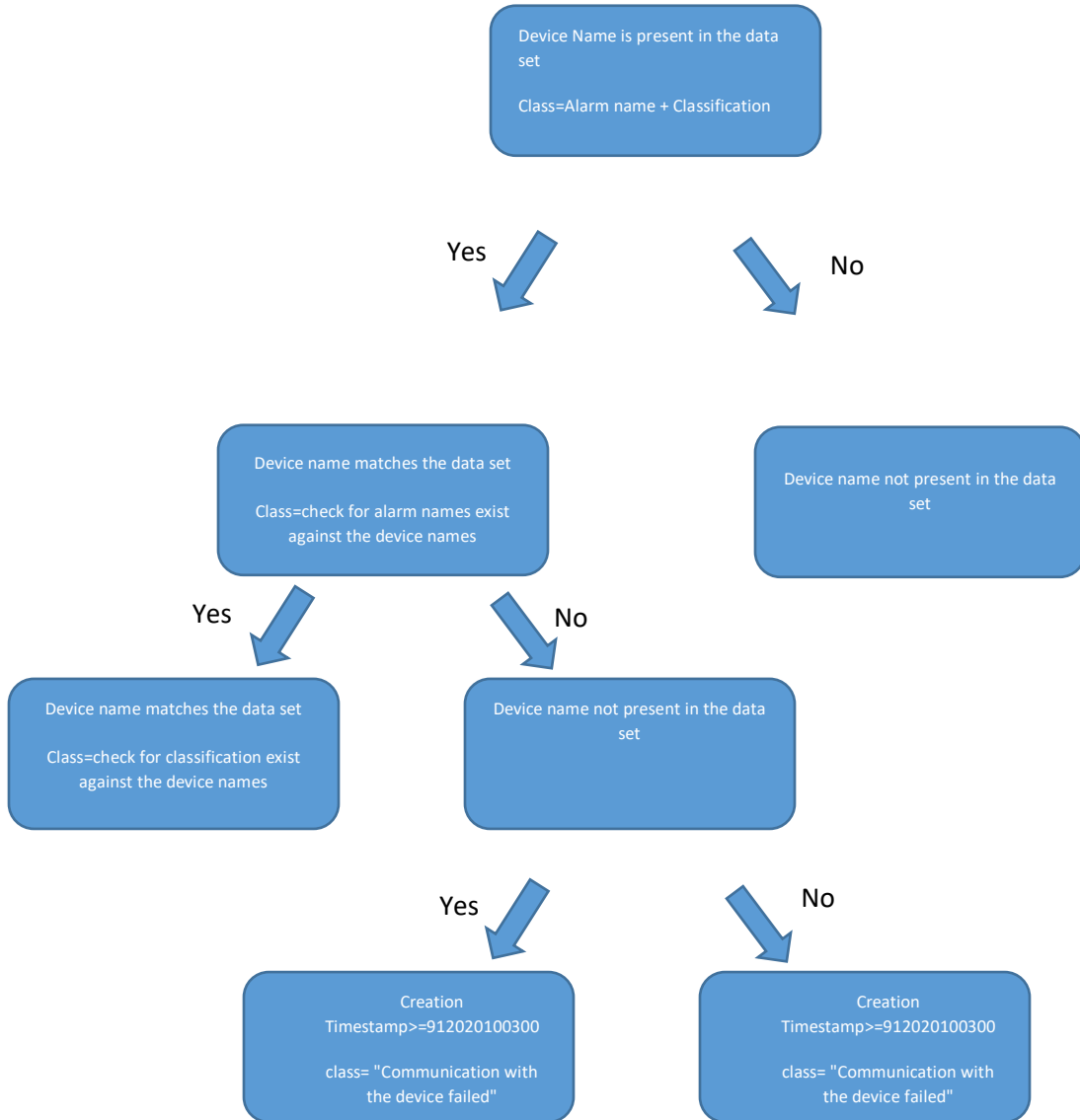


Figure 16 Flow diagram of decision tree working for this example.

The above flow diagram of decision tree is the drawn output which can be viewed using following command:

```
Tree. Export_graphviz(model,out_file='output.dot', feature_names=['alarms','device name'],  
class_names=sorted(y.unique()),label='all',rounded='True',filled='True')
```

The file “output.dot” is then opened with visual studio where graphviz add-on is needed to be installed.

After this the output of the decision tree algorithm is viewed visually.

There is the decision node that is the root node from where all the decisions have started that either the particular device name exist in the dataset. If yes then it will do the other decision of if the alarms are present against the required device name. Then it will check the nature of the alarm either it is an outage alarm or a nin-outag alarm. Then it brings out the output of predicted alarm along with the nature of alarm for the required communication device. Similarly, it predicts the entire testing data set that has fed to the algorithm and provide with required outputs.

Now the teams that are responsible for the respective devices are warned for the future occurrences of the alarms on these machines. So that they may take actions promptly. This future forecasting of the alarms on the telecommunication device provides a proactive approach to the dealings of the fulats in these

devices. Thus increasing the efficiency of the network and which will eventually lead to less customer churn for telecommunication companies.

Results:

Accuracy is calculated for producing the accurate results between the predicted outcomes of the algorithm and the actual results against the dataset used for testing of the algorithm. The score gives the accuracy rate between these variables.

As you can see from earlier topic, since we had two data variables as an input for the algorithm therefore use of simple algorithm like decision tree for prediction in our program was a good decision. It successfully predicted alarms on specific machine at specific timestamp. With more data attributes and other algorithm we can build more advance version of this alarm prediction program.

Beforehand prediction of alarms on specific nodes/machines will turn out to be a major help in telecommunication industries where a slight decrease in the service may result in the customer churn for the company. Field teams may be given alert to prepare for any type of alarm that is to occur in a particular machine.

An alert system may be integrated with this prediction outputs that will generate SMS or emails to the respective in charge of the areas where machines are placed so that early actions are to be taken to avoid

any problem. One more advantage of such program is that outage alarms may also be controlled via these predictions. Outage alarms are caused when the device/machine completely shut down causing a major delay in the communication services provided to the customers, thus resulting in unhappy customers. These outages alarm when appearing on a device means that device is faulty and quick action is required to rectify the problem.

10-fold Cross Validation:

10 fold cross validation process means input data is split into 10 groups of separate data inputs instead of just two groups. For each time one group is been used for testing while all the remaining will be used for training of the algorithm.[19]

Following steps are included for 10-fold validation process:

1. First it is required to shuffle all the data present so that any order can be removed from it.
2. Split your data into 10 small groups
3. Now take one fold for testing and remaining all the folds for training of algorithm
4. Train the algorithm and test on that one fold and note down the results
5. Now repeat whole process by using the next fold.
6. And at the end sum up all the score and get the mean score.

Now, the dataset is divided into 10 equal parts and one part is used for testing of the algorithm that is decision tree. While the other parts contribute in the training of the decision tree algorithm. All the parts one by one are used for the testing purposes of the algorithm. Then the accuracy of each dataset proportion is measured using comparison with predicted and the actual possibilities of the inputs. The accuracies measured are given below:

This flow diagram shows the k times validation method of the data set. Data is divided into 10 parts and then the 10 parts are one by one used for testing the model while the rest is used for the training of the model. The accuracy turned out to be approximately same as the training and testing dataset proportions are equal.

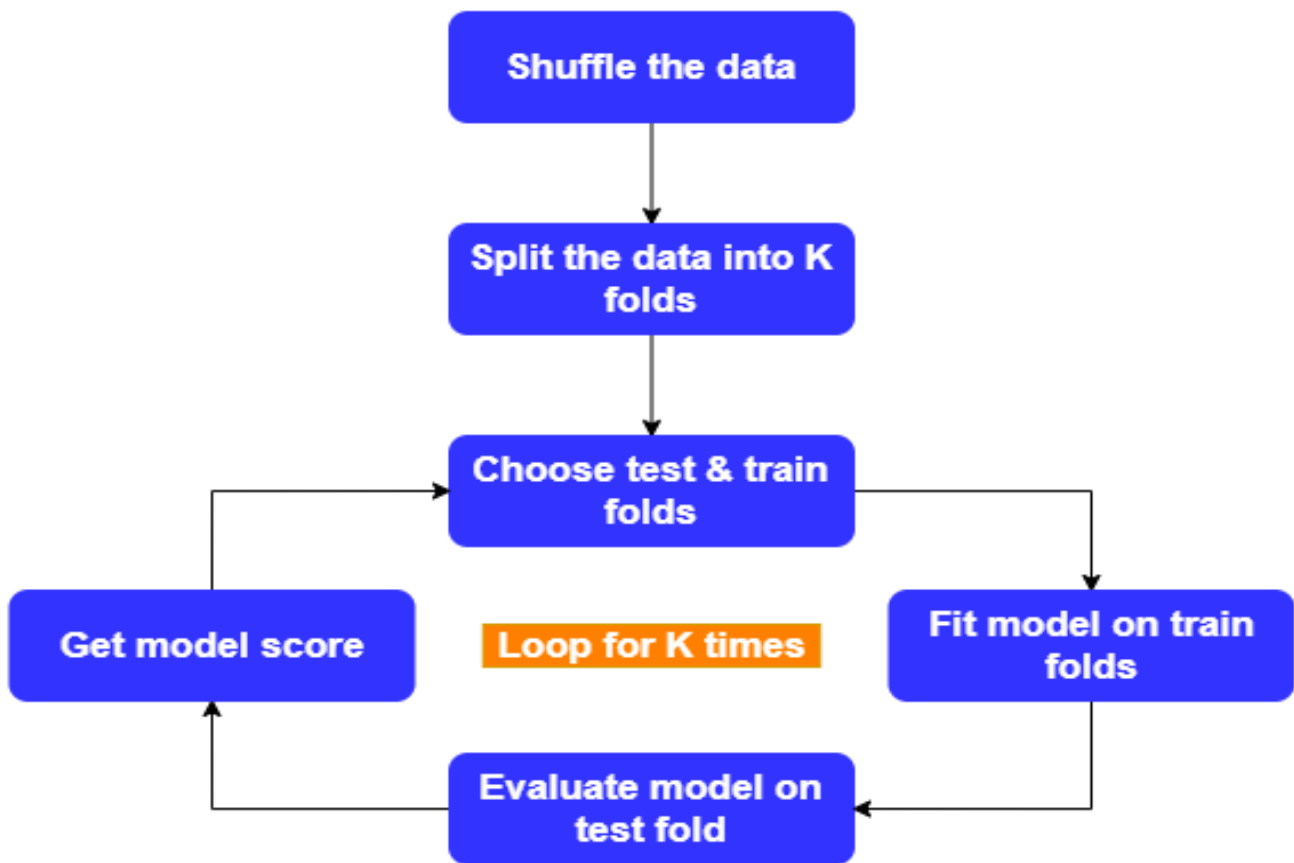


Figure 17 K times cross validation flow diagram [31]

K times cross validation is the process to validate the data set that we have, means the consistency throughout the dataset. Dataset is divided into the dub datasets depending on how many folds that we need to perform on the dataset. In each cycle one of the dataset is used for testing purpose while the

remaining used for testing purposes. Then the accuracy of the model is measured on the findings. If the dataset is consistent then the accuracy should be equal through out the datasets.

k times validation method of the data set. Data is divided into 10 parts and then the 10 parts are one by ones used for testing the model while the rest is used for the training of the model. The accuracy turned out to be approximately same as the training and testing dataset proportions are equal.

The accuracy table is shown below:

S.No	Dataset number	Accuarcy measured
1.	Accuracy1	0.668
2.	Accuracy2	0.706
3.	Accuracy3	0.713
4.	Accuracy4	0.678
5.	Accuracy5	0.688
6.	Accuracy6	0.682
7.	Accuracy7	0.669
8.	Accuracy8	0.669
9.	Accuracy9	0.663
10.	Accuracy10	0.664

Table 2Accuracy measured against 10 cross validation of data set.

These accuracies when plotted on the graph, then we can have following trend among them:

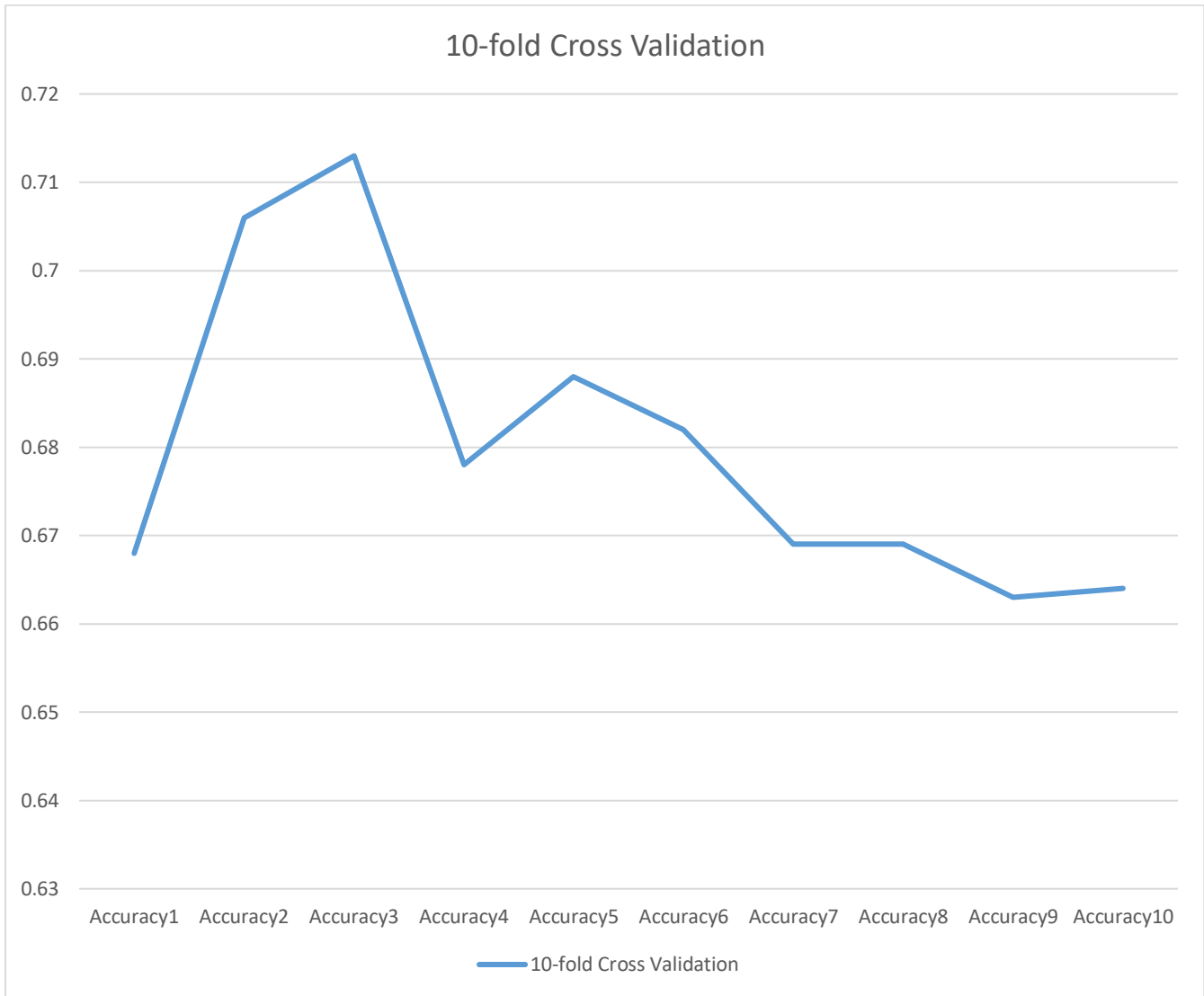


Figure 18 Showing Accuracy trend for 10-fold data analysis

It is observed that there is a consistency in the accuracy for all the 10 fold data used. Since only one fold is used for testing while the remaining were used for the training of data. First the data set is divided into 10 parts then 1 part is used for testing and the rest is used for training purpose. The 10 times decision

tree is trained with 9 parts of data and one by one all the parts are used for testing purpose. The accuracy measured is therefore approximately similar. This shows the consistency of data and algorithm, no matter what part is used for training and testing, the accuracy is similar through out the data.

This accuracy is measured by comparing the training output with the testing outputs. This process is called 10 fold because of the division of data into 10 parts.

Now let's observe accuracy measurements of our algorithm for different proportions of data used for training and testing.

If we train algorithm with 80% of the data and 20% is kept for testing then the accuracy is :

Accuracy=0.998

If we train the algorithm with 20% of the data and 80% is been used for testing then the accuracy is:

Accuracy=0.421

If we train the algorithm with 60% of the data is been used for training and 40% of the data is been used for testing, then accuracy is:

Accuracy=0.489

If we train the algorithm with 40% of the data is been used for training and 60% of the data is been used for testing, then accuracy is:

Accuracy=0.411

If we train the algorithm with 50% of the data is been used for training and 50% of the data is been used for testing, then accuracy is:

Accuracy=0.444

According to this test, the proportion of datasets into training and testing is directly proportional to the accuracy measured. the more the algorithm is trained on huge number of dataset, the more accurately it can predict future occurrence of the alarms whereas the less the data is used for training the algorithm, the less accurate is the result it produced.

S.No	Dataset Proportion	Accuarcy measured
1.	Training: 80% Testing: 20%	0.998
2.	Training: 20% Testing: 80%	0.421
3.	Training: 60% Testing: 40%	0.489
4.	Training: 40% Testing: 60%	0.411
5.	Training: 50% Testing: 50%	0.444

Table 3 Accuracy measured against different proporation of training and testing datasets.

Let's see the graph below:

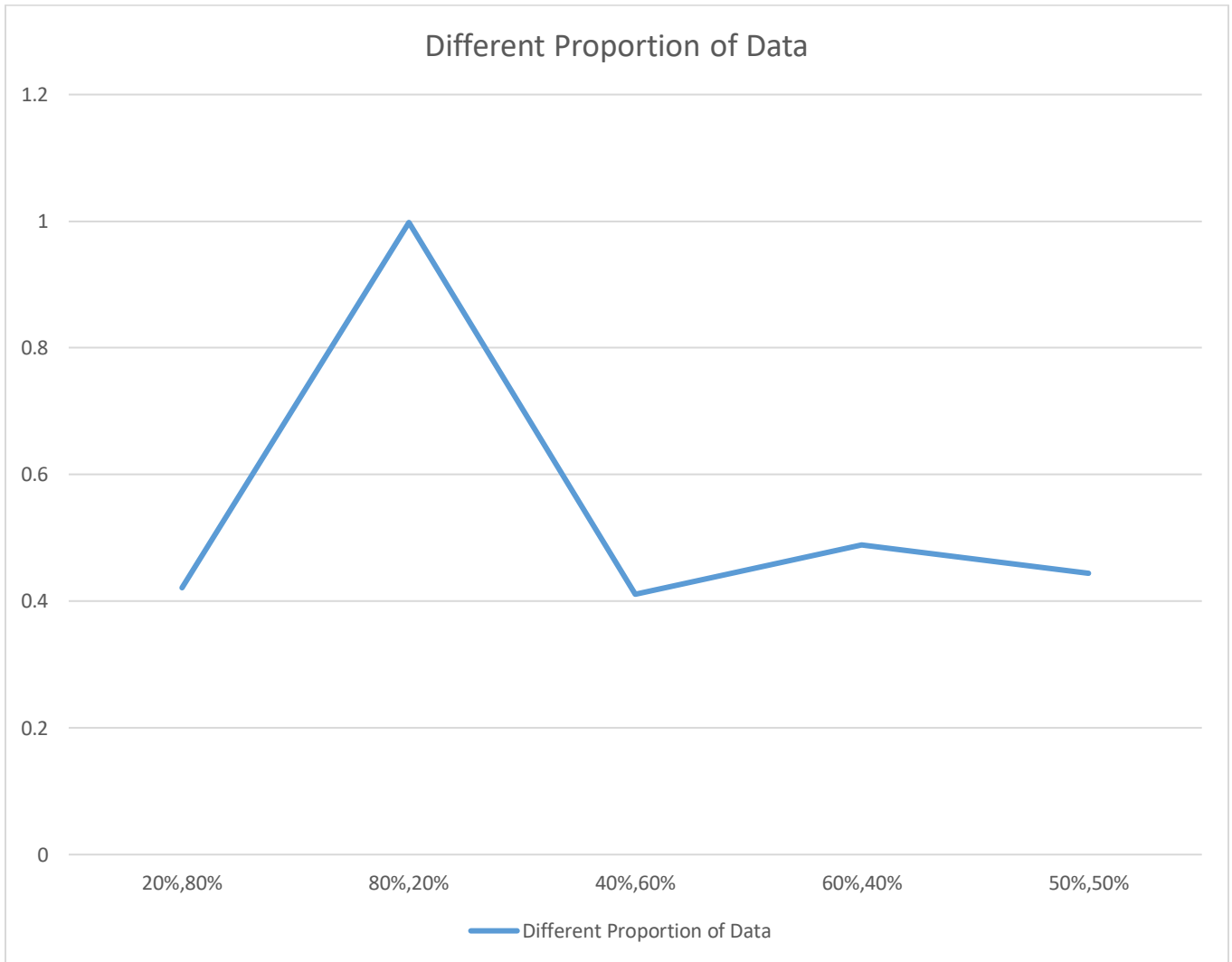


Figure 19 Showing trend between accuracy and different proportions of training & testing data sets

This trend shows the relationship of accuracy with the division of the dataset. The more the dataset is divided in a way that more data is used for training and less is used for testing the algorithm, the better

the accuracy of the algorithm. The more data used for testing while less data used for training will result with the less accuracy. And when the data set division for testing and training is equal then the accuracy is also in the middle of 100.

Hence, accuracy is directly proportional to the training data set means the more the training dataset for the algorithm, the more the accuracy measured between predicted outputs from algorithm and the actual outcome of the alarms.

Whereas, accuracy is inversely proportional to the amount of testing dataset for the algorithm. The more the data is put to testing that means less is used for training, the less the accuracy is achieved. The less the dataset is used for testing, the better the accuracy of the predicted outcomes of the alarms in the network communication systems.

Comparison of Decision Tree Algorithm with other

Algorithms:

	Decision Tree	Support Vector Machine	Naïve Bayes	K-Nearest Neighbor Algorithm	Artificial Neural Network Algorithm
Computational Complexity	Computational complexity is not high	Complex calculation when there are many class labels	Sensitive to pre-processing of data input	Well suited for multimodel classes	Requires high processing time if neural network is large
Easy to understand output	Output is easy to understand and to interpret(i.e using tree diagram)	Suitable when the sample size is smaller than the number of dimensions	Easy to extend to multi-class classification problem	Time to find the nearest neighbors in large training data set can be excessive	Difficult to know how many neurons and layers are necessary
Numerical and categorical data	It can easily handle numerical and categorical data	Not usually employed for continuous numerical variables, mostly for categorical variables	only for categorised data	It is sensitive to noisy or irrelevant attributes	Applicable to wide range of problems in real life
Fast, efficiency	Efficient and easy to implement	When the number of features larger than the number of samples, it is crucial to choose suitable Kernel function	Easy to implement	Performance of algorithm depends on the number of dimensions used	Learning can be slow

Figure 20.5 A comparison of different algorithms with Decision tree algorithm [32] [33]

Above comparison shows that decision tree algorithm is not only efficient and easy to implement, it is also very helpful in understanding of the output. Since the output produced by DT algorithm is visually observed in terms of the tree.

Computational complexity of decision tree is less than that of other algorithms like support vector machine, naïve bayes, neural networks. This feature is very useful for our project as our dataset attributes are not large therefore complexity level of the algorithm should be low. Output of the decision tree is easily understood as it creates a tree type structure of decision nodes and leaf nodes. All the possible outcomes can be easily visualise using some libraries in visual studio. K nearest neighbor is a very sensitive algorithm, a little noise in the data set may provide variation in results. Handling of numerical and categorical data is also easier when used with the decision tree algorithm whereas naïve bayes only works for categorical data. The fastness and efficiency level of the decision tree algorithm is also greater when compared with other algorithms.

All the decision and each step can be easily understood by just looking at the visual tree diagram of the output produced by decision tree algorithm. Hence these were the features of decision tree algorithm that we were attracted towards this algorithm for our project. The visual studio library helps in the creating a tree type structure of the output and therefore it is easier to understand.

Discussion:

We have only worked on one of the use case that predicting the alarms that are to be appearing on a specific machine at specific time. There are several use cases that this project may be extended on. For example, algorithm can be used to calculate life span of a machine so that early actions could be taken by the company to replace the machine in order to have uninterrupted communication.

Faulty cards installed at the machine could also be predicted against specific alarm, if we will provide it along with data set.

This system is helpful when forecasting and predicting of faulty devices takes place. The respective teams can be alarmed or warned by alerts through SMS or emails for future alarms of their devices so that quick actions can be taken place.

A user interface could be designed in order to provide a platform where the user can easily put an input and then the output of the faulty device could be known. This system will tell the health of the devices that is now and in future what is the condition of the network device. And after knowing the health of each of the network devices the health and quality of netrok communication can also be assumed and measures should b etakesn to improve the quality of the networks. Because eventually the better the network quality is, the better the communication devices are, the customers will be happy with the telecommunication company.

As a telecommunication company customer churn is very important, churn rate should be minimum in order to become good telecommunication company with best quality services.

Therefore, in a nutshell the use of decision tree algorithm in order to predict alarms is a good decision as it provided the predicted alarms on the specified device with full accuracy. And altogether this will help in the prospering of the company.

The dataset of alarms and device names worked fine for the algorithm as this algorithm handled this data with efficient way. And provided with results that helped in taking quick actions for the future occurrences of the alarms on the device. The predicted alarms on the respective devices are sent to the respective teams for taking immediate actions with accordance with the nature of alarm.

Conclusion:

In a nutshell, a decision tree is used to predict the alarms that are to be appeared on a specific machine and on a specific time stamp. The decision tree algorithm is able to predict the future occurrences of alarms on the network elements. We have trained the algorithm to the already occurring alarms on different network elements and tested with the names of network element. According to the training the algorithm provided with the set of alarms that are predicted to occur. Then we have calculated accuracy of the prediction made by the algorithm and the already occurred alarms and found the accuracy to be 100%.

Now the predicted values can be shared with respective teams in order to prepare for the fault that will occur in future. For the alarms with nature “outage”, the hardware or the software of the device should be observed thoroughly as these types of alarms make the device offline. When device goes offline the time that is required to bring it online again and also to rectify the device, consumes much time. This time is not beneficial if a company like telecommunication company whose sole purpose is to provide better/uninterrupted services to their customers.

The alarms that have nature of “Non-Outage” are warning alarms or the card faulty alarms. In these alarms services degradation is observed which is also not acceptable in telecommunication companies,

therefore these alarms are also required to be monitored and different type of actions is required for these type of alarms.

Future Work:

There are several use cases that this project may be extended on. For example, algorithm can be used to calculate life span of a machine so that early actions could be taken by the company to replace the machine in order to have uninterrupted communication.

Faulty cards installed at the machine could also be predicted against specific alarm, if we will provide it along with data set.

A user interface could be designed that may provide users with interactive features like search the condition of the particular devices could be monitored. Or network management systems can be made advanced by including a section of health status of the devices that are included in the network. The health status section in these systems get their data updated from these forecasting data mining algorithms and future occurrences of alarms on the particular devices may indicate the future condition of the device.

For example if the device is to go offline/down in next occurrences than it means its health is less than 100% and the notifications to the respective teams are also sent so that actions could be taken to rectify the issues.

References

- [1] P. F.-V. G. H. F. N. Z. M. Mourad Nouioua, "A Survey of Machine Learning for Network Fault Management".
- [2] D. T. a. R. C. J. B. M. Patil, "Predicting burn patient survivability using decision tree in weka environment," in *IEEE International Advance Computing Conference*, 2009.
- [3] a. K. B. P. K. Srimani, "A comparative study of different classifiers on search engine based educational data," *International Journal of Conceptions on computing and Information Technology*, vol. Vol. 2, pp. pp 6-11, 2014.
- [4] "Fault Detection Using Machine Learning Techniques," Cloud Mantra, 28 February 2020. [Online].
- [5] "Data Science vs Machine Learning vs AI vs Deep Learning vs Data Mining: Know the Differences," altexsoft, 26 Jan 2021. [Online].
- [6] "Difference Between Data Mining and Machine Learning," [Online]. Available: Difference Between.net.

- [7] P. F.-V. G. H. F. N. Z. M. Mourad Nouioua, "A Survey of Machine Learning for Network Fault Management".
- [8] M. V.-K. M. H. V. Lozonavu, "Relation discovery of mobile network alarms with sequential pattern mining. In: 2017 International Conference on Computing," in *Networking and Communications (ICNC)*, 2017.
- [9] T. Velmurugan, "A State of Art Analysis of Telecommunication Data by k-Means and k-Medoids Clustering Algorithms," 2018.
- [10] A. a. V. T. Dharmarajan, "Efficiency of k-Means and k-Medoids Clustering Algorithms Using Lung Cancer Dataset," *Int. Journal of Data Mining Techniques and Applications*, vol. 5, pp. 150-156, 2016.
- [11] J. E. Z. A. A. MouradGueroui, "Efficient k-means based clustering scheme for mobile networks cell sites management".
- [12] S. S. A. N. S. H. Yasser Khan, "Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 10, no. 9, 2019.

- [13] J. P. a. B. J. Mirjana Pejić Bach, "Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees," *Journal of Risk and Financial Management*, 2021.
- [14] I. I. T. C. Yue Chenga, "Pattern matching of alarm flood sequences by a modified Smith–Waterman," 2013.
- [15] R. M. M. A. R. A. E. S. A. a. R. S. Saad Gadal, "Machine Learning-Based Anomaly Detection Using K-Mean Array and Sequential Minimal Optimization," 2022.
- [16] Y. L. Zohreh Abtahi Foroushani, "Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique," 2018.
- [17] G. Z. R. A. A. M. B. a. R. M. Ying Zhou, "Research on data mining method of network security situation awareness based on cloud computing," 2022.
- [18] Y. Z. S. R. Y. L. X. Y. D. L. Y. H. a. J. Z. Boyuan Yan, "Dirty-data-based alarm prediction in self-optimizing large-scale optical networks," *Optics Express*, vol. 27, no. 8, pp. 10631-10643, 2019.
- [19] S. G. Ashish Prajapati, "A Survey : Data Mining and Machine Learning Methods for Cyber Security," *International Journal of Scientific Research in Computer Science*, vol. 7, no. 2, pp. 24-34, 2021.

- [20] D. B. a. J. K. M. Bhuyan, "Network anomaly detection: Methods, systems," *IEEE Communications Surveys & Tutorials*, vol. 99, pp. 1-34, 2013.
- [21] S. J. & R. F. Mahzad Mahdavisharif, "Big Data-Aware Intrusion Detection System in Communication Networks: a Deep Learning Approach," *Journal of Grid Computing volume*, 2019.
- [22] M. A. T. G. V. V. V. H. R. K. H. M. W. L. P. & V. P. Padmasiri, "Survey on Deep learning based Network Intrusion Detection and Prevention Systems," 2020.
- [23] N. C. C. Paulo Cortez, "An Intelligent Alarm Management System for Large-Scale Telecommunication Companies," in *14th Portuguese Conference on Artificial Intelligence*, Portugal, 2009.
- [24] (. e. al, "LSTM for Anomaly-Based," 2018.
- [25] H. e. al, "Using Long-Short-Term Memory Based Convolutional Neural Networks for Network Intrusion Detection," 2019.
- [26] Zabee, *PTCL MSAG Rahman Plaza University Road*.
- [27] L. Armstrong, "Decision tree diagrams: what they are and how to use them," 2021.

[28] S. studios, "Huawei Network Management System (NMS)," 2020.

[29] P. n. a. communication, "Intelligent Network Management System".

[30] HP, TeMIP Client User's Guide, 2015.

[31] S. Gunjal, "K Fold Cross Validation".

[32] R. gate, "Functional and structural biomarkers of cognitive outcomes after brain tumor resection," 2020.

[33] S. S. Nikam*, "A Comparative Study of Classification Techniques in Data Mining Algorithms," 2015.