Evolution in Electrical and Electronic Engineering Vol. 2 No. 2 (2021) 444-452

EEEE

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/eeee
e-ISSN : 2756-8458

Cable Fault Detection in DSL Communication System-based on Machine Learning

Nur Liyana Sakun[1], Zuhairiah Zainal Abidin[1]*, Fauziahanim Che Seman[1]

[1]Advanced Telecommunication Research Center (ATRC), Faculty of Electrical and Electronic Engineering,
Universiti Tun Hussein Onn Malaysia , 86400 Parit Raja, Batu Pahat, Johor, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/eeee.2021.02.02.053
Received 10 July 2021; Accepted 14 February 2021; Available online 30 October 2021

Abstract: Digital Subscriber Line (DSL) is a technology that commonly used copper cable as the transmission medium in telephone systems, and it is used widely around the world. In this modern era, the demand for high-speed internet keeps increasing to fulfill customer needs. The very-high-bit-rate Digital Subscriber Line (VDSL) is the latest DSL technology emulated in the copper network, providing internet speed up to 100Mbps. However, the copper network is still vulnerable to electromagnetic interference, which can cause degradation in the performance system to achieve a high-speed data rate. Cable faults are also common problems in copper networks like open, partially open, bridge tap, short and partial short circuits. Currently, there is no online monitoring system that able to detect the cable fault conditions accurately. Hence, this project has the best machine learning algorithm that can provide the best accuracy to identify or classify the cable fault condition compared to the ideal condition. Initially, experimentally, the emulation was conducted before deploying machine learning to evaluate the cable fault classifying accuracy. The prediction of cable condition data was simulated using WEKA Software based on few machine learning algorithms such as decision tree (J48), k-Nearest Neighbour (k-NN), Multilayer Perceptron, Naïve Bayes and Random Forest. This algorithm was tested for a cable length between 100 m, 200 m and 300 m on the line operation parameters (LOP) and loop line test (LLT) parameters. These DSL parameters were identified to indicate the overall performance of the DSL technology. The best algorithm for classifying cable fault conditions by LOP and LLT parameters is selected based on the most accurate percentage. The test results showed that the Random Forest algorithm could give consistency a higher accuracy rate with 99% and above for all cable length distances than the other algorithms.

Keywords: VDSL, Cable Fault Detection, WEKA, Machine Learning

1. Introduction

**Corresponding author: zuhairia@uthm.edu.my*

Digital Subscriber Line (DSL) is a technology which commonly used copper cable as the transmission medium in telephone systems, and it is used widely around the world. In this modern era, the demand for high-speed internet keeps increasing to fulfil the customer needs. Asymmetrical Digital Subscriber Line (ADSL) technology is the technology for wire line and can provide internet access network speeds up to 8 Mbps. Meanwhile, the Very-high-bit-rate Digital Subscriber Line (VDSL) is the latest DSL technology emulated in the copper network. The VDSL is the latest DSL technology providing internet speed up to 100 Mbps. Nowadays, technology internet speed is increasing by time as an example, in 2012 the average speed of download is about 2.47 Mbps [1], and now the average speed of download is about 24.44 Mbps [2].

The VDSL technology is realised based on the TM infrastructure as shown in Figure 1. Figure 1 shows the infrastructure consists of Central office, MSAN cabinet, distribution point and customer house. Fiber optic cable is usually used to connect the central office and the MSAN cabinet and become the infrastructure's backbone. While the copper cable will connect the MSAN cabinet to the customer's house. Although the fiber optic cable can give a higher internet access network, cable installation will be costly. Besides, the copper access network is remaining widespread network deployment, especially in the sub-urban area. This is because the 70% of the Malaysia's communication infrastructure is covered with copper transmission cables networks [3].

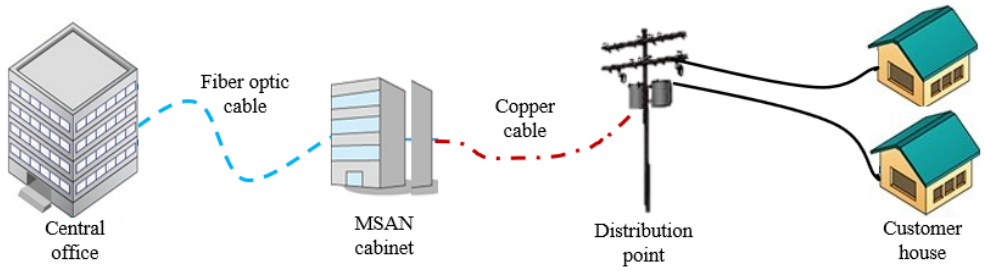


Figure 1: Telephone cabling system

However, cable fault conditions like open, partially open, bridge tap, short and partial short circuit can cause the line impairment that degradation the copper access network line performance. Besides, this cable fault is a common problem that occurred in the TM networks. Currently, there is no online monitoring system that able to detect the cable fault conditions accurately. TM workers detect the cable fault condition in real situations depending on their experience from the past issues. Therefore, based on the recent technology in machine learning, there is a possibility for classifying this type of fault based on the machine learning technique. Hence, this project will find out the best algorithm in machine learning to accurately identify or classify the cable fault condition. Experimentally, the emulation will be conducted before deployment of machine learning to evaluate the cable fault classifying accuracy. This research will solve the cable fault detection at the TM network and improve the performance of the readily available copper cables throughout Malaysia other than installing the fiber optic cable.

1.1 Machine learning on classification

Nowadays, many researchers and industries have been famous and widely used machine learning to process data. Machine learning is a subset of Artificial Intelligence with a computer program assigned to perform data mining tasks. Machine learning deals with data from the entire dataset to evaluate the predict the pattern in the data. This machine also learnt from its experience to improves the predictions performance in executing the tasks. Besides, machine learning has also been applied in a wide variety such as pattern recognition, natural language processing, traffic prediction and data mining. Data mining identifies hidden patterns and relations within the data to predict future trends using various data analysis tools. These tools are a combination of machine learning, database and statistics [4]. Many machine learning types are available to be used as open-source software, such as Orange, WEKA, KNIME and Pentaho.

Authors in [5] used WEKA to classify the data set of alcohol consumption by school students. These researchers have employed a few techniques, including Decision Stump, Random Forest, J48 and Naïve Bayes. The results show that among the algorithms, Decision Stump achieves better classification with 95.44%. Next, in [6], the authors used Orange Software to determine fruit classification types based on the image dataset. The image pattern in the dataset can be filtered using machine learning algorithms like Logistic Regression, Neural Network, Random Forest and others are employed. As a result, Logistic Regression provides the highest classification accuracy of 91% to classify the type of fruit.

While in [7] used Pentaho Software to predict and classify the intrusion by unlabeled traffic data. In [8], KNIME software is used to classify mobile reviews as positive or negative. The approaches utilised include Naïve Bayes, Random Forest, Decision Tree and Support Vector Machine. These approaches were applied on the dataset, and the experimental results show that Decision Tree and SVM give above 88% accuracy. From the analysis made by researchers, the Random tree has given the highest accuracy than other algorithms used.

1.2 WEKA

In this project, WEKA software is used to classify the copper cable faults compared with the ideal condition. All the lab data collected will be trained in machine learning. Waikato Environment for Knowledge Analysis (WEKA) is a machine learning tool for data mining tasks developed at Waikato University in New Zealand. WEKA is an open-source data mining that can be implemented to calculate data pre-processing, classification, clustering and visualisation. The entire calculation is made using a Graphical User Interface (GUI) known by an explorer to help investigate situations from information contained in the dataset. This WEKA is developed in Java language, and it supports the dataset file like ARFF, CSV, C4.5 and JSON files to extract the relevant information from the crude information [5].

2. Materials and Methods

For this work, an experimental methodology has been used. Figure 2 shows the primary process involved for this project work. The measurement laboratory was conducted to gather all the ideal and faulty cable conditions based on DSL parameters and categorisation. These parameters were identified to indicate the overall performance and referred to as information variables. All the collected raw data is then filtered from meaningless data, incomplete (missing), noisy, irrelevant and inconsistent data. The filtered data is then converted into the format used by WEKA, which is .csv. In WEKA, classification algorithms that have been chosen will be tested and analysed. Finally, all the algorithms performance will be evaluated to determine the best classifier.

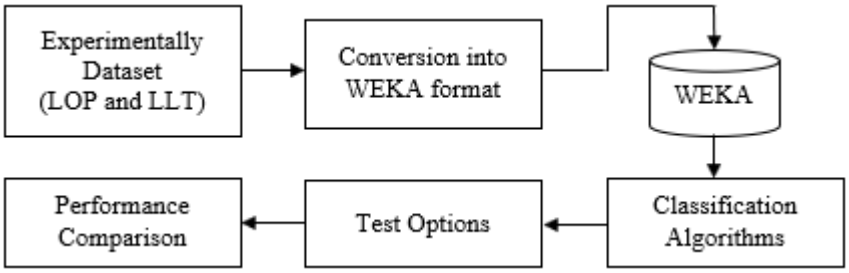


Figure 2: Main process of the project

2.1 DSL parameters and categorisation

The measurement performed in this study depends on the line operation parameters (LOP) and loop line test (LLT) parameters. These are all the DSL parameters identified to indicate the overall performance of the DSL technology. Upstream is an uploading data process from computer to the Internet, referred to as outgoing data such as sending an email or uploading a file. While downstream refers to receiving data by computer from the Internet, likes receiving email and downloading files. Table 1 shows the list of LOP and LLT parameters that will be measured in this study.

Table 1: LOP and LLT parameters

Line operation parameter	Loop line test parameter
Upload actual rate	A-B resistance (Ohm)
Download actual rate	A-G resistance (Ohm)
Maximum upload rate	B-G resistance (Ohm)
Maximum download rate	A-B capacitance (nF)
Upstream SNR margin	A-G capacitance (nF)
Downstream SNR margin	B-G capacitance (nF)
Upstream signal attenuation	A-B AC voltage (V)
Downstream signal attenuation	A-G AC voltage (V)
Upstream out power	B-G AC voltage (V)
Downstream out power	A-B DC voltage (V)
	A-G DC voltage (V)
	B-G DC voltage (V)

2.2 Laboratory setup for measurement

The laboratory setup is assembled as to imitate the existing DSL network provided by TM. Figure 3 shows the schematic of the laboratory set up in UTHM. In the existing condition, mini multi-service access node (Mini MSAN) is usually located in the street cabinet installed in a telephone exchange to connect customers telephone lines to the core network. There are two tag blocks involved in the lab setup. The modem is a hardware device that can convert data into a suitable format for an access network system. Ten (10) modems connected to the ten binder copper cables are act as the customers. The first tag block, located near the Mini MSAN, is also placed in the street cabinet. The second tag block refers to a distribution point in TM networks connected with the customers' house, and at this part, the cable fault emulation will be set up. The Ixia network and switch emulate a real traffic network with the lab setup's required parameters. This due to the real condition where the customers will generate their data traffic by accessing the internet network.

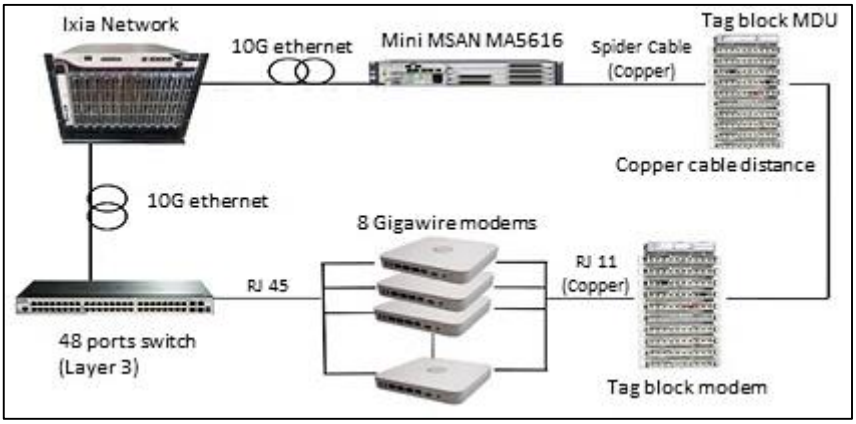


Figure 3: Schematic UTHM lab setup

2.3 WEKA Explorer

In WEKA, all the machine learning algorithms are already embedded in this software. The classification algorithms were chosen for this study's purposes are Multilayer Perceptron, Naïve Bayes, k-NN, J48 and Random Forest. All these algorithms will be used to train the collected data in WEKA. Next, for the testing set, a few test options are available in this software to test the data like use training set, supplied test set, cross-validation folds, and percentage split. In this study, the testing is conducted based on the data training set and 10-fold cross-validation. This 10-fold cross-validation is the defaults value that is suitable to be used [9].

2.4 Analysis of classification algorithm

All the algorithms performance will be evaluated to determine the best classifier according to the five statistical criteria like correctly and incorrectly instance classification, speed, Kappa statistics,

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) and accuracy measurement. The five statistical criteria are explained as follows [5]:

a) Correctly and incorrectly instance classification: The values of correctly and incorrectly predict the class label presented by percentage.
b) Speed: Time is taken to build the model.
c) Kappa Statistics: Refers to a chance-corrected measure that is calculated between classification and true classes.
d) Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE): MAE is the average magnitude of errors, while RMSE is used to measure the difference between the actual values and the estimated values of a model.
e) Accuracy measurement: The parameters involved are TP rate, FP rate, Precision, Recall and F-measure.

Comparing the classification algorithms such as J48, k-Nearest Neighbour (k-NN), Multilayer Perceptron, Naïve Bayes, and Random Forest with the five statistical criteria provide the most suitable algorithms for the dataset.

3. Results and Discussion

All the collected data will be trained with the machine learning algorithm. Machine learning is one of the artificial intelligence applications that can learn and automatically develop from experience. The yield dataset is analysed with various classification algorithms such as J48, k-Nearest Neighbour (k-NN), Multilayer Perceptron, Naïve Bayes, and Random Forest WEKA Software. All the algorithms will analyse the correctly and incorrectly instance classification, speed, Kappa statistics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and accurate measurement. There are nine hundred (900) total data fed into the machine learning, and the division for each cable condition is shown in Table 2.

Table 2: Total of data samples

Cable fault conditions	Samples of data
Ideal	150
Bridge tap	150
Short	150
Partial short	150
Partial open	150
Open	150

3.1 Correctly and Incorrectly Instance Classification

This analysis will provide the classification algorithm's ability to predict the types of cable faults expressed in percentage correctly. Figure 4 presents correctly and incorrectly classified instances among various algorithms. From the bar graph, it can be concluded that classification algorithms like Multilayer Perceptron, J48, and Random Forest have higher correct prediction with 96.22%, 98.89%, and 99.33% respectively. These three algorithms also give the lowest incorrect classified instance with 3.78% for Multilayer Perceptron, 1.11% for J48, and 0.67% for Random Forest.

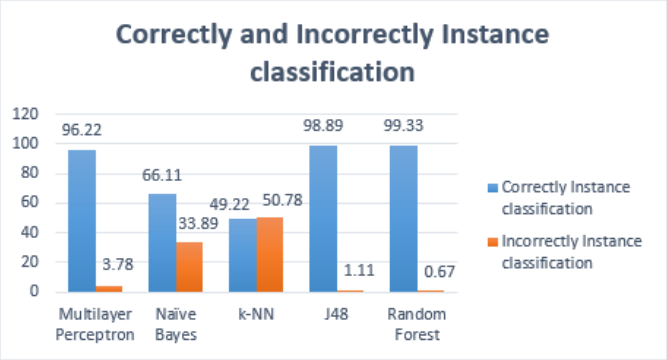


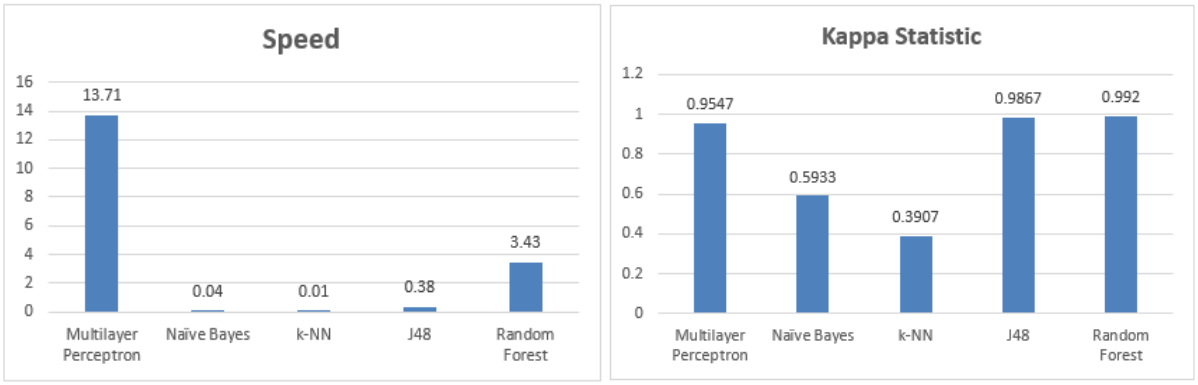
Figure 4: Bar graph of correctly and incorrectly classified instances of algorithms

3.2 Speed

The speed in this experiment is referred to the time taken to set up the model. As mentioned in [10], this speed is vital. The reasonable times needed for faults detection must below 10s. Figure 5 (a) shows that Naïve Bayes, k-NN, and J48 classification algorithms have the lowest time to train the dataset. Depending on the highest correctly classified faults from the previous section, the focus only on the Multilayer Perceptron, J48, and Random Forest. Although Random Forest has the highest accuracy, the training time to classify the cable faults is significantly higher than the J48 algorithm, which is 3.43s and 0.38s, respectively. The longest training time required to classify the data based on the types of faults is the Multilayer Perceptron algorithm, which takes 13.71s for 10-fold cross-validation.

3.3 Kappa Statistics

Kappa is referred to as a chance-corrected measure that is calculated between classification and true classes. For better performance, Kappa statistics' value must greater than zero and nearly to one, as indicates in [5]. Based on the previous analysis, Random Forest gives better performance when compared with J48 algorithms. Figure 5 (b) shows the Kappa Statistics for each of the classification algorithms.



(a) (b)

Figure 5: Bar graph of (a) speed and (b) Kappa statistics

3.4 Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

MAE is the average magnitude of errors. It is equal to the average of absolute differential values between the prediction and absolute observations. In contrast, RMSE is used to measure the difference between the actual values and the model's estimated values. The RMSE showed the standard deviation of the difference between the predicted values and the observed values. The value of RMSE is preferable to be small [11]. From the collected data implemented in WEKA software, the Random Forest algorithm performs better than J48 and others, as shown in Figure 6.

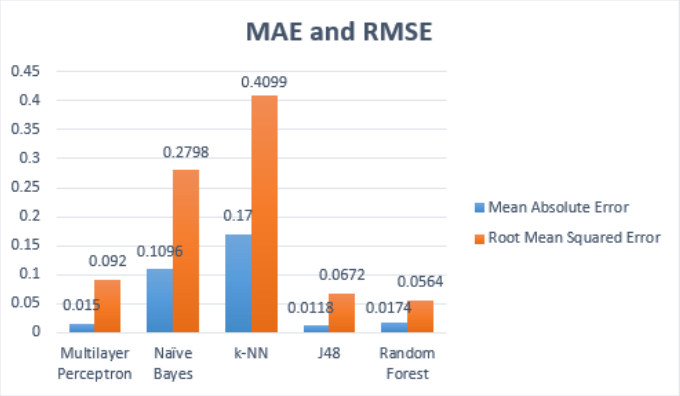


Figure 6: Comparison of MAE and RMSE

3.5 Accuracy Measurement

This section illustrates other parameters that can help with the algorithm's accuracy, such as TP rate, FP rate, precision, recall, and F-measure. These parameters are defined as tabulated in Table 3. Figure 7 shows the accuracy data for each classification algorithm. Multilayer Perceptron, J48, and Random Forest algorithms have shown a good value for the figure's TP rate. Simultaneously, the TP rate for Naïve Bayes and k-NN shows the worst value with 0.661 and 0.492. Comparing the values in F-measure according to the precision and recall, Random Forest shows the highest value with 0.993, respectively. In summary, referring to the TP rate, FP rate, Precision, Recall, and F-Measure parameters, it is clear that the Random Forest algorithm has given the highest performance.

Table 3: Definition of Parameters [13]

No	Parameters	Description
1	TP rate	Known as True Positive. It determines the data that is correctly classified respect to the type of class.
2	FP rate	Known as False Positive. It determines the data that is falsely classified respect to the type of class.
3	Precision	Ratio of correctly data classified to a certain class to total data classified concerning the class type.
4	Recall	Ratio of correctly data classified to all data present in the class (>0.5)
5	F-Measure	Determined by combining the measure of Recall and Precision. F-Measure = $2\frac{Precision*Recall}{Precision+Recall}$

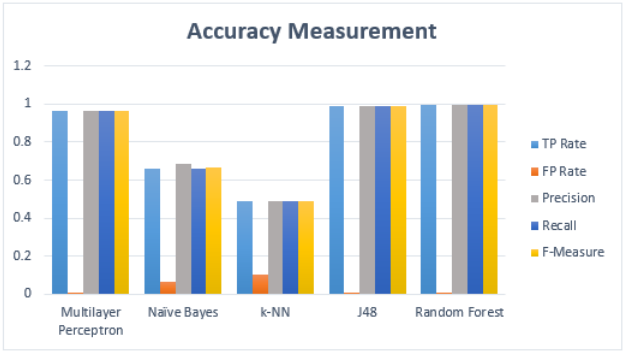


Figure 7: Graphical representation for accuracy measurement

4. Conclusion

This project investigated copper cable performance with various configurations of faults and configured the best classification algorithm to provide the highest precision in classifying the cable fault conditions. In this study, the types of copper cable conditions that have been focused on ideal, bridge tap, short, partial short, open, and partial open. These configurations have been tested for various

distances like 100 m, 200 m, and 300 m. All the LOP and LLT data is collected in *.csv format in the Excel Software and fed into the machine learning. In WEKA Software, all the selected algorithm is evaluated with several statistical criteria described in Chapter 4 to identify the best classifier algorithm. Random Forest classification algorithm performs a better classification with the highest precision, and lowest incorrectly instance classification with 99.33% and 0.67% from various machine learning used. The performance of Random Forest also satisfactory as it only needs 3.43s to build the model.

Acknowledgement

The authors would like to thank the Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia for its support.

References

[1] Malaysia Internet Speed: Average Speed Test Results: Top Cities, Devices, Trends. (2020, February 27). Retrieved October 23, 2020, from https://www.bandwidthplace.com/location/malaysia/

[2] MCMC: Malaysia maintains third spot in Asean fixed broadband ranking. (2020, September 17). Retrieved October 23, 2020, from https://www.thesundaily.my/local/mcmc-malaysia-mantains-third-spot-in-asean-fixed-broadband-ranking-GL4094019

[3] A. Asrokin, M. K. A. Rahim, A. N. Z. Abidin, N. Hashim and S. Ab Azis, "able modelling comparison for twisted-pair copper plant in malaysia," 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE), George Town, 2015, pp. 359-364, doi: 10.1109/ICCSCE.2015.7482212.

[4] S. Asha Kiranmai and A. Jaya Laxmi, "Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy," Prot. Control Mod. Power Syst., vol. 3, no. 1, p. 29, Dec. 2018.

[5] A. K. Pandey, D. S. Rajpoot and D. S. Rajpoot, "A comparative study of classification techniques by utilising WEKA," 2016 International Conference on Signal Processing and Communication (ICSC), Noida, 2016, pp. 219-224.doi: 10.1109/ICSPCom.2016.7980579

[6] D. Vaishnav and B. R. Rao, "Comparison of Machine Learning Algorithms and Fruit Classification using Orange Data Mining Tool," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2018, pp. 603-607, doi: 10.1109/ICICT43934.2018.9034442.

[7] A. Jaiswal, A. S. Manjunatha, B. R. Madhu and M. P. Chidananda, "Predicting unlabeled traffic for intrusion detection using semi-supervised machine learning," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimisation Techniques (ICEECCOT), Mysuru, 2016, pp. 218-222, doi: 10.1109/ICEECCOT.2016.7955218.

[8] C. Chauhan and S. Sehgal, "Sentiment Classification for Mobile Reviews using KNIME," 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 548-553, doi: 10.1109/GUCON.2018.8674946.

[9] A. S. Ahmadu, S. Boukari, E. J. Garba, and K. J. Danjuma, "Simulation of the Framework for Evaluating Academic Performance (FEAP) using WEKA," no. July 2019, 2017.

[10] K. Chen, C. Huang and J. He, "Fault detection, classification and location for transmission lines and distribution systems: a review on the methods," in High Voltage, vol. 1, no. 1, pp. 25-33, 4 2016, doi: 10.1049/hve.2016.0005.

[11] Comparison Begum Cigsar and Deniz Unal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk," no. February 2019, 2018