

# Cable fault classification in ADSL copper access network using machine learning



Nurul Bashirah Ghazali <sup>a,1</sup>, Dang Fillatina Hashim <sup>a,2</sup>, Fauziahanim Che Seman <sup>a,3,\*</sup>,  
Khalid Isa <sup>a,4</sup>, Khairun Nidzam Ramli <sup>a,5</sup>, Zuhairiah Zainal Abidin <sup>a,6</sup>, Saizalmursidi Md Mustam <sup>a,7</sup>,  
Mohammed Al Haek <sup>b,8</sup>

<sup>a</sup> Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Malaysia

<sup>b</sup> TM Research & Development, Cyberjaya, Malaysia

<sup>1</sup> bash.ghazali@gmail.com; <sup>2</sup> fillatina.29@gmail.com; <sup>3</sup> fauziahs@uthm.edu.my; <sup>4</sup> halid@uthm.edu.my; <sup>5</sup> khairun@uthm.edu.my;

<sup>6</sup> zuhairia@uthm.edu.my; <sup>7</sup> saizal@uthm.edu.my; <sup>8</sup> alhaek@um.edu.my

\* corresponding author

## ARTICLE INFO

### Article history

Received June 22, 2021

Revised October 8, 2021

Accepted November 11, 2021

Available online November 30, 2021

### Keywords

Digital subscriber line

Copper access network

WEKA

Classification

Machine learning

## ABSTRACT

Asymmetrical Digital Subscriber Line (ADSL) is the technology widely deployed worldwide, but its performance may be limited with respect to its intrinsic. The nature of the copper cable causes it to be more susceptible to signal degradation and faulty line. Common ADSL line faults are short-wired, open-wired, bridge taps, and uneven pair. However, ADSL technology is still one of the most established networks, and users in the suburban area still depend on the technology to access the internet service. This paper discussed and compared a machine learning algorithm based on Decision Trees (J48), K-Nearest Neighbor, Multi-level Perceptron, Naïve Bayes, Random Forest, and Sequential Minimal Optimization (SMO) for ADSL line impairment that affects the line operation performance concerning their percentage of accuracy. Resulting from classifications done using algorithms as mentioned above, the random forest algorithm gives the highest overall accuracy for the ADSL line impairment dataset. The best algorithm for classifying DSL line impairment is chosen based on the highest accuracy percentage. The accomplishment classification of fault type in the ADSL copper access network project may benefit the telecommunication network provider by remotely assessing the network condition rather than on-site.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## 1. Introduction

One telecommunication technology that persists in being the utmost popular is the digital subscriber line (DSL) [1][2]. Asymmetrical Digital Subscriber Line (ADSL) is one of the earliest DSL technologies that enable internet services that utilize the existing twisted-pair telephone subscriber loop infrastructure [3][4]. Typical ADSL copper access network infrastructure consists of central office (HQ), Multi-Service Access Network (MSAN) cabinet, distribution point (DP), and customer premises [5]-[7]. In the ADSL copper access networks, fiber cables become the backbone infrastructure that connects the central office and MSAN. The copper cables are laid between MSAN and the customer premises through the distribution point (DP). The ADSL network cards and Plain Old Telephone Service Line (POTS) cards are integrated at MSAN, and this is the important key factor that allows both telephone and internet services to be offered using the same copper cables infrastructure [8]. This network topology is usually established at the suburban site where the capital expenditure of the telco needs to reflect the customer's populations in that particular area. This ADSL network configuration can support up to 8.0 Mbps data

rates for distances up to 5km [9]. However, the offered data rates are still subjected to the copper cables' nature where the attenuation degrades rapidly with distances and the presence of external electromagnetic interferences, leading to crosstalk. Furthermore, the most common copper cable faults that occurred at cable jointing are the imbalance impedance line. Note that the effective impedance matching of copper cables is 100 Ohm. Any values above or below will lead to open or short cable faults. Also, bridged taps that the cable fault contributed when an open-circuited twisted pair is connected in shunt with the twisted working pairs. Uneven length occurs when the length of one twisted pair is not the same. All the cable faults described here will affect the ADSL network line parameters: line operation and loop line test. These parameters will indicate the conditions of the copper access networks. The line operation attributes that are only available when the subscriber line accesses the internet service are the speed rate, attenuation, and signal to noise ratio (SNR). In comparison, the loop line test attributes are used to test the electrical indicators of the line from MSAN to the customer house. When the customers' POTS services are faulty, the loop line tests can be performed to identify and localize the copper cable faults [10]. Based on the line operations and loop line test parameters, the ADSL network performance can be observed to indicate possible cable faults that may occur along the subscriber loop line. The copper cable fault classifications may be deployed by using a machine learning algorithm [11]. To date, there are not many papers emphasizing the usage of machine learning for cable fault classification in the ADSL copper access network.

The feasibility of classification using machine learning algorithms is widely demonstrated for other applications such as identifying objects or data, determining organisms, gender classifications, and transmission line faults. Classification categorizes data instances to their specified class based on learned training data [12][13] to forecast targeted output. Several popular machine learning classifications algorithms have been used to categorize and classify various applications, such as Decision trees, Naïve Bayes, Artificial Neural Network (ANN), Random Forest, and k-Nearest Neighbor (k-NN) [14].

Multi-Layer Perceptron Neural Network (MLP), Bayes, and Naïve Bayes machine learning algorithm were deployed to distinguish the type of fault that exists in shunt compensated static synchronous compensator (STATCOM) transmission line by applying a Discrete Wavelet Transform (DWT) as feature extraction method and Naive Bayes (NB) algorithm as classification method [15]. Several types of faults are classified based on the changes in fault resistance. Faults that are applied are LG, LL, DLG and LLLG. The feature extraction using DWT results in the acquisition of the SD and Energy values have been various faults with the resistance of fault 0.001  $\Omega$ . The acquired features are then adopted for the training of the transmission line fault type classifiers. Result acquired from the classifier indicates the excellent performance of the NB classifier for the case of with and without STATCOM, which is 100% accuracy compared to MLP and Bayes, which gives an average of 80% and 20% of accuracy, respectively.

Fault classifications was used to detect normal and abnormal sensed data of Wireless Sensor Network (WSN) using Support Vector Machine (SVM), Convolutional Neural Network (CNN), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Random Forest (RF), and Probabilistic Neural Network (PNN) [16]. There are 40 datasets used in this paper, alongside 9566 instances and 12 dimensions each. The dataset has a column for distinguishing normal and abnormal instances, which is 1 and -1, respectively. The faults that are induced in this paper are gain, offset, stuck-at, out of bounds, spike, and data loss fault. The classifier's accuracy is evaluated using Detection Accuracy (DA), true positive rate (TPR), Matthews Correlation Coefficient (MCC), and F1-score. These accuracy metrics indicate that RF's performance is better than another classifier, which means RF can accurately classify faults as listed above in WSN.

The multilayer perceptron is one of the ANN methods that was often used for classifications. It is a feed-forward network, and it comprises several layers: the input, hidden, and output layers [17]. Multilayer Perceptron Neural Network classifier was used in Adnan *et al.* [18] to detect heart abnormality using P, Q, R, S, and T amplitude of ECG data. The learning algorithm that is used for the training of these data are Back Propagation (BP), Bayesian Regularization (BR), and Levenberg-Marquardt (LM).

The high accuracy of all learning algorithm, which is more than 90%, indicates the capability and reliability of the MLP model to detect heart abnormality. Despite the high accuracy, the BR gives the lowest Mean Square Error (MSE), which is 0.237, compared to other learning algorithms that give MSE values of 4.667 and 4.067 for BP and LM learning algorithms, respectively. The low MSE value for BR means it can give better predictions than other learning algorithms.

Different from multilayer perceptron that used several layers of network nodes to process and classify the data samples, the decision tree classifies data points using tree representations. The decision tree aims to build a model for forecasting the value of a target output by giving a considerable amount of inputs [19]. Each leaf node signifies a test on a feature, each branch denotes a test result, and each leaf node represents a class label. Decision tree's disadvantage is that their training process may lead to overfitting [20]. Several works are done using machine learning algorithms to classify transmission line fault. As demonstrated in Chandra *et al.* [21], a decision tree algorithm was applied for classifications of fault in the transmission line. Using 4262 fault data samples, data are split into 70% training data and 30% testing data. The training and testing of the fault data samples result in 81.23% of accuracy. Thus, it has been concluded that the decision tree algorithm was easy to use and implement for classification purposes.

The closest application of machine learning to the ADSL network is described in Akhikpemelo *et al.* [22]. The fault detection of 132kV transmission line system is done using the Levenberg-Marquardt learning algorithm in Artificial Neural Network (ANN). The data are generated and extracted using MATLAB software. The detection output is represented as 1 or 0 to indicate the presence of a fault and no-fault. The network configuration setting for ANN (15-15-10-5) gives excellent classification results, with 0.99953 correlation coefficient and 0.000145 overall MSE.

Based on the described machine learning algorithm in the paper, the copper cable fault classification is performed by using Naïve Bayes, Random Forest, Multilayer Perceptron, k-NN, and Decision Tree. The purpose of this preliminary work is to provide the feasibility study of the algorithm that can classify cable faults that may occur in the ADSL copper access network. All the algorithms are readily embedded in WEKA, a tool that can provide visualization and classification for many applications [23]-[25]. WEKA is used as the initial step of exploration in classifying ADSL copper access network fault types. The intention of this paper is to get insights on which algorithm that might benefit more from this application. The accuracy of the classification algorithms was achieved up to 97% [23]. These algorithms are applied to cross-validated ADSL network data, where the copper cable faults focus on bridge tap, open, short, and uneven. The rest of the paper is organized as follows: Section 2 explains the proposed method of classifying cable fault types. Section 3 discusses the finding and evaluates the accuracy of all algorithms. Section 4 concludes the findings and the future direction of this work.

## 2. Method

The cable fault classification in the ADSL copper access networks is realized in four stages, as shown in Fig. 1. The first stage involves lab data acquisition. The laboratory setup is established to imitate the actual ADSL copper access network. The specific ADSL network data representing the ideal cable condition and the faulty network are gathered. Only four types of cable faults are emulated in the laboratory measurements, which are open, short, bridge tap, uneven and short. The dataset acquired in the laboratory has been compared with National Telecommunication company dataset, which consists of the same parameters to make sure the benchmarking of the ADSL access network dataset is met and in range. The collected data is fed into the second stage which is called data preprocessing. The raw collected data are prepared in this stage, and only highly correlated parameters and attributes are extracted as the data training set. The data preprocessing of the acquired data also handles any invalid or missing data that may affect the accuracy of the classification in WEKA. The third stage establishes the actual cable fault classification in which the preprocessed data set has been used to train the machine learning algorithm in WEKA. This involves five types of algorithms which are J48, k-NN, Multilayer

Perceptron, Naïve Bayes and Random Forest. The cable fault classification algorithm's accuracy in WEKA is evaluated in the fourth stage.

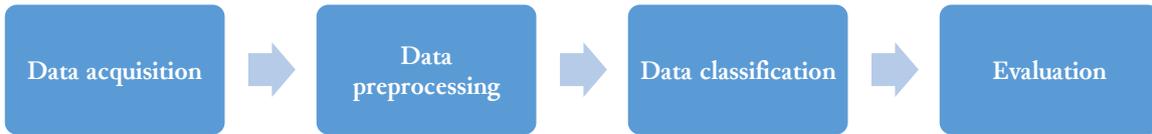


Fig. 1. ADSL copper access network impairment classification process

**2.1. Laboratory Setup and Data Acquisition**

The laboratory setup is developed to imitate the actual ADSL copper access network. Fig. 2 shows the laboratory setup of network data acquisitions, consisting of MSAN, copper cables with various lengths, and modem. The raw lab data are gathered from MSAN using the command prompt Telnet script. The raw lab data are stored in the form of a text file, which is then manually processed and converted into an excel file for ease of use. The UTP Cat 3 twisted copper cables are used to connect MSAN and Modem through tag block. The ten binder copper cable is connected to 10 modems which represents the network line termination at the customer premises. The possible presence of cable faults in the ADSL networks is emulated on the tag block.

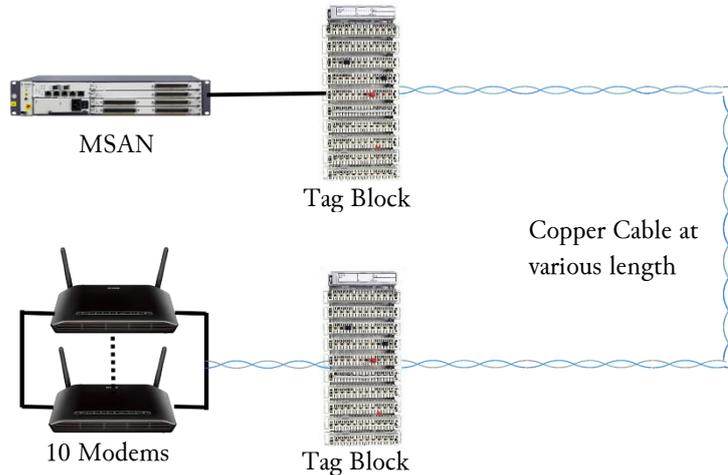


Fig. 2. Lab Setup for ADSL Access Network Data Acquisitions

Fig. 3 shows the cross-sectional area of the UTP Cat3 bundles of 10 twisted pairs. There are ten twisted pairs, and each pair consists of Tip (A) and Ring (B).

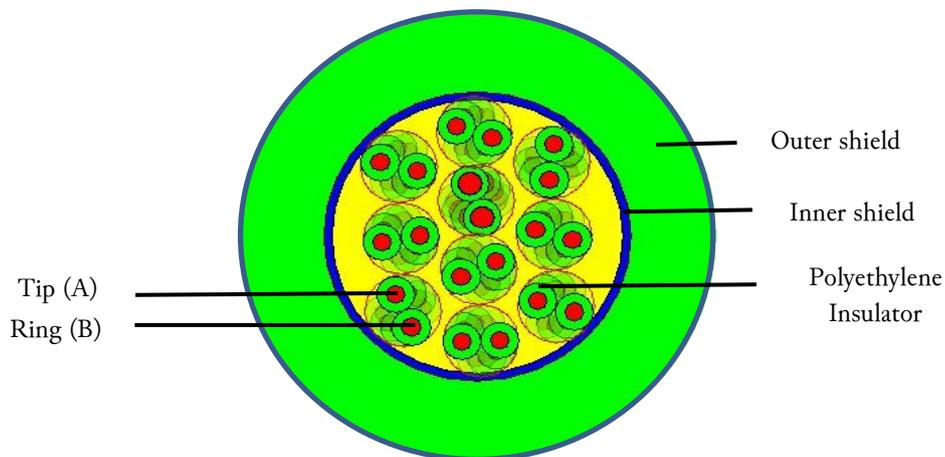


Fig. 3. Cross-sectional area of 10 pairs of twisted pair copper cable in a single copper cable

The UTP Cat3 cable can support voice and data transmission at 8Mbps data speeds. Each twisted pair can be modeled as the RC lumped component model [26], as illustrated in Fig. 4. Each A and B component has the same component of RC but possibly with different values when the cable fault occurs along the line [27]. Each A and B is connected to the Ground (G), which are represented by component Resistances,  $R_{A-G}$  and  $R_{B-G}$ ; Capacitances,  $C_{A-G}$  and  $C_{B-G}$ ; and Voltage,  $V_{A-G}$ , and  $V_{A-G}$ . Even though the other components are still available, they were neglected in the electrical schematic due to insignificant changes when different cable faults are emulated. In ideal cases, the effective impedance,  $R_s$  is 100 Ohm, and obviously, cable faults such as imbalance impedance, uneven length, and bridge tap can be assessed from Components  $R_{A-B}$ , hence also affecting the capacitance and voltage components.

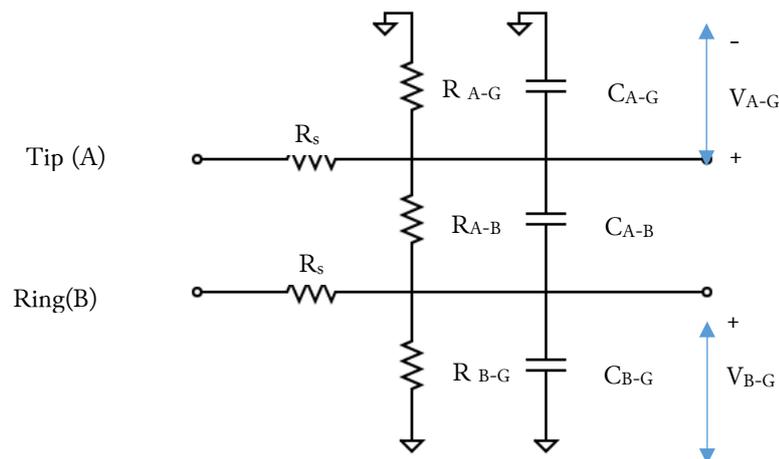


Fig. 4. Twisted pair, Tip, Ring and Ground component using RLC lumped component elements.

As highlighted previously, the ideal cable conditions and cable faults such as bridge tap and imbalance line emulation are conducted on the tag block side. The copper cables length (between the two tag block) varies at various lengths from minimum 20 m to maximum 4000 m. These cable lengths are chosen to represent the actual minimum and maximum distance possible between MSAN and the customer premises in ADSL network topology. In these cases, if the location of the customer premises are relatively far from MSAN, the attenuation and signal-to-noise ratio (SNR) will be higher, but the achievable data rates will be lower.

The emulation of different cable faults is conducted by modifying the RLC lumped component. Emulation open wired cable fault, the emulation are conducted by disconnecting A and B. Open circuit fault type usually causes a slight increase of capacitance value between tip and ring, while decreasing the value of resistance between tip and ring of the line. Emulation of short-wired impairment is performed by connecting its twisted pair A and B directly to each other. Short circuit fault causes the capacitance and resistance between the tip and ring of the line to be extremely low. The bridge tap impairment type is emulated by adding another copper cable connected in parallel to existing A and B at the tag block near the modem. The bridge tap causes a slight drop in line performance due to the increase of signal attenuation and decreasing of attainable rate of the line, while the resistance and capacitance stay in a good range of value.

In reality, the bridge tap fault occurs when the customer did unnecessary additional telephone wiring. Uneven line fault occurs when the length of A and B is not the same. Therefore this fault is emulated by increasing the length of either A or B by 100 m. Uneven fault causes an increment in capacitance value, while other line performance parameters and resistance stay in a good range of value. The acquired lab data may contain missing data and may interfere with the classification process in WEKA. Therefore, these raw data will be preprocessed, and details are explained in the next section.

## 2.2. Data preprocessing

WEKA are equipped with the utilization of learning algorithms that can be easily used to a dataset. It also provides a variety of tools for transforming datasets, such as the algorithms for discretization and sampling. Other than that, WEKA allows data to be preprocessed before feeding it to a learning algorithm and evaluates the classifier results and their efficiency without having to write any program code. The workbench also provides regression, classification, clustering, association rule mining, and attribute selection methods for solving data mining problems. In addition, many data visualization facilities and data preprocessing tools are provided for the purpose of data exploration. Since the irrelevant values such as symbols or missing values in the dataset may interfere with the classification process in WEKA software, the acquired lab data will need to undergo preprocessing stage, which enables the data to be transformed into a suitable format for machine learning algorithms. The data that contains symbols are manually removed since its presence in data might cause certain algorithms in WEKA classified cannot be used. Other than handling irrelevant data, primary parameters that are suspected to be correlated with the types of fault that occurred in the ADSL copper access network are selected, while others are removed from the dataset. After undergoing the preprocessing stage, 420 samples of ADSL lab data are obtained. The line operation parameters and loop line test parameters which are represented by 26 attributes, are gathered as tabulated in Table 1. The loop line test attributes are extracted based on the RLC lumped components which are previously discussed in Figure 4. The cable faults classifications are initiated by labeling all related cable fault types. Table 2 shows the divisions of data samples for each class of ideal conditions of 4 types of cable faults.

**Table 1.** ADSL lab emulated line performance data attributes

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

**Table 2.** Samples of cable impairment types

Cable impairment types	Samples
Ideal	169
Open	40
Short	12
Bridge tap	152
Uneven	47

After undergoing data preprocessing to handle irrelevant data, the cleansed data are used for classification. The classifier or algorithms are chosen to classify ADSL copper access network fault type. Table 3 shows the parameter setting of each of the algorithms used.

**Table 3.** Parameter setting of classification algorithms.

Algorithm	Parameter
1 Naïve Bayes	Batch size = 100, debug = false, display model in old format = false, do not check capabilities = false, number of decimal places = 2, use kernel estimator = false, use supervised discretization = false.
2 Multi-Layer Perceptron	GUI = false, auto build = true, batch size = 100, debug = false, decay = false, do not check capabilities = false, hidden layers = a, learning rate = 0.3, momentum = 0.2, nominal to binary filter = true, normalize attributes = true, normalize numeric class = true, number of decimal places = 2, reset = false, resume = false, seed = 0, training time = 500, validation set size = 0, validation threshold = 20.
3 K-nearest neighbors	K = 20, batch size = 100, cross validate = false, debug = false, distance weighting = no distance weighting, do not check capabilities = false, mean squared = false, nearest neighbour search algorithm = LinearNNsearch, number of decimal places = 2, window size = 0
4 Random Forest	Bag size percent = 100, batch size = 100, break ties , randomly = false, calculate out of bag = false, compute attribute importance = false, debug = false, do not check capabilities = false, maximum depth = 0, number decimal places = 2, number execution slots = 1, number features = 0, number iterations = 100, output out of bag complexity statistics = false, print classifiers = false, seed = 1, store out of bag predictions = False
5 J48	Batch size = 100, binary splits = false, collapse tree = true, confidence factor = 0.25, debug = false, do not check capabilities = false, do not make split point, actual value = false, minimum number object = 2, number decimal places = 2, number folds = 3, reduced error pruning = false, save instance data = false, seed = 1, sub tree raising = true, unpruned = false, use Laplace = false, use mdl correction = true

### 3. Result and Discussion

Table 4 depicts the classification performance of ADSL data based on K-fold cross-validation method. In WEKA software, by default, a 10-fold setting for cross-validation is chosen [28]. The percentage accuracy of each ADSL line fault type is made by comparing the classified instances from the total samples of each tested case. The performance of the algorithms are evaluated based on total accuracy percentage, which is the percentage of correctly classified instances. All samples are used for classification and are evaluated using the cross-validation method, in which the dataset is resampled into 'k' samples. The evaluation of every sample will then result in the average accuracy of the dataset.

The classification of ADSL line impairment data done using J48 algorithms results from the highest accuracy percentage of 91.67% for short-wired impairment type. As for k-NN algorithms, the line impairments that produce high percentage accuracy is open wired line impairment, which gives out 72.50% of accuracy. It can be observed that multilayer perceptron and random forest algorithm bring about 85.11% and 95.74% of accuracy. The accuracy of uneven line impairment classification using random forest is noticeably higher than multilayer perceptron. Considering the overall accuracy of these algorithms, the random forest classifier gives the highest accuracy, 71.67%, with 301 correctly classified samples. The algorithm with the lowest accuracy is the k-NN algorithm, with only 51.43% accuracy and 216 correctly classified samples.

Even though the random forest has the highest accuracy, the training time of the cable impairment classifications is slightly higher than the k-NN classification algorithm, which is 0.34s and 0.00s, respectively. The multilayer perceptron algorithm takes the longest training time to classify the data based on its impairment type, which takes 3.74s to train 10-fold cross-validation. Training of MLP can be a bit of a gradual process [29] because the number of training epochs and time elapsed for total epochs indicates when the training was completed [30].

Table 4. WEKA classification result

Algorithm	Total samples	Fault type	Training Time (s)	Correct Classification		Total accuracy percentage
				No. of Samples	Accuracy (%)	
J48	169	Ideal	0.02	109	64.50	69.52
	152	Bridge Tap		98	64.47	
	40	Open		33	82.50	
	47	Uneven		41	87.23	
	12	Short		11	91.67	
k-NN	169	Ideal	0	74	43.79	51.43
	152	Bridge Tap		75	49.34	
	40	Open		29	72.50	
	47	Uneven		32	68.09	
	12	Short		6	50.00	
Multilayer Perceptron	169	Ideal	3.74	102	60.36	63.33
	152	Bridge Tap		84	55.26	
	40	Open		31	77.50	
	47	Uneven		40	85.11	
	12	Short		9	75.00	
Naïve Bayes	169	Ideal	0.01	74	43.79	61.43
	152	Bridge Tap		116	76.32	
	40	Open		21	52.50	
	47	Uneven		38	80.85	
	12	Short		9	75.00	
Random Forest	169	Ideal	0.34	111	65.68	71.67
	152	Bridge Tap		100	65.79	
	40	Open		35	87.50	
	47	Uneven		45	95.74	
	12	Short		10	83.33	

The result of classifications done using WEKA in Table 4 indicates that's random forest algorithm gives a promising result for the application of classifying fault types in ADSL copper access networks. Due to a small number of short fault type samples, the J48 decision tree algorithm can provide a high percentage of accuracy [31]. The small number of instances increases the reliability of the algorithms. These are due to the algorithm characteristics that provide imputation technique, making it suitable for short wired line impairment parameters.

KNN is an algorithm, which the classifications are achieved by calculating the closest distance between data attributes [32]. Thus, given that most of the values of the attributes between each class are rather close to each other may cause the classification using this algorithm results in a bit low accuracy compared to J48 algorithms. The high accuracy resulting from the random forest method may be due to its ability to handle an abundant number of parameters due to the existing feature selection in the model development process [33]. The random forest algorithms can be used with many variables and do not require complex parameter tuning to achieve high classification accuracy.

Ensemble classifier is a method that combines the predictions of several classifiers into one classifier model. Several weak classifier models are trained and combined using either the voting or averaging method. Ensemble a classifier such as random forest that uses the bagging method is best for this dataset due to its ability to consider small changes produced by weak learners to its decision [34][35]. The aspect that may contribute to the low accuracy of the k-NN is the disposing of new data classes that is based on a plain majority voting rule, where the majority voting rule may overlook the adjacency of the data, this is intolerable when the distance of every nearest neighbor varies greatly against the distance of the test data [36]. Another critical factor that may affect its accuracy is the value of  $k$  chosen for k-NN classification. Choosing a high value of  $k$  may result in misclassification of a new point, while choosing a small value of  $k$  may lead to overfitting [37].

The nearest neighbor search algorithm is used by kNN in WEKA is linear NN search or brute force method. The Brute force method calculates a new point's distance to another point in training data, sorts out the distances, and takes on the k nearest for a majority vote. Thus, it does not need other training processes, just deal with prediction complexity, resulting in 0 seconds of the training process for kNN [38].

## 5. Conclusion

This paper presents the copper cable faults classification with WEKA by using several machine-learning algorithms, namely, Decision tree, k-NN, Multilayer Perceptron, Random Forest, Naïve Bayes, and SMO. The results showed that the random forest algorithm provides the highest accuracy of about 71% and the shortest training time. This is followed by the Decision Tree (J48). Based on the existing literature, the accuracy is considerably low because the accuracy of the machine learning classification is normally higher than 80%. This may be due to an imbalanced distribution of lab data for each tested class, and further data preprocessing to tackle specific ADSL lab data is required. Considering this matter, the accuracy of the cable fault detection should be improved based on random forest and J48 by developing a machine learning algorithm using Python. In the future, the algorithms that show promising accuracy may be explored more using Python Programming, which consists of a variety of libraries suitable for machine learning development and may allow flexibility of development of certain algorithms. New ADSL and VDSL data may be acquired to test the reliability of the classification model chosen by WEKA and further explored by Python. This may give an insight into the efficiency of a chosen algorithm in determining fault types for different copper access network technology. This research on copper access networks might benefit telecommunication sectors that still have customers that use copper access networks and may only need remote line condition assessment rather than on-site troubleshooting.

## Declarations

**Author contribution.** All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

**Funding statement.** None of the authors have received any funding or grants from any institution or funding body for the research.

**Conflict of interest.** The authors declare no conflict of interest.

**Additional information.** No additional information is available for this paper.

## References

- [1] T. Hendrickx, B. Cule, P. Meysman, S. Naulaerts, K. Laukens, and B. Goethals, "Mining Association Rules in Graphs Based on Frequent Cohesive Itemsets," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9078, no. 3, 2015, pp. 637–648. doi: [10.1007/978-3-319-18032-8\\_50](https://doi.org/10.1007/978-3-319-18032-8_50).
- [2] T. Bai, H. Zhang, J. Zhang, C. Xu, A. F. Al Rawi, and L. Hanzo, "Impulsive Noise Mitigation in Digital Subscriber Lines: The State-of-the-Art and Research Opportunities," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 145–151, May 2019. doi: [10.1109/MCOM.2019.1800858](https://doi.org/10.1109/MCOM.2019.1800858).
- [3] T. Bai, H. Zhang, J. Zhang, C. Xu, A. F. Al Rawi, and L. Hanzo, "Impulsive Noise Mitigation in Digital Subscriber Lines: The State-of-the-Art and Research Opportunities," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 145–151, May 2019. doi: [10.1109/MCOM.2019.1800858](https://doi.org/10.1109/MCOM.2019.1800858).
- [4] G. D. Ranasinghe, D. Yearling, M. Girolami, and A. K. Parlikad, "Using Expert Knowledge to Generate Data for Broadband Line Prognostics Under Limited Failure Data Availability," *IFAC-PapersOnLine*, vol. 53, no. 3, pp. 265–270, 2020. doi: [10.1016/j.ifacol.2020.11.043](https://doi.org/10.1016/j.ifacol.2020.11.043).
- [5] D. Heath, "Chart of ADSL and ADSL2+ Speed Versus Distance | Increase Broadband Speed", Increase Broadband Speed | Tips to improve and speed up your broadband, 2019. [Online]. Available: <https://www.increasebroadbandspeed.co.uk/2012/graph-ADSL-speed-versus-distance>. [Accessed: 02- Oct-2021].

- [6] Z. Abdellaoui, Y. Dieudonne, and A. Aleya, "Design, implementation and evaluation of a Fiber To The Home (FTTH) access network based on a Giga Passive Optical Network GPON," *Array*, vol. 10, no. March, p. 100058, Jul. 2021. doi: [10.1016/j.array.2021.100058](https://doi.org/10.1016/j.array.2021.100058).
- [7] S. Neeraj, K. Sanjay, S. Manoj, and S. Atulkumar, "Introduction of Fiber To The Home Technology," no. March, 2019. doi: [10.13140/RG.2.2.23634.20169](https://doi.org/10.13140/RG.2.2.23634.20169).
- [8] K. Featherly, "Digital Subscriber Line," in *Encyclopedia of New Media*, vol. 1, no. 10, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2012, pp. 190–194.
- [9] G.992.1 : Asymmetric digital subscriber line (ADSL) transceivers, E 17821, International Telecommunication Union (ITU), July 1999. [Online]. Available: <https://www.itu.int/rec/T-REC-G.992.1-199907-1/en>.
- [10] "Access Network Products Troubleshooting Guide (pdf) 01," Huawei, 31-Oct-2018. [Online]. Available: <https://support.huawei.com/enterprise/my/doc/EDOC1100047614/62a34014/pots-user-loop-line-test>. [Accessed: 12-Feb-2020].
- [11] M. S. Coutinho *et al.*, "Machine learning-based system for fault detection on anchor rods of cable-stayed power transmission towers," *Electr. Power Syst. Res.*, vol. 194, p. 107106, 2021. doi: [10.1016/j.epsr.2021.107106](https://doi.org/10.1016/j.epsr.2021.107106).
- [12] A. Soofi and A. Awan, "Classification Techniques in Machine Learning: Applications and Issues," *J. Basic Appl. Sci.*, vol. 13, pp. 459–465, Aug. 2017. doi: [10.6000/1927-5129.2017.13.76](https://doi.org/10.6000/1927-5129.2017.13.76).
- [13] K. Hameed, D. Chai, and A. Rassau, "Class distribution-aware adaptive margins and cluster embedding for classification of fruit and vegetables at supermarket self-checkouts," *Neurocomputing*, vol. 461, pp. 292–309, 2021. doi: [10.1016/j.neucom.2021.07.040](https://doi.org/10.1016/j.neucom.2021.07.040).
- [14] Neelakantan . P, "Analyzing the best machine learning algorithm for plant disease classification," *Mater. Today Proc.*, 2021. doi: [10.1016/j.matpr.2021.07.358](https://doi.org/10.1016/j.matpr.2021.07.358).
- [15] E. Aker, M. L. Othman, V. Veerasamy, I. bin Aris, N. I. A. Wahab, and H. Hizam, "Fault Detection and Classification of Shunt Compensated Transmission Line Using Discrete Wavelet Transform and Naive Bayes Classifier," *Energies*, vol. 13, no. 1, p. 243, Jan. 2020. doi: [10.3390/en13010243](https://doi.org/10.3390/en13010243).
- [16] Z. Noshad *et al.*, "Fault Detection in Wireless Sensor Networks through the Random Forest Classifier," *Sensors*, vol. 19, no. 7, p. 1568, Apr. 2019. doi: [10.3390/s19071568](https://doi.org/10.3390/s19071568).
- [17] I. N. Yulita, R. Rosadi, S. Purwani, and M. Suryani, "Multi-Layer Perceptron for Sleep Stage Classification," *J. Phys. Conf. Ser.*, vol. 1028, no. 1, p. 012212, Jun. 2018. doi: [10.1088/1742-6596/1028/1/012212](https://doi.org/10.1088/1742-6596/1028/1/012212).
- [18] J. Adnan *et al.*, "Heart abnormality activity detection using multilayer perceptron (MLP) network," in *AIP Conference Proceedings*, 2018, vol. 2016, no. September, p. 020013. doi: [10.1063/1.5055415](https://doi.org/10.1063/1.5055415).
- [19] M. A. Hafeez, M. Rashid, H. Tariq, Z. U. Abideen, S. S. Alotaibi, and M. H. Sinky, "Performance Improvement of Decision Tree: A Robust Classifier Using Tabu Search Algorithm," *Appl. Sci.*, vol. 11, no. 15, p. 6728, Jul. 2021. doi: [10.3390/app11156728](https://doi.org/10.3390/app11156728).
- [20] A. M. Ahmed, A. Rizaner, and A. H. Ulusoy, "A novel decision tree classification based on post-pruning with Bayes minimum risk," *PLoS One*, vol. 13, no. 4, p. e0194168, Apr. 2018. doi: [10.1371/journal.pone.0194168](https://doi.org/10.1371/journal.pone.0194168).
- [21] A. Chandra, A. Pani, P. K. Sahu, B. B. Majumdar, and S. Sharma, "Identifying large freight traffic generators and investigating the impacts on travel pattern: A decision tree approach for last-mile delivery management," *Res. Transp. Bus. Manag.*, p. 100695, 2021. doi: [10.1016/j.rtbm.2021.100695](https://doi.org/10.1016/j.rtbm.2021.100695).
- [22] A. Akhikpemelo, J. E. Evbogbai, and M. S. Okundamiya, "Fault Detection on a 132kV Transmission Line Using Artificial Neural Network," *Int. Rev. Electr. Eng.*, vol. 14, no. June, pp. 1–6, 2019. doi: [10.15866/iree.v14i3.17054](https://doi.org/10.15866/iree.v14i3.17054).
- [23] J. Abbineni and O. Thalluri, "Software Defect Detection Using Machine Learning Techniques," *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 471–475. doi: [10.1109/ICOEI.2018.8553830](https://doi.org/10.1109/ICOEI.2018.8553830).

- [24] H. Talabani and E. Avci, "Performance Comparison of SVM Kernel Types on Child Autism Disease Database," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, pp. 1-5. doi: [10.1109/IDAP.2018.8620924](https://doi.org/10.1109/IDAP.2018.8620924).
- [25] D. Merlini and M. Rossini, "Text categorization with WEKA: A survey," *Mach. Learn. with Appl.*, vol. 4, p. 100033, 2021. doi: <https://doi.org/10.1016/j.mlwa.2021.100033>.
- [26] Valencia, R., 1999. Outside Plant Telephone Cable Testing & Fault Locating. [online] Multimedia.3m.com. Available at: <http://multimedia.3m.com/mws/media/2827280/outside-plant-telephone-cable-testing-fault-locating.pdf> [Accessed 12 March 2020].
- [27] Rousselot, E. (n.d.). Guide to the Last Mile A Benchmark for VDSL Testing. [ebook] Pennsylvania: Megger Group Limited, p.1. Available at: <https://cdn.thomasnet.com/ccp/00145012/54340.pdf> [Accessed 4 Mar. 2020].
- [28] P. Feng and L. Feng, "Sequence based prediction of pattern recognition receptors by using feature selection technique," *Int. J. Biol. Macromol.*, vol. 162, pp. 931-934, 2020. doi: [10.1016/j.ijbiomac.2020.06.234](https://doi.org/10.1016/j.ijbiomac.2020.06.234).
- [29] D. Li, F. Huang, L. Yan, Z. Cao, J. Chen, and Z. Ye, "Landslide susceptibility prediction using particle-swarm-optimized multilayer perceptron: Comparisons with multilayer-perceptron-only, bp neural network, and information value models," *Appl. Sci.*, vol. 9, no. 18, p. 3664, 2019. doi: [0.3390/app9183664](https://doi.org/10.3390/app9183664).
- [30] H. Alla, L. Moumoun, and Y. Balouki, "A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction," *Sci. Program.*, vol. 2021, pp. 1-12, Jun. 2021. doi: [10.1155/2021/5558918](https://doi.org/10.1155/2021/5558918).
- [31] Ö. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1-5. doi: [10.1109/ASYU50717.2020.9259880](https://doi.org/10.1109/ASYU50717.2020.9259880).
- [32] L. Xiong and Y. Yao, "Study on an adaptive thermal comfort model with K-nearest-neighbors (KNN) algorithm," *Build. Environ.*, vol. 202, p. 108026, 2021, doi: [10.1016/j.buildenv.2021.108026](https://doi.org/10.1016/j.buildenv.2021.108026).
- [33] A. Nur, S. Syarifandi, and S. Amin, "Implementation of Text Mining Classification as a Model in the Conclusion of Tafsir Bil Ma' tsur and Bil RA' YI Contents," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 2789-2795, 2019. doi: [10.35940/ijeat.A9780.109119](https://doi.org/10.35940/ijeat.A9780.109119).
- [34] S. Jukic, M. Saracevic, A. Subasi, and J. Kevric, "Comparison of Ensemble Machine Learning Methods for Automated Classification of Focal and Non-Focal Epileptic EEG Signals," *Mathematics*, vol. 8, no. 9, p. 1481, Sep. 2020. doi: [10.3390/math8091481](https://doi.org/10.3390/math8091481).
- [35] H. Sayadi, N. Patel, S. M. P.D., A. Sasan, S. Rafatirad and H. Homayoun, "Ensemble Learning for Effective Run-Time Hardware-Based Malware Detection: A Comprehensive Analysis and Classification," *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1-6. doi: [10.1109/DAC.2018.8465828](https://doi.org/10.1109/DAC.2018.8465828).
- [36] R. Diaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest.," *BMC Bioinformatics*, vol. 7, p. 3, Jan. 2006. doi: [10.1186/1471-2105-7-3](https://doi.org/10.1186/1471-2105-7-3).
- [37] K. U. Syaliman, E. B. Nababan, and O. S. Sitompul, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *J. Phys. Conf. Ser.*, vol. 978, no. 1, p. 012047, Mar. 2018. doi: [10.1088/1742-6596/978/1/012047](https://doi.org/10.1088/1742-6596/978/1/012047).
- [38] Sovit Ranjan Rath, "An Introduction to k-Nearest Neighbors in Machine Learning," *Debugger Cafe*, 2019. <https://debuggercafe.com/an-introduction-to-k-nearest-neighbors-in-machine-learning/> (accessed Oct. 01, 2021).