# Fault Detection in Power Transmission Lines Using AI Model

Assistant Prof Dr. T. Dharani
*Department of ISE*
*Rajarajeswari College of Engineering*
Bangalore, India
drtdharani@gmail.com

Shrinidhi
*Department of ISE*
*Rajarajeswari College of Engineering*
Bangalore, India
shrinidhi0904@gmail.com

Rakshitha R
*Department of ISE*
*Rajarajeswari College of Engineering*
Bangalore, India
rakshitharajaram9@gmail.com.

Vaishnavi D S
*Department of ISE*
*Rajarajeswari College of Engineering*
Bangalore, India
vaishnavisrinivassrinivas@gmail.com

Shreya B R
*Department of ISE*
*Rajarajeswari College of Engineering*
Bangalore, India
shreyabr202@gmail.com

***Abstract— Unexpected disturbances in electrical power transmission lines can arise due to numerous unpredictable conditions. Such faults interrupt the flow of electricity and must be detected quickly to prevent major system failures. Traditional fault-finding methods rely heavily on manual inspection and human interpretation, which is slow and prone to mistakes. Therefore, automated and intelligent fault-diagnosis techniques have gained significant attention in recent years.***

***This research focuses on developing an AI-based model to identify faults in power transmission lines using Machine Learning (ML). A cleaned and integrated Kaggle dataset is used for training and testing. Multiple ML algorithms—Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF)—are compared for classification performance. Based on evaluation metrics, the Naïve Bayes model delivers the best results with an accuracy of 97.77%, recall of 97.09%, precision of 98.64%, and F1-score of 97.86%. These findings show that NB can serve as a strong and efficient tool for fault detection in transmission systems.***

***Keywords – Fault Detection, Machine Learning, Classification, Accuracy, Power Transmission.***

## I. Introduction

To assure a consistent supply of electricity and electric power systems are essential for modern living, and even a small disturbance in transmission lines can result in major outages. Transmission lines carry electricity over long distances, making them vulnerable to environmental changes, equipment damage, and operational failures. When a fault occurs, the continuity of power supply is affected, leading to instability and economic losses. Identifying and correcting such faults as early as possible is therefore crucial. Transmission lines carry electricity over long distances, making them vulnerable to environmental changes, equipment damage, and operational failures Power transmission networks play a vital role in delivering electricity from generation stations to consumers. These networks operate across vast geographical areas and are constantly exposed to environmental variations, aging equipment, mechanical failures, and unpredictable external influences. As a result, faults in transmission lines are inevitable and, if not detected promptly, can lead to voltage instability, line outages, large-scale power interruptions, and financial losses for utilities and consumers.

Traditionally, fault detection in transmission lines depends heavily on manual monitoring, protective relays, and periodic inspections. While these methods have been used for decades, they suffer from several limitations. Manual inspection is slow, labour-intensive, and sometimes impractical in remote or hazardous locations. Conventional protective devices often rely on predefined thresholds and may not efficiently distinguish between normal disturbances and actual faults, especially in complex and modern power grids. In addition, the increasing load demand and interconnection of multiple power sources make fault analysis more challenging.

With the advancement of digital technologies and the increasing availability of measurement data, the power sector is gradually shifting towards smarter, automated diagnostic solutions. Artificial Intelligence (AI) and Machine Learning (ML) have emerged as highly promising tools for analyzing power system behaviour due to their ability to learn patterns, process large datasets, and make accurate predictions. Machine Learning models can study current and voltage variations, identify hidden correlations, and classify different categories of faults without relying solely on heuristic or rule-based methods.

Earlier research has explored various ML algorithms such as Support Vector Machine (SVM), Decision Trees, Random Forests, and neural network models for identifying power system anomalies. These studies demonstrate that ML-based approaches are capable of improving detection speed, minimizing false alarms, and providing more reliable fault classification when compared to conventional methods. However, the choice of algorithm and dataset significantly influences the performance of the diagnostic system. In this study, we focus on building an efficient ML-based fault detection model using real-time current and voltage features obtained from an open-source dataset. The research involves data preprocessing, feature analysis, model training, and performance comparison among three widely used algorithms—Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF). The aim is to identify the model that offers the best accuracy and robustness for diagnosing faults in power transmission lines. Such an automated system can support power utilities in reducing downtime, improving reliability, and advancing

Several studies have explored the use of machine learning for fault detection in transmission lines. One such work developed a fault detection model using a multi-layer perceptron and compared its performance with other commonly used algorithms. The researchers evaluated their approach on a dataset containing more than 12,000 samples of three-phase current and voltage measurements. Their findings showed that the multi-layer perceptron delivered better accuracy than Random Forest, Decision Tree, and SVM models, outperforming them by 0.6%, 2.7%, and 1.9% respectively.

The study also analyzed different input combinations for the ML models to identify which configuration produced the highest accuracy. In total, 24 machine learning models were trained, and the most accurate one—with the best input features—was selected for further use. Another related work introduced an SVM-based method using a Gaussian Kernel for identifying line faults. This technique requires only one cycle of pre-fault and post-fault current data, making it computationally efficient... The method was tested on a PSCAD-simulated 400 kV, 50 Hz power network consisting of two areas. The algorithm dynamically adjusts after each cycle to incorporate updated post-fault current values. Tests were conducted under varying resistance levels, fault distances, and different types of faults. Accurately identifying faults at an early stage is essential because failures not only disrupt power system reliability but can also lead to financial losses and safety hazards for both equipment and personnel. Although faults cannot be completely eliminated, their timely detection significantly reduces severe damage in power systems.

Because of the growing complexity of modern power systems, investing in fault detection and classification (FDC) research has become very important. Most of today’s approaches rely on sampled current and voltage data collected from the network, which is then processed to identify faults. With the rise of Machine Learning and other AI techniques, data-driven methods are now replacing traditional rule-based systems, as they are more flexible and often provide better performance. These AI models excel because they can learn patterns directly from data and adapt to different fault conditions, making them more general and reliable. In this work, we aim to develop a strong and adaptable AI-based solution for detecting and classifying faults in power transmission lines.

## PROCESS FLOW

This section explains the workflow followed for detecting different types of faults in power transmission lines. The dataset used in this study, which includes various fault conditions with corresponding current and voltage values, is collected from the Kaggle platform. Once the data is obtained, it goes through several preprocessing steps such as data cleaning, integration, and splitting to prepare it for model development..Three machine learning models—Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF)—are built for comparison. After constructing these models, the dataset is divided into two parts: one portion is used for training the models, and the other portion is used for testing them… The whole system is illustrated in Fig. 1.

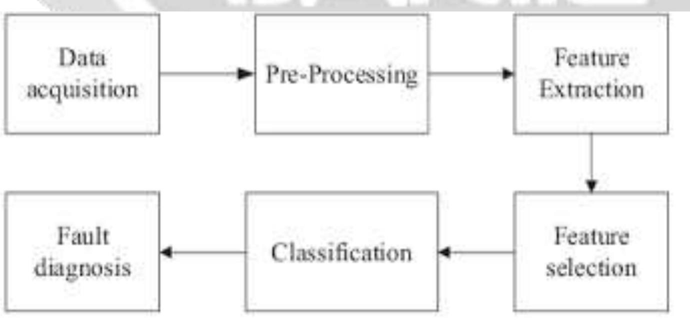


Fig. 1. Research flow of fault detection in power transmission lines.

### *Data Acquisition*

*The delivery of electricity from power stations to homes and industries depends heavily on the condition of components like power lines, poles, and insulators. When these components fail, two major types of faults can occur—Open Circuit (OC) faults and Short Circuit (SC) faults. Both of these can appear as either symmetrical or asymmetrical faults.OC faults usually happen when the voltage becomes too high, whereas SC faults have become more frequent due to rising current levels. A typical symmetrical fault is a three-phase short circuit, where all three phases are affected at the same time, with or without involving the ground.*

Asymmetrical faults, on the other hand, involve only certain phases. Examples include:

Single line-to-ground (L-G) fault – occurs when one phase comes into contact with the ground, usually due to insulator failure.
Double line-to-ground (L-L-G) fault – happens when two phases and the ground are shorted together.
Line-to-line (L-L) fault – occurs when any two phases short-circuit.
Three-phase to ground (L-L-L-G) fault – all three phases shorted along with the ground.
The dataset shown in Fig. 2, sourced from Kaggle, is used for training the AI model to classify these faults. In the dataset, the first four columns indicate the type of fault:
No fault: [0000] ,L-G fault: [1001],L-L fault: [0011],L-L-G fault: [1011],L-L-L fault: [0111].

Fig. 3 depicts the breakdown of the obtained data (which comprises six different fault categories).

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | G | C | B | A | Ia | Ib | Ic | Va | Vb | Vc |
| 2 | 1 | 0 | 0 | 1 | -151.292 | -9.67745 | 85.8002 | 0.40075 | -0.13293 | -0.26781 |
| 3 | 1 | 0 | 0 | 1 | -336.186 | -76.2833 | 18.3289 | 0.31275 | -0.12363 | -0.1891 |
| 4 | 1 | 0 | 0 | 1 | -502.892 | -174.648 | -80.9247 | 0.26573 | -0.1143 | -0.15143 |
| 5 | 1 | 0 | 0 | 1 | -593.942 | -217.703 | -124.892 | 0.23551 | -0.10494 | -0.13057 |
| 6 | 1 | 0 | 0 | 1 | -643.664 | -224.159 | -132.283 | 0.20954 | -0.09555 | -0.11398 |
| 7 | 1 | 0 | 0 | 1 | -632.313 | -181.715 | -90.7955 | 0.19312 | -0.08614 | -0.10697 |
| 8 | 1 | 0 | 0 | 1 | -557.392 | -119.469 | -29.5294 | 0.21 | -0.07671 | -0.13329 |
| 9 | 1 | 0 | 0 | 1 | -458.8 | -96.3189 | -7.38185 | 0.27365 | -0.06726 | -0.20639 |
| 10 | 1 | 0 | 0 | 1 | -385.669 | -97.9898 | -10.0768 | 0.33465 | -0.0578 | -0.27685 |

Fig. 2. Sample Data

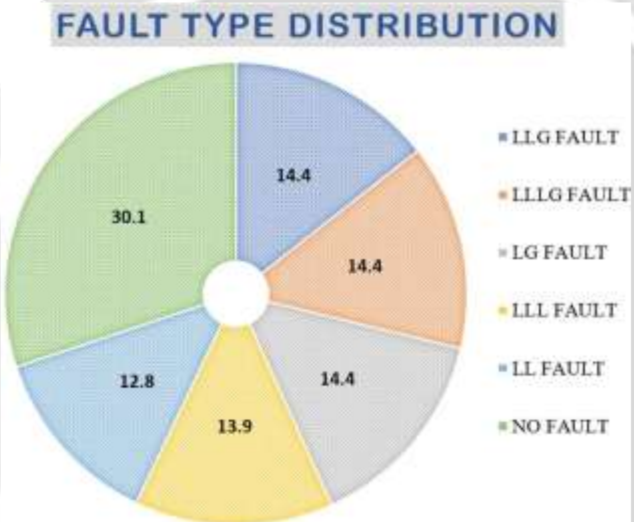


Fig. 3. Fault data distribution

*A. Data Processing*

Data processing is one of the most crucial steps at the beginning of any machine learning project. At this stage, the main goal is to make sure the dataset is clean, accurate, and suitable for analysis. Pre-processing helps fix issues such as duplicate entries, inconsistent formats, noisy data, missing labels, imbalance between classes, and unnecessary features. Although this step can be time-consuming, it plays a major role in improving the quality of the final model. In this study, careful attention was given to cleaning and preparing the dataset before model development..

Missing values are one of the key reasons for poor classification performance. Real-world datasets often contain incomplete entries, not only in medical fields but across all domains. A common way to handle missing data is through imputation, where the missing values are replaced based on certain rules. However, the best imputation method depends on factors like dataset size, the percentage of missing data, and the reason for the missing values. In many cases, it is better to work with complete datasets to avoid any artificial adjustments. After verification, our collected dataset was found to have no null or missing entries.

The relationship between ML and data integration has grown in importance alongside the explosion in data size and variety [16]. Data Integration is critical for ML to thrive since it requires data from as many diverse sources as feasible. Simultaneously, ML in data integration results in lower costs and more accuracy. The collected data of the first four columns represent the type of fault and it is hard to process and classify the type of fault. In this integration step, the four columns will be merged and produce the new column called fault type which is shown in Fig. 2. This step helps to convert four target variables to one target variable.

| Ia | Ib | Ic | Va | Vb | Vc | Fault Type |
|---|---|---|---|---|---|---|
| -151.292 | -9.67745 | 85.80016 | 0.40075 | -0.13293 | -0.26781 | 1001 |
| -336.186 | -76.2833 | 18.3289 | 0.312732 | -0.12363 | -0.1891 | 1001 |
| -502.892 | -174.648 | -80.9247 | 0.265728 | -0.1143 | -0.15143 | 1001 |
| -593.942 | -217.703 | -124.892 | 0.235511 | -0.10494 | -0.13057 | 1001 |
| -643.664 | -224.159 | -132.283 | 0.209537 | -0.09555 | -0.11398 | 1001 |
| -632.313 | -181.715 | -90.7955 | 0.193116 | -0.08614 | -0.10697 | 1001 |
| -557.392 | -119.469 | -29.5294 | 0.210004 | -0.07671 | -0.13329 | 1001 |
| -458.8 | -96.3189 | -7.38185 | 0.273652 | -0.06726 | -0.20639 | 1001 |
| -385.669 | -97.9898 | -10.0768 | 0.334649 | -0.0578 | -0.27685 | 1001 |

Fig. 4. Outcome of data integration

To build a reliable machine learning model, the dataset must be divided into different parts. Usually, one portion of the data is used to train the model, while the remaining portion is used to test how well the model performs. This step is essential in data science because it helps prevent the model from learning only the training data and ensures it can perform well on new, unseen data.

Figure 4 shows the overall process of splitting data for model development. In a basic two-part split, the training set is used to teach the model and help it learn important patterns. This set also helps in choosing the best model and tuning its parameters. Once the training phase is complete, the testing set is used to evaluate the final model and check its accuracy.in Table I.

TABLE. I. DATA SPLITTING

| Total | Train | Test |
|---|---|---|
| 7861 | 6289 | 1572 |

### B. *AI Model*

Three different AI models are used in this research to identify the fault type. All models are detailed in the below sections.

Support Vector Machine (SVM) is a popular machine learning technique used for both classification and regression tasks. It works on the principle of supervised learning, where data points are separated into distinct classes. To achieve this, SVM converts the original data into a higher-dimensional space using nonlinear mapping. This transformation helps the algorithm identify a function that can correctly predict the category of both current and unseen data.

In an ideal scenario, if the right nonlinear mapping is applied, two classes of data can always be separated using a hyperplane. This hyperplane, known as the decision boundary, divides the data points into different classes. Although many possible hyperplanes can separate the data, SVM selects the one that provides the maximum margin — the widest possible separation between the two classes. A larger margin generally leads to better classification performance.

analysis. The recall value is attained by the ML model and it is arranged from higher to lower value. The order of the ML model is NB (97.09%), RF (96.08%), and SVM (95.48%). The
third metric is considered precision. The NB gives a higher precision rate of 98.64%, RF provides the second largest precision score of 94.81%, and SVM delivers lower precision of 94.21%. Finally, F1-Score is taken, the value of 94.85%, 97.86%, and 95.44% obtained by SVM, NB, and RF. The metrics value of all three ML models is detailed in table 2.

TABLE. II. DATA SPLITTING

| MODEL | SVM | NB | RF |
|---|---|---|---|
| ACCURACY | 94.5929 | 97.7792 | 95.229 |
| RECALL | 95.4823 | 97.0979 | 96.088 |
| PRECISION | 94.2169 | 98.6486 | 94.813 |
| F1-SCORE | 94.8454 | 97.8672 | 95.4463 |

Table II is converted into a bar graph for easy comparison Random Forest (RF) Classifier

The Random Forest (RF) algorithm is an ensemble learning method composed of a large number of CART (Classification and Regression Trees). Each tree is trained using a random bootstrap sample drawn from the original training dataset. The samples that are not selected during the bootstrap process are known as out-of-bag (OOB) samples and are used to estimate the generalization error of the model.

The OOB error provides a reliable estimate of model performance without the need for a separate validation dataset. Each decision tree in the forest gives an individual prediction, and the final output is determined by majority voting among all trees.

RF is known for its robustness, resistance to overfitting, and high accuracy, particularly when handling noisy and high-dimensional data. It also performs implicit feature selection using the random subspace method, which enhances its predictive stability and reliability in fault detection and classification tasks.

---

## II. RESULTS AND DISCUSSION

The performance of three machine learning models—Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF)—was evaluated using four key metrics: accuracy, recall, precision, and F1-score.

Among the three models, the NB classifier achieved the highest accuracy of 97.77%, indicating superior fault classification capability. RF attained an accuracy of 95.22%, while SVM produced a slightly lower accuracy of 94.59%.

In terms of recall, NB again outperformed the other models with a value of 97.09%, followed by RF at 96.08%, and SVM at 95.48%. A higher recall indicates that the model is more effective at identifying actual faults.

For the precision metric, NB achieved the best result with 98.64%, confirming that it made the most reliable positive predictions. RF recorded a precision of 94.81%, and SVM obtained 94.21%.

The F1-score, which combines precision and recall into a single performance indicator, also highlighted the strength of the NB model with a value of 97.86%. RF and SVM achieved F1-scores of 95.44% and 94.85%, respectively.
all three ML model metrics. Fig. 5 shows the ML model comparison. The three various ML models such as SVM, NB, and RF are represented in different colours like green, blue and grey color. The height of the blue bar will be higher than the other two, and it represents the NB will give better results for fault classification.

The Naïve Bayes (NB) classifier is based on Bayes' Theorem and follows the assumption that all predictor variables are independent of each other. Even though this assumption is often unrealistic in real-world datasets, NB performs remarkably well in many classification tasks, especially when handling large-scale data.

Feature importance in the NB approach is evaluated using the Gini impurity criterion. The Gini index is a non-parametric metric that measures how well a variable separates different classes without depending on any specific data distribution. For a node n in a binary decision tree, the Gini index is defined as:

$$\mathrm{Gini}(n) = 1 - \sum (p_k)^2$$

where  is the relative frequency of class k. A lower Gini value indicates better purity of the node. Therefore, a greater reduction in the Gini index signifies that the corresponding feature contributes significantly to class separation.

The NB classifier is simple in structure yet highly efficient. Despite its simplicity, it often competes with and even outperforms more complex machine learning techniques. It offers an effective probabilistic framework for learning from data and is widely used due to its fast training and prediction capabilities. The posterior probability is calculated using Bayes' theorem as:

$$P(U|V) = \frac{P(V|U) \, P(U)}{P(V)}$$

where  is the posterior probability,  is the likelihood,  is the prior probability, and  is the evidence.

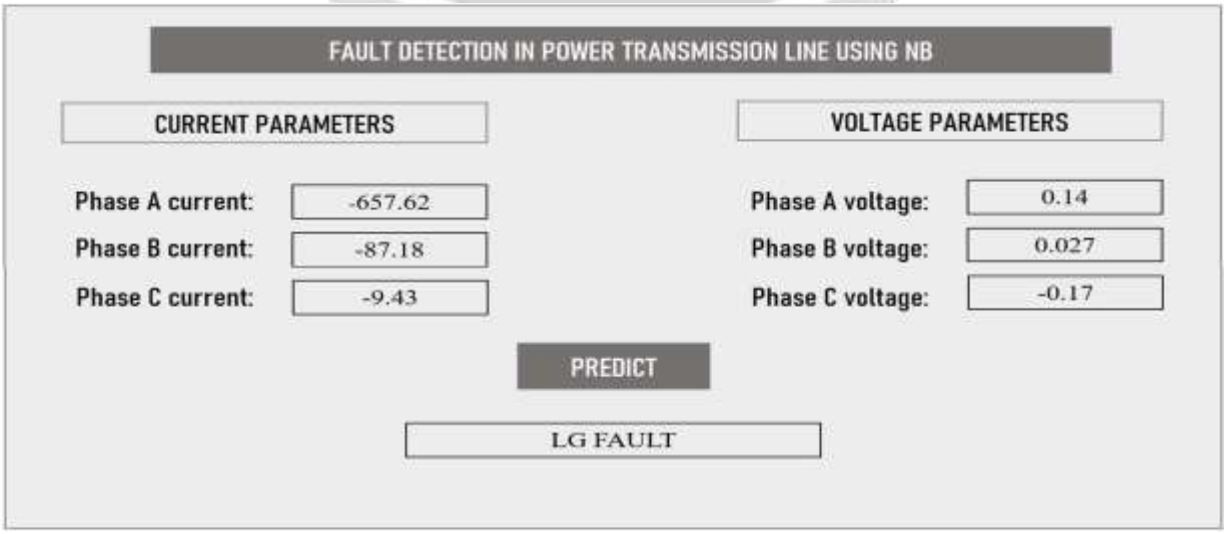


Fig. 6. GUI working on fault detection by deploying NB

## II. CONCLUSION

In this study, the performance of three widely used machine learning classifiers—Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF)—was evaluated for fault classification. The comparison was carried out using standard performance metrics such as accuracy, recall, precision, and F1-score. Among all the models, the Naïve Bayes classifier consistently demonstrated superior performance across all evaluation measures. It achieved the highest accuracy, precision, recall, and F1-score, indicating its strong capability in correctly identifying and classifying fault conditions.

The Random Forest model also showed reliable and stable performance, ranking second across most metrics, due to its robustness and ensemble-based structure. The Support Vector Machine, although effective, produced comparatively lower results than NB and RF for this specific dataset and application.

Overall, the results confirm that the Naïve Bayes classifier is the most suitable model for the proposed fault classification system due to its high accuracy, simplicity, and computational efficiency. This makes NB a strong candidate for real-time fault detection applications. Future work can focus on testing the model with larger real-world datasets and exploring hybrid or deep learning approaches to further enhance classification performance.

## REFERENCES

[1]. Liu, J., Zhao, Z., Ji, J., & Hu, M. (2020). Research and [10]. Saini, M., Sultan, A. R., Mustafa, M. W., & Rahimuddin, R. (2020). Faults detection and classification on parallel transmission lines using modified Clarke's transformation- ANN approach. PRZEGLĄD ELEKTROTECHNICZNY, ISSN 0033-2097, R. 96 NR 4/2020, 96(NR 4), 23-27.

[2]. Kaur, H., & Kaur, M. (2020). Fault classification in a transmission line using Levenberg –Marquardt algorithm based artificial neural network. In Data Communication and Networks: Proceedings of GUCON 2019 (pp. 119-135). Springer Singapore.

[3]. Jayamaha, D. K. J. S., Lidula, N. W. A., & Rajapakse, A. D. (2019). Wavelet-multi resolution analysis based ANN architecture for fault detection and localization in DC microgrids. IEEE Access, 7, 145371- 145384.

[4]. Prakash, E. S. (2021, May 22). Electrical fault detection and classification. Kaggle. https://www.kaggle.com/datasets/esathyaprakash/electrical-fault-detectionclassification?resource=download&select=detect_dataset.cs

[5]. Breiman, L. (2001). Random forests. Machine learning, 45, 5-3

[6]. Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.

application of wireless sensor network technology in power transmission and distribution system. Intelligent and Converged Networks, 1(2), 199-220.

[7]. Qoria, T. (2020). Grid-forming control to achieve a 100% power electronics interfaced power transmission systems (Doctoral dissertation, HESAM Université).

[8]. Chaari, M. Z., & Al-Maadeed, S. (2020, February). Wireless power transmission for the internet of things (iot). In 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT) (pp. 549-554). IEEE.

[9]. Ehsani, M., Singh, K. V., Bansal, H. O., & Mehrjardi, R. T. (2021). State of the art and trends in electric and hybrid electric vehicles. Proceedings of the IEEE, 109(6), 967-984.

[10]. Guo, E., Jagota, V., Makhatha, M. E., & Kumar, P. (2022). Study on fault identification of mechanical dynamic nonlinear transmission system. Nonlinear Engineering, 10(1), 518-525.

[11]. Shifat, T. A., & Hur, J. W. (2021). ANN assisted multi sensor information fusion for BLDC motor fault diagnosis. IEEE Access, 9, 9429-9441.

[12]. Leh, N. A. M., Zain, F. M., Muhammad, Z., Abd Hamid, S., & Rosli, A. D. (2020, August). Fault detection method using ANN for power transmission line. In 2020 10th IEEE international conference on control system, computing and engineering (ICCSCE) (pp. 79-84). IEEE.

[13]. Frizzo Stefenon, S., Waldrigues Branco, N., Nied, A., Wildgrube Bertol, D., Cristian Finardi, E., Sartori, A., ... & Bartnik Grebogi, R. (2020). Analysis of training techniques of ANN for classification of insulators in electrical power systems. IET Generation, Transmission & Distribution, 14(8), 1591-1597.Alagarsamy, "Machine to machine communication enabled internet of things: a review", International Journal of Reconfigurable and Embedded Systems, 2022, 11(2), pp. 126-1346, IEEE.

[14]. Cherif, B. D. E., Bendiabdellah, A., & Tabbakh, M. (2020). An automatic diagnosis of an inverter IGBT open-circuit fault based on HHT-ANN. Electric Power Components and Systems, 48(6-7), 589-602.

[15]. Saini, M., Sultan, A. R., Mustafa, M. W., & Rahimuddin, R. (2020). Faults detection and classification on parallel transmission lines using modified Clarke's transformation- ANN approach. PRZEGLĄD ELEKTROTECHNICZNY, ISSN 0033-2097, R. 96 NR 4/2020, 96(NR 4), 23-27.

[16]. Kaur, H., & Kaur, M. (2020). Fault classification in a transmission line using Levenberg –Marquardt algorithm based artificial neural network. In Data Communication and Networks: Proceedings of GUCON 2019 (pp. 119-135). Springer Singapore.

[17]. Jayamaha, D. K. J. S., Lidula, N. W. A., & Rajapakse, A. D.
(2019). Wavelet-multi resolution analysis based ANN architecture for fault detection and localization in DC microgrids. IEEE Access, 7, 145371- 145384.

[18]. Prakash, E. S. (2021, May 22). Electrical fault detection and classification. Kaggle. https://www.kaggle.com/datasets/esathyaprakash/electrical-fault-detectionclassification?resource=download&select=detect_dataset.cs

[19]. Breiman, L. (2001). Random forests. Machine learning, 45, 5-3

[20]. Tyralis, H., Papacharalampous, G., & Langousis, A. (2019). A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910.