Defence Science Journal, Vol. 75, No. 5, September 2025, pp. 603-614, DOI : 10.14429/dsj.21069

Machine Learning-Based Fault Analysis: Transforming Correlated Fault Data in Distributed Generation Systems

Kamlesh Singh Bisht*, Nafees Ahamad, and Saurabh Awasthi

Department of Electrical and Electronics & Communication Engineering, DIT University, Dehradun - 248 009, India
**E-mail: ksbisht99@gmail.com*

ABSTRACT

Distributed generation systems offer several advantages over centralised power systems. However, faults within the system can lead to stability loss, reclosure failures, and voltage fluctuations, necessitating special attention to safeguard the system's components. Identifying the location and nature of faults is crucial in preventing adverse effects on the system's overall functionality. Various machine learning (ML) techniques have been proposed for fault location and categorization in distributed generation systems. While many of these techniques effectively pinpoint the fault's location, accurately determining the fault type remains a challenge. This work presents a novel approach to enhance fault classification efficiency by transforming correlated fault data using Principal Component Analysis (PCA) in ML. The suggested findings demonstrate that the proposed method can significantly improve the performance of specific fault classification algorithms in power systems. The novelty adopted in reducing the dimensions before subjecting the data to the ML algorithm has given higher accuracy in identifying fault types in a faster time frame, thereby enhancing the security and stability of distributed generation systems. Notably, the suggested method achieved a faster time frame. This paper has explored excellent accuracy with two of the ML algorithms namely the; Random Forest Classifier and K-Nearest Neighbors underscoring their further potential for improving system protection.

Keywords: Distribution network; Distributed generation; Power system modeling; Fault classification; Machine learning (ML); Principal Component Analysis (PCA); Feature extraction

NOMENCLATURE

μ	: Mean values
σ	: Standard deviations
λ	: Eigenvalues of the matrix

1. INTRODUCTION

The concept of the smart grid is rapidly evolving due to various limitations associated with conventional power plants primarily due to the substantial transmission costs and losses incurred when delivering power from these stations to end-users. Distributed Generation (DG) plays a significant role in these smart grids, offering inherent advantages such as reduced transmission expenses, enhanced electricity supply to end-users, diminished distribution costs, heightened efficiency, improved reliability, decreased maintenance expenditures, and a more environmentally conscious energy culture[1]. The integration of DG into the smart grid necessitates an upgrade of the entire power transmission and distribution system from the conventional, centralized power generation model, which relies on massive power plants generating electricity centrally and distributing it over long distances via the grid.

Distributed Generation aims to generate energy at or near the point of consumption[2,] usually at a lower voltage level and typically under 50 MW. This concept is not only limited to generating power at scattered locations but also involves integrating energy storage technologies and various renewable energy sources. Here a comprehensive strategy is employed to achieve a reduction in transmission cost, stability, reliability, and enhanced efficiency, utilizing a wide range of energy sources such as nuclear, solar, tidal, wind, and mini-hydro-power, all integrated with the established conventional centralized power generation system.

Mitigation of certain challenges must be addressed for the successful incorporation of DG into a smart grid. One major challenge is managing the entire distribution system in a healthy state without "islanding" during outages. This requires precise and reliable fault detection at the earliest possible moment when a fault occurs. Therefore, accurately identifying faults in terms of their location and type is essential for taking timely corrective actions and preserving the grid's integrity.

Faults in distribution systems can arise from various factors, such as short circuits, equipment malfunctions, operational errors, overloads, vegetation growth, and ageing[3]. Line faults are the most prevalent in distribution systems[4,] with common types including single-line-to-ground (LG), line-to-line (LL), double-line-to-ground (LLG), and balanced three-phase faults (LLLG)[5]. The probability of the LG fault has been reported to be as high as 70 % in a distributed system, followed by 15 % for LL faults, 10 % for LLG faults, and a mere 5 % for LLLG faults, which, although rare, are the most severe and detrimental to the health of the entire grid[6].

Received : 18 November 2024, Revised : 27 March 2025
Accepted : 20 June 2025, Online published : 01 September 2025

Conventional fault location techniques[7-12] and the conventional fault classification techniques[13-17] are not precise and have inherent limitations. However, with Artificial Intelligence (AI) and Machine Learning (ML), better, faster, and more precise methods have been developed using various techniques. For example, Awasthi[18], *et al.* proposed a fault-type identification method using a shallow neural network in a real-time distributed generation system with two distributed generators and a utility grid. The dataset, derived from MATLAB simulations, was used to compare results from the training, validation, and testing phases. Although this approach is novel, however, the classification accuracy obtained is merely 90 %, highlighting a key limitation of this technique.

In contrast, Alsafasfeh[19], *et al.* proposed a three-stage scheme for fault detection and classification. This approach involves using phase current signatures for fault detection. A two-stage classification process comprises training and testing, and Principal Component Analysis (PCA) for classification. Although classification accuracy of 94.54 % was achieved with a single template for each fault signature in the training set, the paper fails to address the computational speed and time required when using two templates per fault in the training set. Furthermore, no ML algorithm was utilized in the classification process.

Ajagekar[20], *et al.* developed a hybrid Quantum Computing (QC)-based deep learning framework for fault diagnosis, using a Conditional Restricted Boltzmann Machine (CRBM)-based network to extract suitable features from time-series data.
The method demonstrated some improvement in response time over traditional methods like Decision Tree (DT) or Artificial Neural Network (ANN) but the performance with missed data for triple line and triple line-to-ground faults requires further improvement. A fuzzy logic and neural network-based technique was explored in[[21] for fault analysis in electric power systems, reporting accuracy rates of 89 % for fault type identification and 93 % for fault location. However, these accuracy rates are insufficient for the reliable operation of smart grids. A hybrid approach[22] combines a Support Vector Machine (SVM) with transient voltage and current data to determine fault distance. This method, however, only identifies the fault distance and not the fault classification. SVM was evaluated for fault diagnosis using data from a 132/11kV grid substation, demonstrating commendable performance in both classification and regression tasks[23]. Tawab[24], *et al.* employed Discrete Wavelet Transform (DWT) for fault classification and combined DWT with SVM to identify fault locations. The results obtained from the IEEE 33-bus system with distributed generators are highly promising, but the computational complexity of wavelet transform-based techniques may pose challenges for real-time applications.

Other researchers have explored various methods, with some claiming high reliability but lacking precisin, while others achieve precision but fall short of reliability[25-30]. In response to these limitations, this paper proposes a novel method for fault classification in power systems by transforming correlated fault data using PCA to reduce the dimensionality of the feature space from eighteen fault variables (voltage and current) to just six. Our method demonstrates significantly improved efficiency, with performance gains of up to 100 % compared to the same approach without transformation.

This paper is organized into six sections. Section 1 presents a comprehensive literature review, followed by the system background in Section 2. In Section 3, the proposed methodology is detailed, covering system modelling, data collection, normalization, and data transformation using PCA. Section 4 discusses various ML classification algorithms. Experimental results are presented in Section 5, and the paper concludes with Section 6.

2. SYSTEM BACKGROUND

A 66 kV distribution line, 100 km in length, is considered here, connected to a grid rated at 100 MVA, 132 kV, and 50 Hz. A 100 MVA, 132/66 kV step-down transformer is installed between the grid and the distribution line to match the distribution line voltage. The system also includes two star-grounded distributed generators, each rated at 100 kVA, 3.3 kV, and 50 Hz, which are connected to the distribution line through 100 kVA, 3.3/66 kV, and 50 Hz transformers. Additionally, the distribution system is connected to a 10 kW load to dissipate power. The 100 Km line is divided into 2 km intervals for fault generation, and the fault parameters are recorded for these locations. The complete system is illustrated in Fig. 1. In the event of a fault, ensuring the distributed generation's optimal operation, protection, and stability is paramount.

Though fault parameters have been recorded at regular intervals of 2 Km only however the probability of occurrence of a fault can be anywhere in between these intervals too along the entire distribution line or near the generators. Hence higher the resolution of fault intervals the better performance of the algorithm would be achieved but at the cost of more computational time essentially, all faults fall into two main categories: open-circuit faults and short-circuit faults. Open-circuit faults happen when one, two, or all three conductors become open, often referred to as series faults because they occur in series with the line. Conversely, short-circuit faults arise from the short-circuiting of phases and can be further classified into symmetrical and unsymmetrical faults. Symmetrical faults include three-phase short circuits and three-phase-to-ground faults. In contrast, unsymmetrical faults encompass single-line-to-ground, double-line-to-ground, and line-to-line faults. Symmetrical faults maintain system balance before and after their occurrence, while unsymmetrical faults cause discrepancies between pre-fault and post-fault

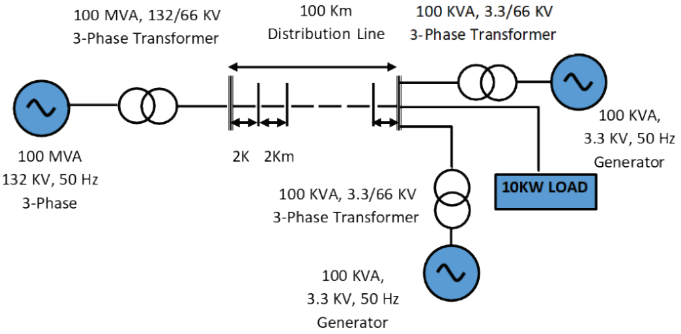


Figure 1. Single line diagram of the investigated distributed generation system.

conditions. Therefore, the primary objective of this research is to accurately identify the fault type in distributed generation systems to ensure proper operation.

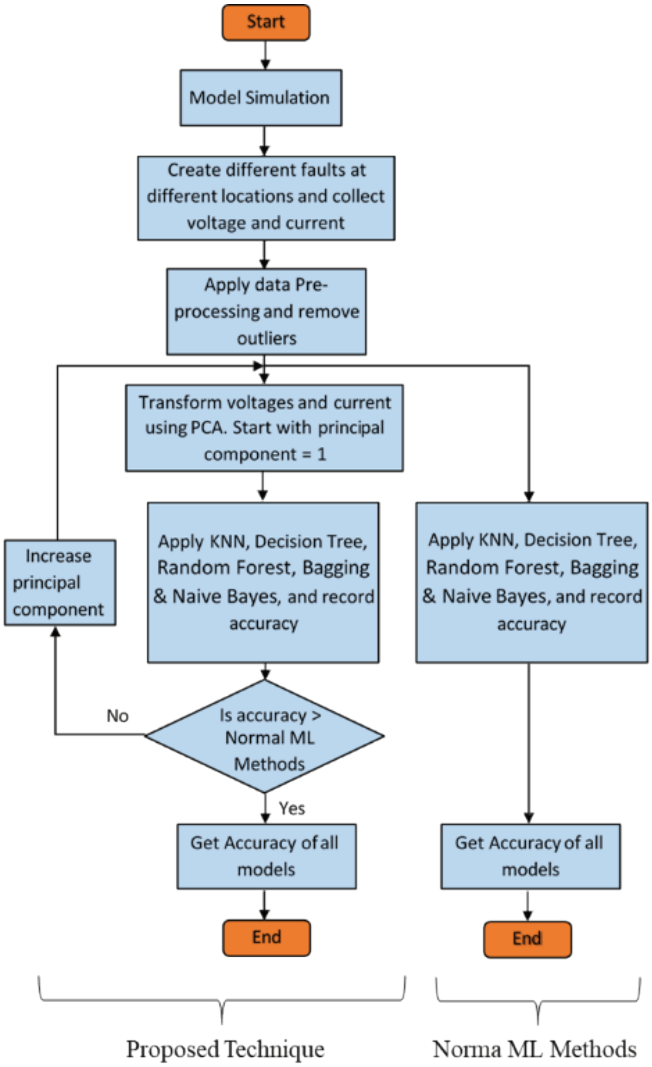


Figure 2. Proposed algorithm for fault classification.

3. METHODOLOGY

To accurately identify fault types in distributed generation systems, this paper introduces a novel method to enhance fault classification efficiency by transforming correlated fault voltages and currents using Principal Component Analysis (PCA). The proposed approach consists of four main steps, as outlined below, with a detailed algorithm depicted in Fig. 2.

Step 1: Data generation and pre-processing.
Step 2: Transformation of correlated voltages and currents using PCA and result extraction using various ML Methods.
Step 3: Obtaining results without transformation using various ML Methods for comparison.
Step 4: Comparison of results.

In Step 1, data is generated using a MATLAB Simulink model for the proposed system shown in Fig. 1, this data is further normalized for better accuracy. In Step 2, the correlated voltages and currents are transformed using PCA, and various machine-learning algorithms are applied for fault identification. Step 3 involves the process of fault identification in the system when no transformation is applied so that the result can be compared after the application of the proposed method and show the comparison. Finally, in Step 4, a comparative analysis is conducted to assess the accuracy of different techniques. The specific details of each phase are further elaborated in the subsequent sections.

3.1 System Modelling

System modelling is a crucial step in the proposed study. While the ideal source of data for machine learning and testing would be from a real powerhouse, collected over an extended period for all fault types, the approach is impractical due to its very high time-consuming nature and the difficulty of obtaining reliable data for all fault types. Therefore, MATLAB has been chosen as the platform to generate various faults at specified intervals and locations. The selection of location is critical and should be closely resembling with those of a practical power system as much as possible. The distributed power system depicted in Fig. 1 has been modelled in MATLAB, as shown in Fig. 3.

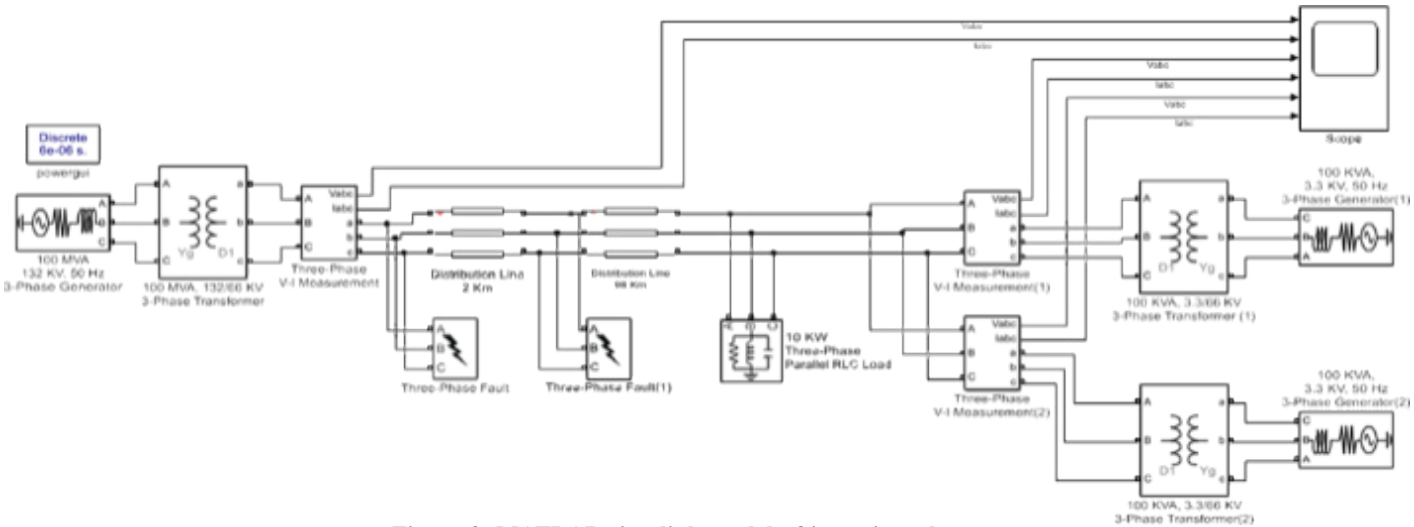


Figure 3. MATLAB simulink model of investigated system.

3.2 Fault Data Collection

All types of faults (single line-to-ground: RG, YG, BG; line-to-line: RY, RB, YB; double line-to-ground: RY-G, RB-G, YB-G; and three-phase fault-to-ground: RYB-G; where A, B, and C represent the three phases and G denotes ground) are simulated in the distributed generation system at specified locations (every 2 km interval). Voltage and current measurements are collected for each fault type and location by conducting short-circuit and load flow tests.

3.3 Fault Data Collection

To improve the performance of the ML model, data normalization is applied to standardize the values onto a consistent scale. Generally, there are four main techniques for normalization: scaling to a range, clipping, log scaling, and z-score normalization. In the research article by Singh[32], *et al.*, normalization is discussed in detail using a modified method also, which differs from the conventional approach where data is typically normalized using a single method. In our proposed system, we determine the upper and lower bounds of fault currents and voltages, which appear to follow an approximately uniform distribution. Specifically, we apply min-max scaling, often referred to as 'scaling to a range.' For voltage and current, this process can be expressed as:

$$f_{normalized} = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \tag{1}$$

where 'f' represents either fault voltage or current. This scaling method transforms voltage and current values to a common scale within the range of 0 to 1.

3.4 Data Transformation using Principal Component Analysis

Data dimensionality can primarily be reduced through two methods: feature selection or data transformation, also known as feature extraction. In feature selection, only the features that produce the best results are selected from the dataset. Conversely, feature extraction involves transforming the dataset into new, lower-dimensional features. Principal Component Analysis (PCA) is the most widely used machine learning technique for feature extraction in unsupervised learning. It is a statistical procedure that uses orthogonal transformations to convert correlated features into a set of linearly uncorrelated features called Principal Components (PCs). The steps involved in PCA include normalization, computation of the covariance matrix to identify correlations, computation of eigenvalues and eigenvectors to find the PCs, creation of a feature vector to decide which PCs to retain, and finally recasting the data along the PCs' axis. Using this method and the corresponding Eqn., the original data can be projected onto a new axis called the Principal Component. The steps involved are explained below:

Step 1: Standardization of fault voltages and currents so that their mean is zero and their standard deviation is one. The standardization is given by the following Eqn.:

$$V_{s\,\text{tandard}} = \frac{V - \mu_V}{\sigma_V} \tag{2}$$

$$I_{s\,\text{tandard}} = \frac{I - \mu_V}{\sigma_V} \tag{3}$$

where:

- V and I represent the fault voltages and currents, respectively.
- μV and μI are the mean values of voltage and current.
- σ_V and σ_I are the standard deviations of voltage and current.

Step 2: Estimation of the covariance, which shows the strength of the relationship between variables. For example, the covariance between two phase voltages VR and VY is computed as:

$$\text{cov}(V_R, V_Y) = \frac{\sum_{i=1}^{n}(V_{R_i} - \overline{V_R})(V_{Y_i} - \overline{V_Y})}{n-1} \tag{4}$$

where $\overline{V_R}$ and $\overline{V_Y}$ are the mean values of phase R and Y voltages. A positive covariance indicates that as V_R increases, V_Y also increases, while a negative covariance shows that as V_R increases, V_Y decreases. Zero covariance means there is no relationship between V_R and V_Y.

Step 3: Computation of eigenvalues and eigenvectors to identify the principal components This can be represented as follows:

Let V be a matrix of size $n \times n$ and X be a non-zero vector of $n \times 1$, then if the following equations hold:

$$VX = \lambda X \tag{5}$$

$$VX - \lambda I = 0 \tag{6}$$

Here, λ represents the eigenvalues of matrix V and X is the eigenvector. The above equation can be rearranged as:

$$(V - \lambda I)X = 0 \tag{7}$$

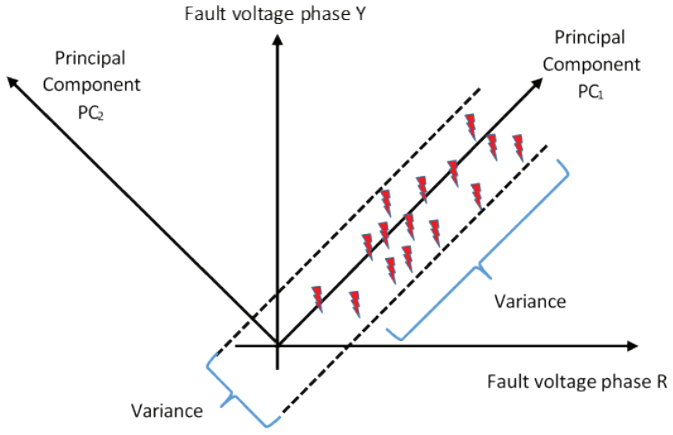


Figure 4. Voltage transformation using principal components.

Eqn. (7) is valid only when (V −λI) is a singular matrix, meaning its determinate is zero. The eigenvector can then be found using equation (5).

The entire process is illustrated in Fig. 4, where the x-axis represents the voltage of phase R, and the y-axis represents the voltage of phase Y. Various types of faults are depicted in the figure, and their variance is distributed along two principal axes, represented by PC1 and PC2, respectively.

4. MACHINE LEARNING CLASSIFIERS

In the proposed research, various machine learning (ML) classifiers are employed, both with and without data transformation, and their results are compared for accuracy. A brief explanation of these classifiers is provided as follows:

4.1 K-Nearest Neighbour (K-NN)

K-nearest neighbour (K-NN)[32] is a common classification method based on feature similarity. The algorithm works on the basic concept of storing existing data and classifying new cases based on their similarity to the data in the database. The appropriate value of k is selected through parameter tuning and is chosen as the closest match to the new sample from the database. Graphically, in the context of faults, this can be represented as shown in Fig. 5. The grid fault voltages (V_{GR} and V_{GY}) have been classified as line-to-line (LL) and line-to-ground (LG) faults, as illustrated in Fig. 5(a) now when a new fault occurs, it is assigned as an LG fault based on its distance from the existing classified points, as shown in Fig. 5(b). The distance between a new fault and the two existing categories can be calculated using various metrics such as Euclidean Distance, Manhattan Distance, or Minkowski Distance. In our research, we used Euclidean Distance between two fault points, which is given as:

$$\text{Distance}(1,2) = \sqrt{(V_{GR1} - V_{GR1})^2 + (V_{GY2} - V_{GY1})^2} \quad (8)$$

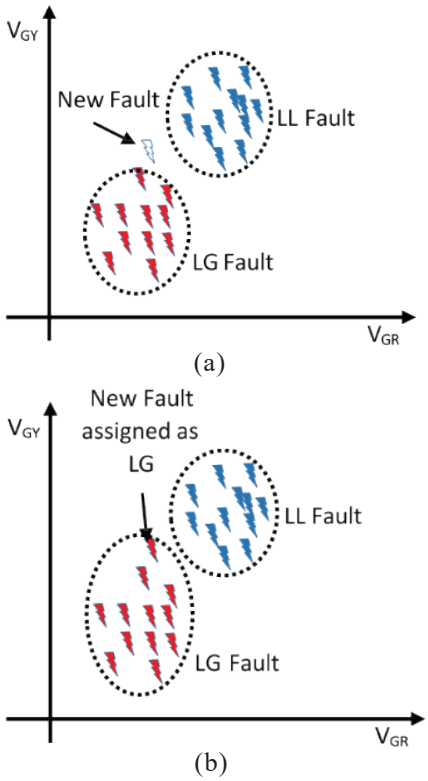


Figure 5. New fault classification using K-Nearest neighbour, (a) Before K-NN; and (b) After K-NN.

4.2 Decision Tree

The decision tree is a straightforward and easily visualized algorithm that can handle both linear and non-linear data without requiring scaling[9]. It operates as a tree-structured classifier, making decisions by traversing through a series of nodes and branches, where each branch represents a possible action based on true or false decisions. While the decision tree is primarily used for classification tasks, it can also be applied to regression problems; however, it is generally more effective for classification. A training dataset with a high degree of entropy is fed into the root node, where information gain is applied to reduce the entropy as the data progresses through the tree. A decision tree consists of two types of nodes: decision nodes and leaf nodes. The root node is initially provided with the entire dataset, and based on the decisions made, the data is split into subsets with lower entropy. This process continues until a leaf node, which represents the final decision, is reached. The fault classification process using a decision tree is illustrated in Fig. 6. In this figure, a comparison between a normal tree and a decision tree algorithm is also shown, where the data is initially classified into symmetrical and unsymmetrical faults and then further categorized into specific fault types.

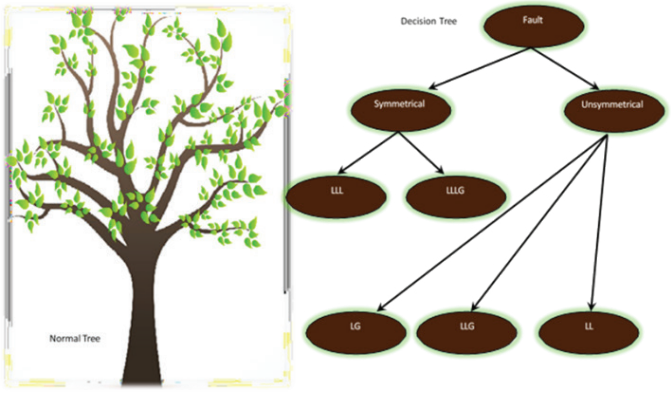


Figure 6. Fault classification using decision tree algorithm.

4.3 Random Forest

The Random Forest (RF)[13], is essentially an extension of the decision tree, consisting of a collection of decision trees, hence the term 'forest'. More specifically, an RF is an ensemble classifier that utilizes the decision tree algorithm in a randomized manner. Unlike a single decision tree, where all variables are considered, an RF uses only a subset of the variables at each step to make node decisions. Initially, a bootstrap dataset is generated with randomness introduced in two ways: first, by randomly selecting samples to include in the bootstrap dataset, and second, by randomly selecting a pair of attributes at each step and choosing one of them to place at a node. The decision regarding which attributes to place at a node is made by carefully evaluating which attribute will better split the data. In essence, the RF acts as a decision-making ensemble, where multiple trees vote on the best classification, with the most frequently selected option being chosen as the final classification.

4.4 Bagging Classifier

The Bagging classifier, also known as bootstrap aggregation, employs a method where 'X' datasets are randomly selected with replacement from the original dataset, forming what is referred to as a bootstrap sample set[33]. These bootstrap sample sets are then used to train individual classifiers.

As a result, 'X' trained classifiers are generated, and these classifiers are ultimately combined to create a final ensemble model. This ensemble model makes decisions by leveraging a voting mechanism across all the individual classifiers.

4.5 Naive Bayes

Naive Bayes[34] is a machine learning algorithm used in supervised learning, which essentially employs Bayes' probability theorem. The fundamental assumption of this method is that the variables used are all naive, or independent of each other, which is a prerequisite for applying the theorem. Although the Naive Bayes method is a simple learning algorithm, the independence criterion is often violated in practice. Bayes' theorem is mathematically stated as follow:

$$P(A/B) = \frac{P(\mathrm{A}\bigcap B)}{P(B)} = \frac{P(A).P(B/A)}{P(B)} \quad (9)$$

where

P(A): The probability of A occurring
P(B): The probability of B occurring
P (A/B): The probability of A given B P(($B \mid A$): The probability of B given A
P ($A \cap B$): The probability of both A and B occurring

5. EXPERIMENTAL RESULTS & DISCUSSION

The proposed distribution network model of Fig. 1 has been developed, consisting of two identical distributed generators (DG1 and DG2), a grid, and a 100 km distribution line, as shown in Fig. 3. The voltage and current of the system under normal (healthy) conditions are depicted in Fig. 7. Since the voltage and current of DG2 are identical to those of DG1, they are not shown in this figure. These voltages and currents are sinusoidal and do not exhibit any faults. This model was used to simulate ten types of faults (RG, YG, BG; RY, RB, YB; RY-G, RB-G, YB-G; RYB-G;). These simulated faults were evenly distributed along the 100 km distribution line at 2 km intervals.

The output waveform for the most severe fault (RYB-G) at the 2 km point on the distribution line is shown in Fig. 8. As mentioned earlier, the voltage and current waveforms for DG2 are identical to those of DG1, so the waveform for DG2 is not shown here. During the short-circuit condition, the grid voltage tends to drop to almost zero, as shown in Fig. 8(a) while the current becomes very high, as depicted in Fig. 8(b). Similarly, the DG1 voltage drops to zero during the fault, and the current waveform becomes distorted, as shown in Fig. 8(c) and Fig. 8(d). The grid voltage and current for all three phases and the voltages and currents for all three phases of DG1 and DG2 were recorded for ten different fault types at 2 km intervals along the distribution line.

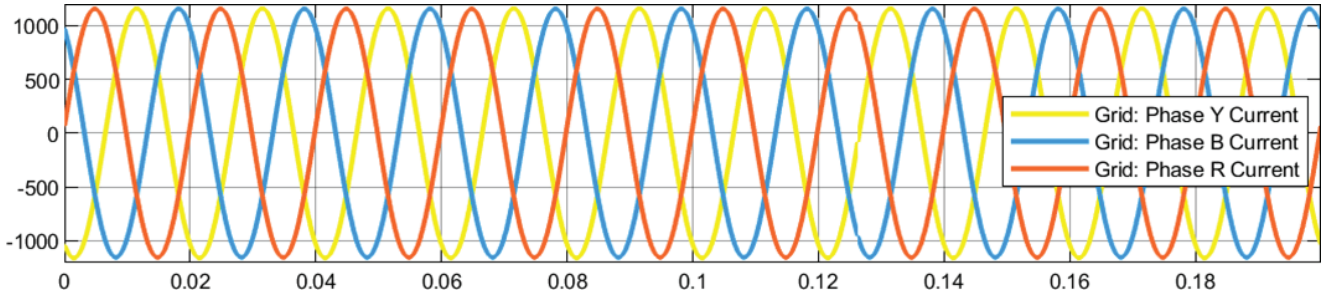
This process resulted in nine fault voltage samples and nine fault current samples for each fault type, yielding 490 samples of 18 data points each. These samples were then used in the proposed machine-learning model. Scatter plots were generated for each fault's voltage and current values to gain a more comprehensive understanding of the fault data. These scatter plots, shown in Fig. 9, visually confirm the absence of outliers in the collected fault data. After confirming the absence of outliers in our recorded data, the correlated fault voltages and currents were transformed using Principal Components (PC). To optimize the results, we considered a PC range from 1 to 8 and employed various classification algorithms, including K-Nearest Neighbour (K-NN), Decision Tree, Random Forest Classifier, Bagging Classifier, and Naive Bayes. For the K-NN algorithm, the nearest neighbor parameter K was varied from 1 to 9. It was observed that when K = 1 and PC = 4, 5, or 6, the training and test accuracy of K-NN reached 100 %, as shown in Fig. 10. Therefore, K = 1 was selected for K-NN.

A comparison of classification accuracy using the proposed method for different PC values, alongside classification without transformation for various algorithms, is summarized in Table 1. Without data transformation, classification accuracy ranges from 94 % to 99 %, with K- NN and Random Forest Classifier achieving accuracies of 99 % and 98 %, respectively. However, with the proposed method—specifically for K = 1 and PC = 4—K-NN reaches 100 % accuracy, while the other algorithms attain a minimum accuracy of 95 %. Notably, employing PC = 6, as indicated in Table 1, further enhances the classification accuracy, achieving a minimum accuracy of 98 %. These accuracies are also presented in Fig. 11. As illustrated in the figure, all algorithms achieve nearly 100 per cent accuracy when the Principal Component (PC) is equal to four or higher. Notably, K-NN achieves 100 % accuracy for both test and training data when K equals 1 and PC is set to 4. For further clarification, the confusion matrices illustrating the correspondence between true faults and predicted faults for the first two classification algorithms are displayed in Fig. 12.

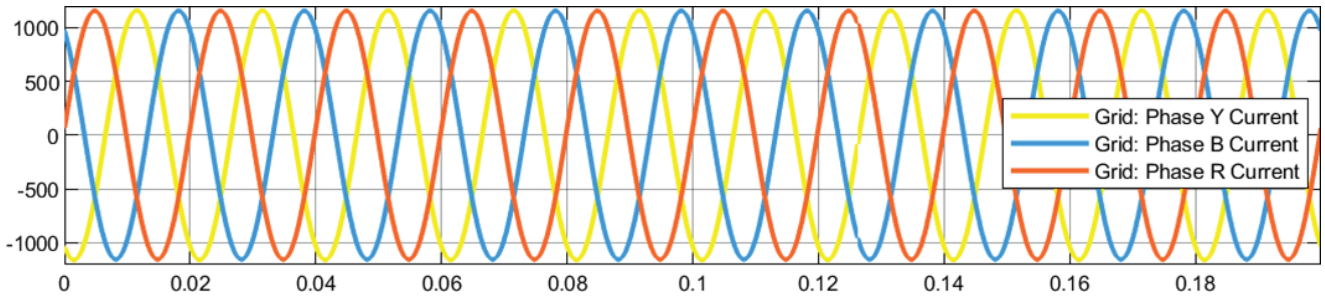
It is evident that the K-NN algorithm accurately classifies both true and predicted values, as shown in Fig. 12(a). Conversely, the Decision Tree algorithm achieves 98 % accuracy, with only three fault values misclassified, as shown in Fig. 12(b). Hence, the results suggest that the minimum value of PC can be chosen as 4 with K = 1. However, when PC = 6 and K = 1 in K-NN, it attains 100 % accuracy. The

Table 1. Accuracy of different classifiers with and without data transformation

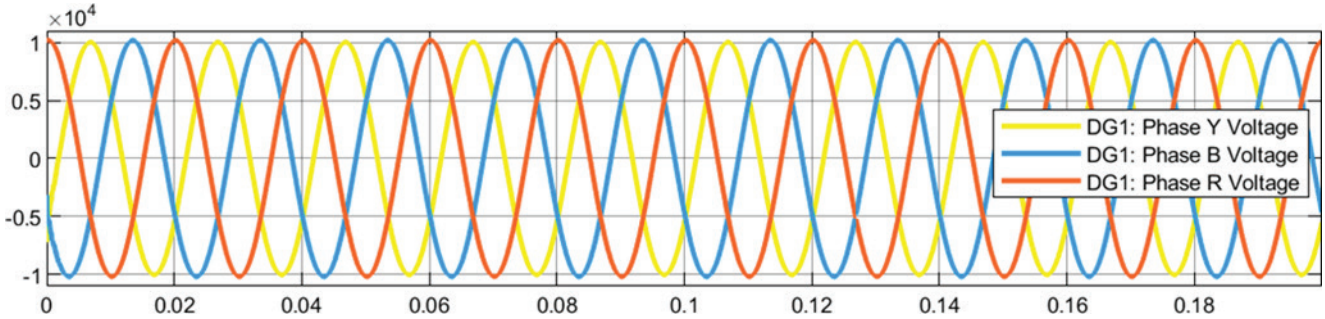
Algorithm	Accuracy without transformation	Accuracy after data transformation (using PC value as)							
		1	2	3	4	5	6	7	8
K-nearest neighbors	98.98	54.08	83.67	84.69	1	1	1	1	98.98
Decision tree	94.90	54.08	74.49	81.63	95.92	96.94	97.96	97.96	97.96
Random forest classifier	97.96	54.08	76.53	85.71	97.96	98.98	1	1	98.98
Bagging classifier	97.98	37.76	58.16	70.41	98.99	97.96	98.98	98.98	98.98
Naive bayes	93.88	32.65	50.00	60.20	96.94	95.92	97.96	97.96	97.96



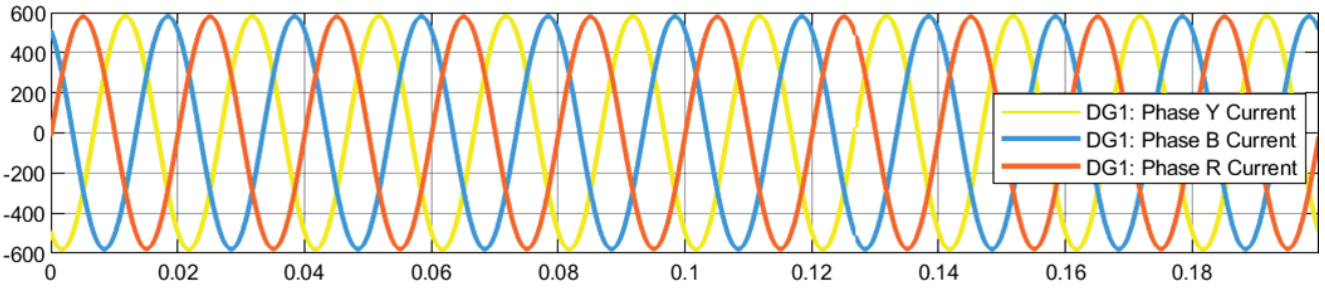
(a)



(b)

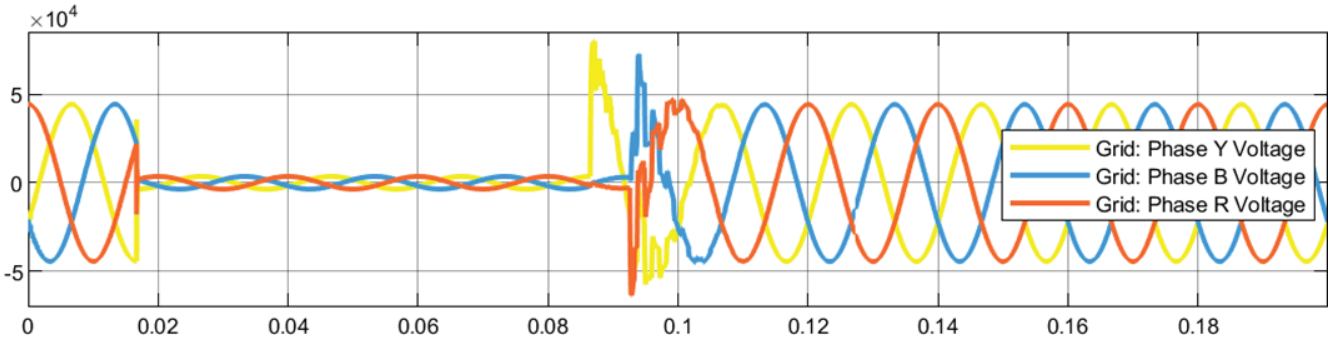


(c)



(d)

Figure 7. Grid, DG1 voltage and currents under normal conditions, (a) Grid voltage: Normal conditions; (b) Grid current: Normal conditions; (c) DG1 voltage: Normal conditions; and (d) DG1 current: Normal conditions.



(a)

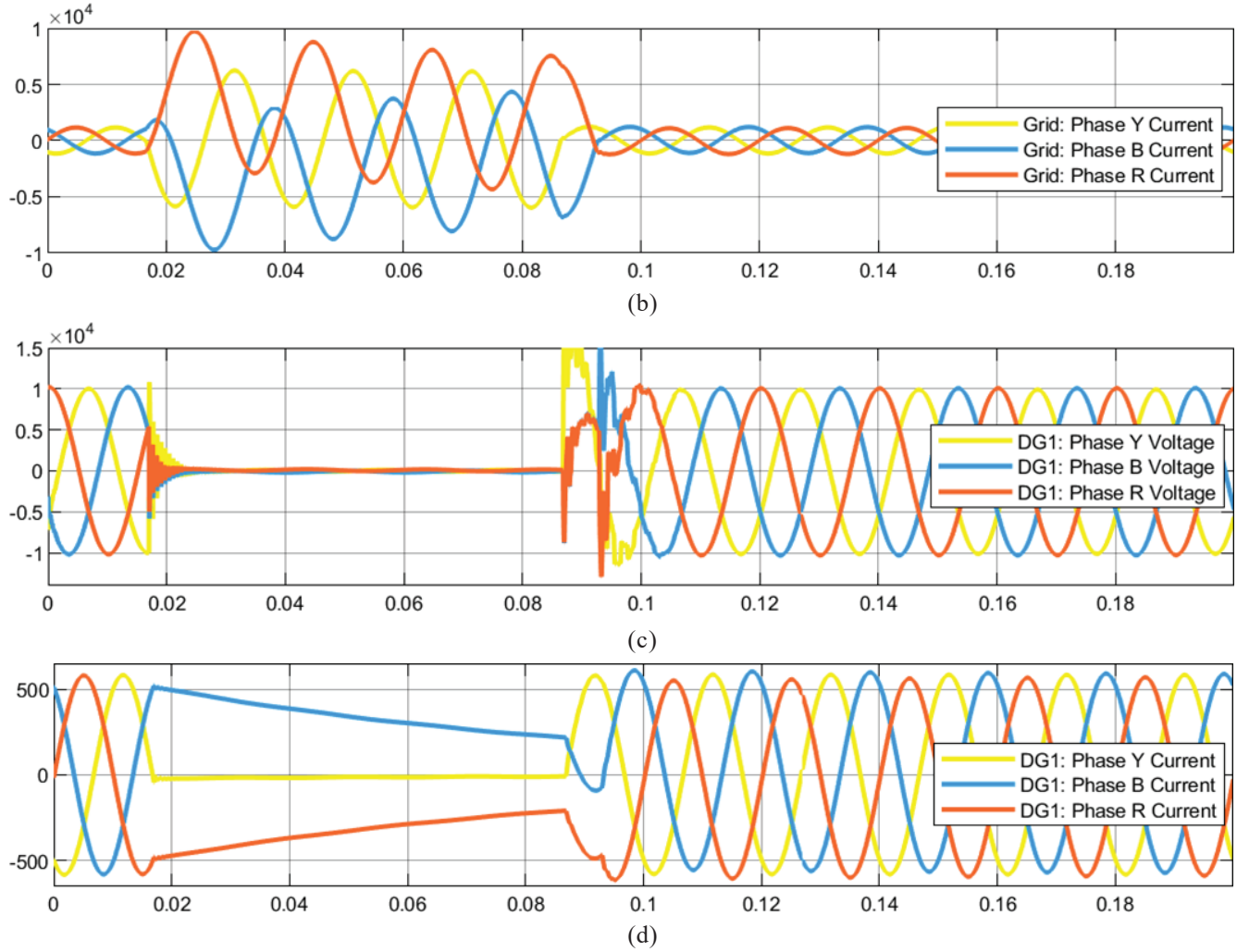
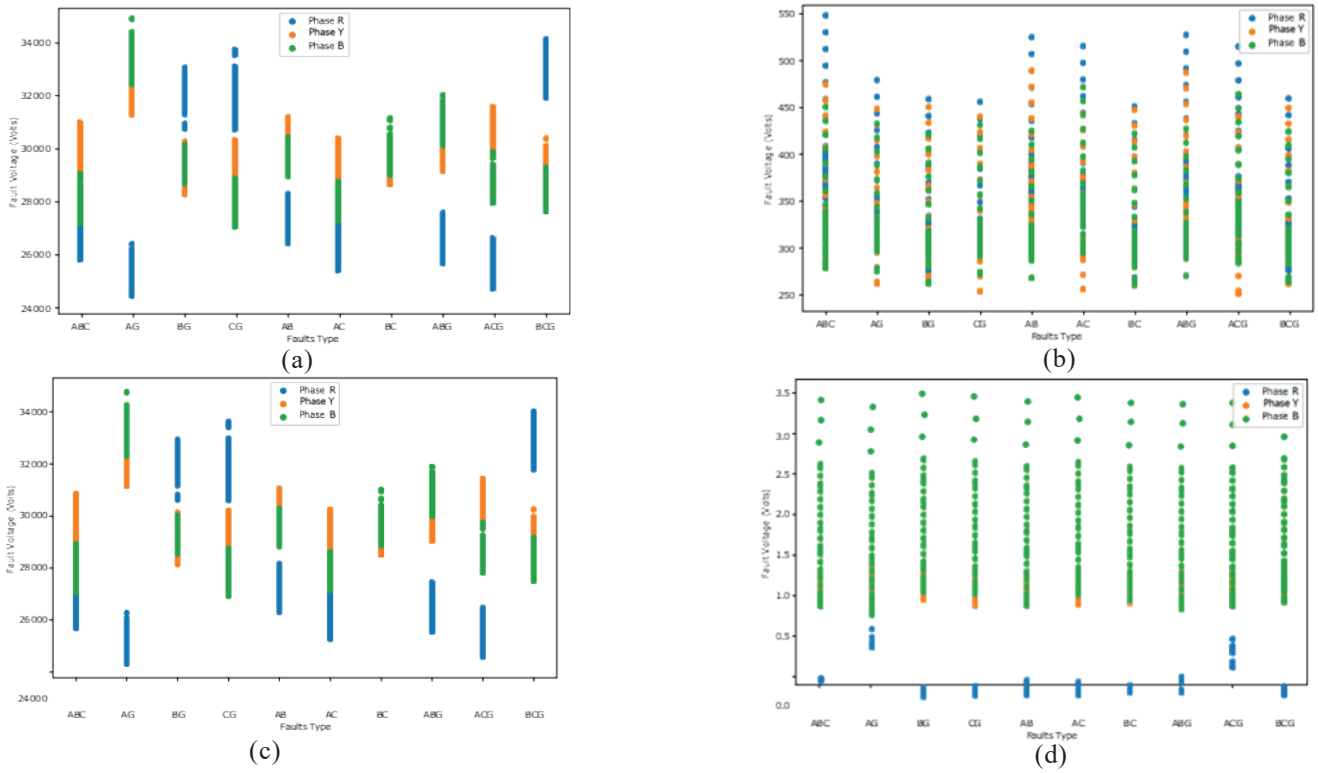
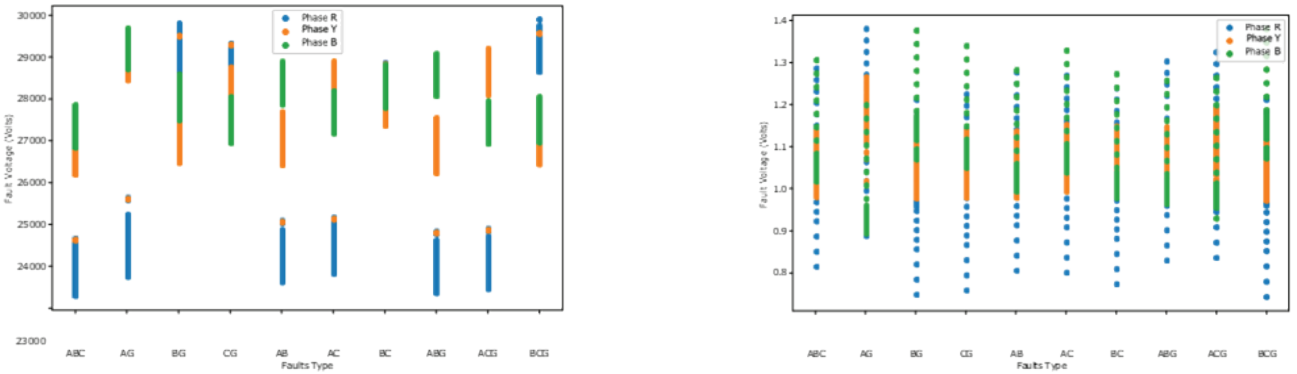


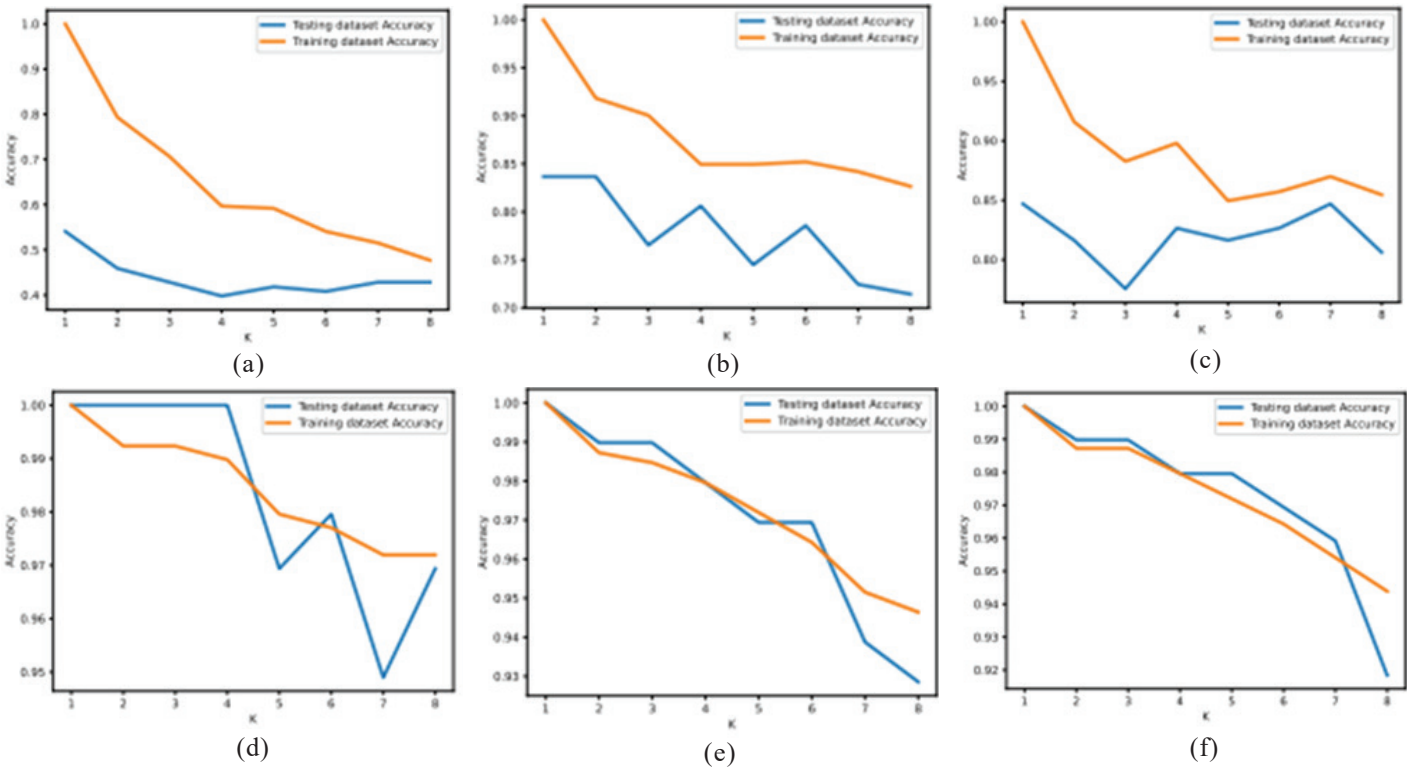
Figure 8. Grid, DG1 voltage and currents under fault conditions, (a) Grid voltage: LLLG fault conditions; (b) Grid current: LLLG fault conditions; (c) DG1 voltage: LLLG fault conditions; and (d) DG1 current: LLLG fault conditions.





(e) (f)

Figure 9. Scatter plot of fault voltages and currents for grid (Phases R, Y, B), DG1 (Phases R, Y, B), and DG2 (Phases R, Y, B), (a) Grid fault voltages: Phase R, Y, B; (b) Grid fault currents: Phase R, Y, B; (c) DG1 fault voltages: Phase R, Y, B; (d) DG1 fault currents: Phase R, Y, B; (e) DG2 fault voltages: Phase R, Y, B; and (f) DG2 fault currents: Phase R, Y, B.



(a) (b) (c)

(d) (e) (f)

Figure 10. Variation of testing and training data accuracy with different values of K in K-NN, (a) PC=1: Accuracy vs K; (b) PC=2: Accuracy vs K; (c) PC=3: Accuracy vs K; (d) PC=4: Accuracy vs K; (e) PC=5: Accuracy vs K; and (f) PC=6: Accuracy vs K.

random forest classifier also achieves 100 % accuracy, and the performance of other algorithms improves as well. Furthermore, the proposed method reduces the dimensionality of the original dataset from 490×18 to 490×6 through fault data transformation. In essence, it utilizes only 6 features instead of all 18. These findings indicate that fault data transformation using PCA can be a valuable technique for enhancing the performance of specific classification algorithms in power system data analysis.

6. CONCLUSION

This study explored a novel fault data transformation technique leveraging machine learning (ML) to enhance fault classification accuracy in distributed power systems. The research involved developing a comprehensive distribution network model, comprising two identical distributed generators, a grid, and a 100 km distribution line. Ten different faults were introduced into the system at every 2 km of the distribution line, and various voltages and currents were recorded to classify the faults using different ML techniques. The resulting data formed a 490 × 18 matrix, which was reduced to 490 × 6 using the proposed data transformation technique with Principal Component Analysis (PCA). The introduction of PCA as a data transformation tool proved to be a crucial factor

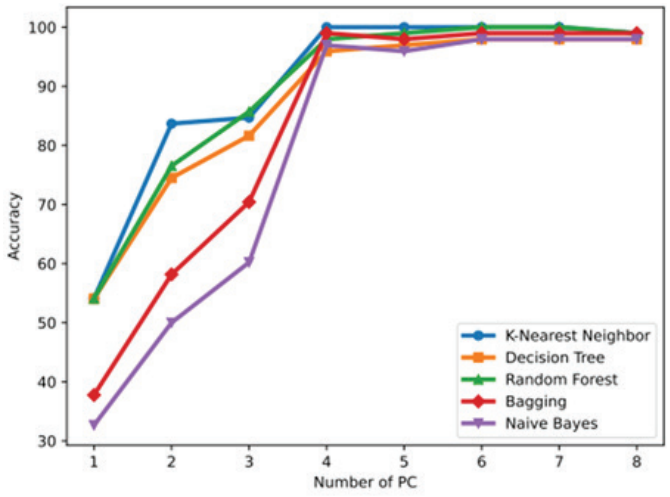
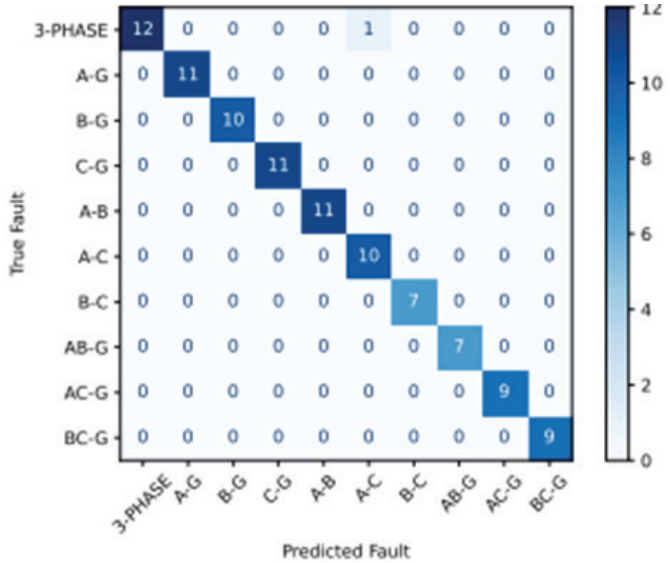
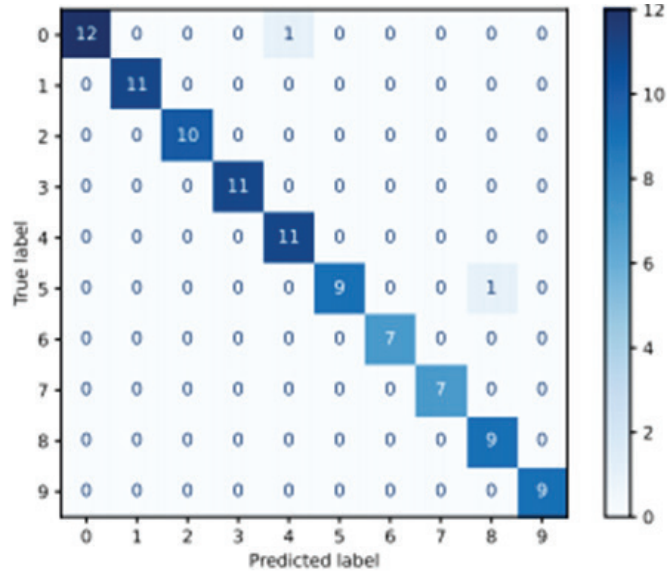


Figure 11. Classification accuracy after data transformation with varying principal components.



(a)



(b)

Figure 12. Confusion matrices, (a) K-NN; and (b) Decision tree.

in improving the performance of various classification algorithms. Our key findings underscore the following:

- **K-Nearest Neighbours (K-NN):** In the absence of PCA, K-NN exhibited an accuracy of 98.98 %. After data transformation using PCA with four principal components (PCs), the accuracy significantly improved, achieving a near-perfect 100 %. When PCA was further extended to six PCs, the accuracy remained consistently high at 100 %, indicating the potency of data transformation in refining K-NN's classification performance.
- **Decision Tree:** The unaided Decision Tree algorithm demonstrated an accuracy of 94.90 %. With data transformation using four PCs, the accuracy surged to 95.92 %, and with six PCs, it further improved to an impressive 97.96 %.
- **Random Forest Classifier:** The original Random Forest Classifier displayed an accuracy of 97.96 %. The introduction of data transformation using PCA with four PCs preserved the accuracy at 97.96 %, and with six PCs, it achieved a perfect classification accuracy of 100 %.
- **Bagging Classifier:** The Bagging Classifier achieved an accuracy of 97.98 %. With the transformation integrating four PCs, the accuracy improved to 98.99 %. When six PCs were employed, the accuracy remained high at 98.98 %.
- **Naive Bayes:** In its original state, Naive Bayes attained an accuracy of 93.88 %. The inclusion of data transformation using PCA with four PCs resulted in an increased accuracy of 96.94 %, with a further enhancement observed when six PCs were applied, reaching 97.96 %.

This analysis highlights the substantial influence of fault data transformation (voltages and currents) using PCA on elevating classification accuracy across multiple algorithms. The research reaffirms that data transformation serves as a valuable tool to enhance classification accuracy and reduce dimensionality, promising to optimize the performance of a range of algorithms in the realm of power system data analysis.

REFERENCES

1. Chiradeja P, Ramakumar R. An approach to quantify the technical benefits of distributed generation. IEEE Transactions on Energy Conversion. 2004 Dec;19(4):764–73.
2. Patrick, Zhu J, Qiu ZW, Wing, Yeung DS. Comparision of different classifiers in fault detection in microgrid. International Conference on Machine Learning and Cybernetics. 2011 Jul 1;1210–3.
3. Okoli C, Anyaka B, Nwokedi C, Anya V. Application of composite method for determining fault location on electrical power distribution lines. Journal of Electrical and Computer Engineering. 2020 Sep 1;2020:1–9.
4. Kezunovic M. Smart fault location for smart grids. IEEE Transactions on Smart Grid. 2011 Mar;2(1):11–22.
5. Swetapadma A, Yadav A. Fuzzy inference system approach for locating series, shunt, and simultaneous series-shunt faults in double circuit transmission lines. Computational

Intelligence and Neuroscience. 2015;2015:1–12.
6. Gururajapathy SS, Mokhlis H, Illias HA. Fault location and detection techniques in power distribution systems with distributed generation: A review. Renewable and Sustainable Energy Reviews [Internet]. 2017 Jul;74:949–58. Available from: https://www.sciencedirect.com/science/article/pii/S1364032117303386
7. Lee H, Mousa AM. GPS travelling wave fault locator systems: investigation into the anomalous measurements related to lightning strikes. IEEE Transactions on Power Delivery. 1996 Jul;11(3):1214–23.
8. bao JZ, Mao ZTY. A fault location and realization method for overhead high voltage power transmission. procedia engineering. 2011;15:964–8.
9. P.-L. Tu, J.-Y. Chung, A new decision-tree classification algorithm for machine learning. In Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92, Arlington, VA, USA: IEEE Comput. Soc. Press, 1992, 370–377. doi: 10.1109/TAI.1992.246431.
10. G. Ma, L. Jiang, K. Zhou, G. Xu. A method of line fault location based on traveling wave theory. *IJCA*, vol. 9, no. 2, pp. 261–270, Feb. 2016, doi: 10.14257/ijca.2016.9.2.25.
11. Sant MT, Y.G. Paithankar. Online digital fault locator for overhead transmission line. Proceedings of the Institution of Electrical Engineers. 1979 Jan 1;126(11):1181–1.
12. Lin S, He ZY, Li XP, Qian QQ. Travelling wave time–frequency characteristic-based fault location method for transmission lines. IET Generation, Transmission & Distribution. 2012;6(8):764.
13. N. M. Abdulkareem, A. M. Abdulazeez. Machine learning classification based on radom forest algorithm: A review. International Journal of Science and Business, 2021; 5(2): 128–142.
14. Chowdhury BH, Kunyu Wang. Fault classification using Kohonen feature mapping. Proceedings of International Conference on Intelligent System Application to Power Systems. :194–8.
15. Chowdhury FN, Aravena JL. A modular methodology for fast fault detection and classification in power systems. IEEE Transactions on Control Systems Technology. 1998;6(5):623–34.
16. Ferrero A, Sangiovanni S, Zappitelli E. A fuzzy-set approach to fault-type identification in digital relaying. IEEE Transactions on Power Delivery. 1995;10(1):169–75.
17. V. Malathi, N.S. Marimuthu. Multi-class support vector machine approach for fault classification in power transmission line. 2008 Nov 1;
18. Awasthi S, Singh G, Ahamad N. Identification of type of a fault in distribution system using shallow neural network with distributed generation. Energy Engineering. 2023 Jan 1;120(4):811–29.
19. Alsafasfeh QH, Abdel-Qader I, Harb AM. Fault classification and localization in power systems using fault signatures and principal components analysis. Energy and Power Engineering. 2012;04(06):506–22.
20. Ajagekar A, You F. Quantum computing based hybrid deep learning for fault diagnosis in electrical power systems. Applied Energy. 2021 Dec;303:117628.
21. Babayomi OO, Oluseyi PO. Intelligent fault diagnosis in a power distribution network. Advances in Electrical Engineering. 2016 Oct 19;2016:1–10.
22. Salat R, Osowski S. Accurate Fault location in the power transmission line using support vector machine approach. IEEE Transactions on Power Systems. 2004 May;19(2):979–86.
23. Agrawal R, D. Thukaram. Identification of fault location in power distribution system with distributed generation using support vector machines. 2013 Feb 1.
24. El-Tawab S, Mohamed HS, Amr Refky, Abdel-Aziz AM. Self-Healing of active distribution networks by accurate fault detection, classification, and location. Journal of Electrical and Computer Engineering. 2022 Mar 31;2022:1–14.
25. S. Awasthi. Identification of fault and its location using Neural Network in Distribution Network with Distribution GenerationA. Ph.D. thesis, Indian Institute of Technology, Roorkee, 2014.
26. Zayandehroodi H, Mohamed A, Farhoodnea M, Mohammadjafari M. An optimal radial basis function neural network for fault location in a distribution network with high penetration of DG units. Measurement. 2013 Nov;46(9):3319–27. Accessed: Mar. 12, 2025. [Online]. Available: https://www.academia.edu/14323393/An_optimal_radial_basis_function_neural_network_for_fault_location_in_a_distribution_network_with_high_penetration_of_DG_units
27. R.C. Dugan, T. E. McDermott. Operating conflicts for distributed generation on distribution systems. In 2001 Rural Electric Power Conference. Papers Presented at the 45th Annual Conference (Cat. No.01CH37214). Little Rock, AR, USA: IEEE, 2001, p. A3/1-A3/6. doi: 10.1109/REPCON.2001.949511.
28. Hernández JC, De la Cruz J, Ogayar B. Electrical protection for the grid-interconnection of photovoltaic-distributed generation. Electric Power Systems Research. 2012 Aug;89:85–99.
29. Yang H, Liu X, Guo Y, Zhang P. Fault location of active distribution networks based on the golden section method. Mathematical Problems in Engineering. 2020 Feb 8;2020:1–9. doi: 10.1155/2020/6937319.
30. Masoud Najafzadeh, Jaber Pouladi, Daghigh A, Beiza J, Shahmohamadi S. A new method for fault detection and location in a low-resistance grounded power distribution network using voltage phasor of D-PMUs data. International Transactions on Electrical Energy Systems. 2023 Jul 22;2023:1–17.doi: 10.1155/2023/1754305.
31. Singh D, Singh B. Feature wise normalization: An effective way of normalizing data. Pattern Recognition. 2022 Feb;122:108307. doi: 10.1016/j.patcog.2021.108307.
32. Boateng EY, Otoo J, Abaye DA. Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: A review. Journal of Data Analysis and Information

Processing. 2020;08(04):341–57.
doi: 10.4236/jdaip.2020.84020.
33. Yaman E, Subasi A. Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. BioMed Research International. 2019 Oct 31;2019:1–13.
doi: 10.1155/2019/9152506.
34. Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. 2010, doi: 10.1007/978-0-387-30164-8.

CONTRIBUTORS

Mr Kamlesh Singh Bisht is pursuing his PhD from DIT University, Dehradun and working as EMU in DRDO
In the present study he has actively contributed to the literature review, writing of the paper and compilation of data.

Dr Nafees Ahamad obtained PhD from the Department of Electrical Engineering at DIT University, India. He is working as a Assistant Professor and Associate Head at Department of Electrical, and Electronic Communication Engineering, DIT University, India. His areas of interest include: Control systems, order reduction, controller design, system engineering, control and automation, electrical machine design, and optimization. In the present study he involved in the conceptualization, applying ML algorithm and finalization of this paper

Mr Saurabh Awasthi obtained MTech in Alternate hydro energy systems from IIT Roorkee and working as an Assistant Professor in the Department of Electrical & Electronics & Communication Engineering, DIT University Dehradun. His area of interest lies in Distributed generation, smart grid, renewable energy systems, power system operation & protection.
His contribution to the paper was towards data generation for fault analysis for further ML application.