RANDOM FOREST-BASED TRANSMISSION-LINE FAULT DETECTION AND CLASSIFICATION USING MEASURED AND DERIVED ELECTRICAL FEATURES

Miss. IPSARANI RATH
B. Tech Scholar
Department of Electrical and Electronics Engineering
Roland Institute of Technology, Berhampur

Mr. GOURI SANKAR PANDA
Assistant Professor
Department of Electrical and Electronics Engineering
Roland Institute of Technology, Berhampur

Mr. HARDEEP KUMAR SETHI
B. Tech Scholar
Department of Electrical and Electronics Engineering
Roland Institute of Technology, Berhampur

Mr. SRIDHAR PATRO
B. Tech Scholar
Department of Electrical and Electronics Engineering
Roland Institute of Technology, Berhampur

Mr. SUNIL KUMAR BADATYA
B. Tech Scholar
Department of Electrical and Electronics Engineering
Roland Institute of Technology, Berhampur

Abstract - Transmission lines are critical components of modern power systems, and faults occurring on these lines can lead to equipment damage, power interruptions, voltage disturbances, and system instability if they are not detected and cleared rapidly. Conventional protection techniques based on electrical quantities such as voltage, current, and impedance are reliable, but their performance can be affected by changing operating conditions, fault resistance, power swings, noise, and variations in network configuration. This work presents an AI-based fault detection and classification system for transmission lines using machine-learning techniques. A dataset containing 5000 samples and 15 electrical input features is considered, representing Normal, LG, LL, LLG, and LLL operating conditions. The input parameters include three-phase voltage and current measurements, RMS voltage and current, line-to-line quantities, and apparent impedance. A Random Forest classifier with 200 decision trees is trained using 80% of the dataset and evaluated using the remaining 20%. The developed model achieves an overall accuracy of 97.90%, demonstrating effective identification of different transmission-line conditions. The model achieves particularly strong performance for Normal, LG, and LLL conditions, while minor confusion is observed between LL and LLG faults. Feature-importance analysis indicates that apparent impedance, RMS voltage, and RMS current are among the most influential parameters for fault classification. The trained model is also capable of predicting the condition of a new transmission-line sample along with the probability associated with each fault class. The results demonstrate that machine-learning-based fault classification can provide a fast, accurate, and intelligent approach for transmission-line protection, monitoring, and fault diagnosis, with potential applications in modern smart-grid and digital protection systems.

Keywords— Transmission-line protection; fault detection; fault classification; Random Forest; machine learning; electrical feature analysis; apparent impedance.

I. INTRODUCTION

The transmission line is one of the most important components of a modern power system, responsible for transferring electrical energy from generating stations to distribution networks over long distances. Due to their extensive geographical exposure and continuous operation under varying environmental and loading conditions, transmission lines are susceptible to different types of faults such as single line-to-ground, line-to-line, double line-to-ground, and three-phase faults as shown in Fig. 1. These faults can cause excessive currents, voltage disturbances, equipment damage, power outages, and instability in the power system if they are not detected and cleared rapidly. Therefore, accurate and fast fault detection and classification are essential for maintaining the reliability, stability, and security of electrical power systems.[1]



Figure 1: A TYPICAL EXAMPLE OF A FAULT IN TRANSMISSION LINE

Conventional transmission-line protection systems primarily use relay-based techniques that analyze electrical quantities such as voltage, current, impedance, and phase angles to identify abnormal operating conditions. Although these methods are reliable and widely implemented, their performance can be affected by fault resistance, fault location, system operating conditions, power

swings, noise, and variations in renewable energy penetration. In addition, conventional methods may have limitations in distinguishing between different fault types under complex and rapidly changing grid conditions. The increasing complexity of modern power systems has therefore created a need for more intelligent and adaptive fault diagnosis techniques.

Artificial Intelligence (AI) provides an effective approach for addressing these challenges by enabling protection systems to learn patterns from large amounts of electrical data. AI-based techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Random Forests, and deep learning algorithms can process voltage and current signals and identify fault-related patterns with high accuracy. These methods can be trained using simulated or real-time fault data to detect the occurrence of a fault and classify its type rapidly. AI-based approaches can also improve their capability to handle nonlinear relationships, noisy measurements, changing system conditions, and complex fault scenarios.[2]

The proposed work, titled "AI-Based Fault Detection and Classification in Transmission Lines," focuses on developing an intelligent protection framework capable of detecting transmission-line faults and accurately classifying their types using electrical system parameters. The system can utilize measured voltage and current signals as input features and apply an appropriate AI algorithm to distinguish normal operating conditions from faulty conditions. Once a fault is detected, the trained model can classify whether the fault is single line-to-ground, line-to-line, double line-to-ground, or three-phase. Such an intelligent system can contribute to faster fault identification, improved protection performance, reduced equipment damage, and enhanced reliability of the overall power transmission network.

The integration of AI with conventional protection techniques represents a significant step toward the development of intelligent and adaptive power-system protection. By combining real-time electrical measurements with machine-learning capabilities, the proposed approach has the potential to provide rapid, accurate, and reliable fault diagnosis. This makes AI-based fault detection and classification a promising technology for modern smart grids, where reliable operation, fast decision-making, and improved system resilience are increasingly important.[3]

II. LITERATURE REVIEW

The detection and classification of faults in transmission lines have been extensively studied because faults can significantly affect the reliability, stability, and security of electrical power systems. Conventional protection techniques generally depend on overcurrent, distance, differential, and impedance-based relays. Among these, distance protection is widely used for transmission lines because it estimates the apparent impedance between the relay location and the fault point. The basic relationship is given by:

$$Z_{app} = \frac{V}{I} \tag{1}$$

where Z_{app} is the apparent impedance, V is the measured voltage, and I is the measured current. Under fault conditions, the apparent impedance changes significantly, allowing the relay to determine whether a fault has occurred. However, the accuracy of conventional impedance-based methods can be affected by fault resistance, power swings, source impedance variation, and changing operating conditions.[4]

With the development of digital protection systems, signal-processing techniques have been increasingly applied to transmission-line fault analysis. Methods such as Fourier Transform, Wavelet Transform, and Short-Time Fourier Transform are used to extract useful information from voltage and current signals. The Discrete Fourier Transform (DFT), for example, can be expressed as:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \tag{2}$$

where x(n) represents the sampled electrical signal, N is the number of samples, and X(k) represents its frequency-domain components. Although these techniques are effective in extracting fault features, their performance may depend strongly on signal quality, sampling frequency, and the selection of suitable features.

Wavelet-based approaches have gained considerable attention because transmission-line fault signals are non-stationary and contain transient components immediately after fault inception. The Continuous Wavelet Transform can be represented as:

$$W(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{3}$$

where a is the scale parameter, b is the translation parameter, x(t) is the input signal, and ψ(t) is the mother wavelet. Wavelet-based methods can capture both time and frequency characteristics of fault transients, making them suitable for fault detection and

classification. However, selecting an appropriate mother wavelet and decomposition level can influence the performance of the protection system.[5]

The introduction of Artificial Intelligence (AI) has provided new opportunities for improving transmission-line fault diagnosis. Artificial Neural Networks (ANNs) have been widely investigated because they can learn complex nonlinear relationships between electrical measurements and fault conditions. A basic neuron can be represented by:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{4}$$

where x_i represents the input features, w_i represents the corresponding weights, b is the bias, f is the activation function, and y is the output. During training, the network adjusts its weights to minimize the difference between the predicted and actual fault classes. ANN-based systems can provide rapid classification once trained, although their performance depends on the quality and diversity of the training dataset.[6]

Support Vector Machine (SVM) techniques have also been applied to transmission-line fault classification. SVM attempts to determine an optimal decision boundary between different fault classes. For a linearly separable dataset, the decision function can be expressed as:

$$f(x) = w^T x + b \tag{5}$$

where w is the weight vector and b are the bias. Kernel functions can be used when the fault classes are not linearly separable. SVM-based methods are particularly useful when the available training dataset is relatively small, but selecting appropriate kernel parameters and input features is important for achieving high classification accuracy.

Decision-tree and ensemble learning techniques such as Random Forest have also been explored for power-system fault classification. These methods can classify faults based on combinations of voltage, current, sequence components, or extracted transient features. Their major advantage is their ability to handle nonlinear relationships and multiple input variables without requiring extensive mathematical modelling of the transmission system. However, their performance is influenced by the quality of the training data and the representation of different fault conditions.

More recently, deep-learning approaches have attracted significant attention because they can automatically extract relevant features from raw or minimally processed electrical signals. Convolutional Neural Networks (CNNs), for example, can identify spatial and temporal patterns in transformed voltage and current signals. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly suitable for sequential electrical measurements because they can capture temporal dependencies in fault signals. A simplified LSTM structure uses input, forget, and output gates to regulate information flow through the network, making it suitable for analysing dynamic power-system behaviour.[7]

The performance of AI-based fault-classification systems is generally evaluated using metrics such as accuracy, precision, recall, and F1-score. Classification accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{6}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. For multi-class fault classification, a confusion matrix is commonly used to determine how effectively the AI model distinguishes between different fault categories.

Overall, the literature indicates a gradual transition from conventional relay-based protection toward intelligent protection systems combining signal processing, machine learning, and deep learning. Conventional methods provide reliable and well-established protection, while AI-based approaches offer improved adaptability and the ability to recognize complex fault patterns. However, challenges remain in obtaining representative datasets, handling noisy measurements, detecting high-resistance faults, maintaining performance under changing operating conditions, and ensuring that AI decisions can be implemented within the strict response-time requirements of protection systems. Therefore, the development of an accurate and computationally efficient AI-based fault detection and classification system remains an important research area for enhancing the reliability and resilience of modern transmission networks.[8]

III. PROBLEM STATEMENT

Transmission lines are continuously exposed to various electrical, environmental, and operational disturbances that can result in faults such as single line-to-ground, line-to-line, double line-to-ground, and three-phase faults. If these faults are not detected and classified rapidly, they may lead to equipment damage, power interruptions, voltage instability, and in severe cases, cascading failures in the power system. Conventional protection methods based on impedance, current, and voltage measurements are

effective under normal operating conditions but may face challenges in accurately identifying faults when fault resistance, loading conditions, source impedance, and network configurations vary. Therefore, there is a need for an intelligent protection system capable of rapidly detecting faults and accurately identifying their types under different operating conditions. The proposed work addresses this problem by developing an AI-based fault detection and classification system for transmission lines using electrical parameters such as voltage and current signals. The system aims to learn the characteristic patterns associated with different fault conditions and provide reliable fault identification with high accuracy and fast response.

IV. METHODOLOGY IMPLEMENTED

The methodology adopted for AI-Based Fault Detection and Classification in Transmission Lines consists of several sequential stages, beginning with the development of a transmission-line model and ending with the evaluation of the AI-based protection system. The overall methodology is designed to detect the occurrence of a fault and subsequently classify its type using electrical parameters such as three-phase voltage and current.

4.1. Transmission Line System Modelling

A suitable transmission-line model is developed using simulation software such as MATLAB/Simulink. The model consists of a three-phase source, transmission line, load, measurement units, and fault block. Different operating conditions are considered to represent both normal and faulty conditions. The transmission-line impedance can be represented as:

$$Z = R + jX \tag{7}$$

where R is the resistance and X is the reactance of the transmission line.[9]

4.2. Fault Generation

Different types of faults are introduced at selected locations on the transmission line. The major fault conditions considered are:

- Single Line-to-Ground (LG) fault
- Line-to-Line (LL) fault
- Double Line-to-Ground (LLG) fault
- Three-Phase (LLL) fault
- Three-Phase-to-Ground (LLLG) fault, if required

Fault conditions are simulated by varying parameters such as fault location, fault resistance, and fault initiation time. This generates a comprehensive dataset containing different fault scenarios.[10]

4.3. Data Acquisition

Three-phase voltage and current signals are measured at the selected observation point. For each operating condition, the corresponding voltage and current waveforms are recorded. The current and voltage relationship can be represented by:

$$I = \frac{V}{Z} \tag{8}$$

During a fault, the impedance seen by the system changes, resulting in significant variations in current and voltage. These variations provide important information for identifying the occurrence and type of fault.[11]

4.4. Signal Preprocessing

The collected electrical signals are pre-processed before being supplied to the AI model. This stage may include noise removal, normalization, sampling, segmentation, and conversion of signals into suitable input features. Normalization can be expressed as:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{9}$$

where x is the original input value, while x_{min} and x_{max} are the minimum and maximum values of the corresponding feature. Preprocessing improves the consistency of the input data and helps the AI model learn efficiently.[12]

4.5. Feature Extraction

Important characteristics are extracted from the voltage and current signals. The features may include RMS voltage, RMS current, peak values, phase angles, sequence components, voltage deviations, current deviations, and transient characteristics. For example, the RMS value of a periodic signal is given by:

$$X_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{10}$$

These extracted features are used to distinguish normal operating conditions from different fault conditions.[13]

4.6. Dataset Preparation

The generated data are divided into different categories according to the operating condition and fault type. The dataset is generally divided into training, validation, and testing datasets. The training dataset is used to develop the AI model, the validation dataset is used for parameter tuning, and the testing dataset is used to evaluate the final performance of the trained model.

4.7. AI Model Development

An appropriate machine-learning or deep-learning algorithm is developed for fault detection and classification. Algorithms such as Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree, Random Forest, or other suitable models can be considered. The AI model learns the relationship between the input electrical parameters and the corresponding fault classes.

For an ANN, the output of a neuron can be represented as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{11}$$

where x_i represents the input features, w_i represents the weights, b is the bias, and f is the activation function.[14]

4.8. Fault Detection

The first stage of AI-based diagnosis determines whether the transmission line is operating normally or under a faulty condition. The model receives the measured electrical parameters and produces a corresponding output. If the output indicates a fault, the system proceeds to the classification stage.

4.9. Fault Classification

After detecting a fault, the AI model identifies its specific type. The classifier distinguishes between LG, LL, LLG, LLL, and other predefined fault categories. This two-stage approach improves the interpretability of the protection system by separately determining whether a fault exists and what type of fault has occurred.[15]

4.10. Model Training and Optimization

The AI model is trained using the prepared dataset by minimizing the difference between the actual fault class and the predicted class. A general loss function for classification can be represented as:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{12}$$

where C is the number of classes, y_i is the actual class label, and $\hat{y}_i$ is the predicted probability for that class. Model parameters are adjusted iteratively to minimize the loss function and improve classification performance.

4.11. Performance Evaluation

The developed model is tested using previously unseen data. Its performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{13}$$

A confusion matrix is also used to determine whether the AI model is correctly distinguishing among the different fault categories. The final objective of this methodology is to develop an intelligent and fast fault-diagnosis system capable of accurately detecting and classifying transmission-line faults under different operating conditions.[16]

V. MODEL IMPLEMENTATION

The dataset used for the AI-Based Fault Detection and Classification in Transmission Lines consists of three-phase electrical measurements collected under normal and different fault operating conditions. The dataset contains 5,000 observations, with 1,000 observations corresponding to each operating condition: Normal, Single Line-to-Ground (LG) fault, Line-to-Line (LL) fault, Double Line-to-Ground (LLG) fault, and Three-Phase (LLL) fault. The dataset provides the necessary electrical parameters required for developing and evaluating a machine-learning-based transmission-line protection system.

Va	Vb	Vc	Ia	Ib	Ic	V_RMS	I_RMS	Vab	Vbc	Vca	Iab	Ibc	Ica	Z_App	Fault
1.025726	1.004909	1.008259	1.056542	0.983707	0.983661	1.013006	1.008555	0.020817	0.00335	0.017467	0.072835	4.57E-05	0.072881	1.004413	Normal
0.995759	0.947611	0.980158	0.978616	0.970966	0.980932	0.974716	0.976847	0.048148	0.032547	0.015601	0.00765	0.009967	0.002317	0.997818	Normal
0.983095	1.020741	0.976845	0.990117	0.998432	0.966832	0.99375	0.985218	0.037646	0.043896	0.00625	0.008315	0.031599	0.023284	1.00866	Normal
1.003021	0.957796	0.958651	0.991509	1.012941	1.026283	0.973385	1.010346	0.045225	0.000855	0.04437	0.021432	0.013342	0.034775	0.963417	Normal
1.017182	0.974052	0.99809	0.982263	0.986318	1.037861	0.996598	1.002467	0.04313	0.024038	0.019093	0.004054	0.051544	0.055598	0.994145	Normal
0.989696	1.006322	0.98149	0.951211	1.031604	1.071448	0.992556	1.019315	0.016626	0.024832	0.008207	0.080392	0.039844	0.120237	0.973748	Normal
1.000201	1.011417	0.945408	1.031799	1.032169	0.980664	0.986097	1.015166	0.011216	0.066009	0.054793	0.00037	0.051505	0.051135	0.971366	Normal
0.984801	0.999651	1.011288	1.003309	0.976265	0.986128	0.998639	0.988631	0.01485	0.011636	0.026487	0.027044	0.009863	0.017182	1.010123	Normal
1.002495	0.997198	0.998489	1.001043	0.995263	0.990872	0.999397	0.995735	0.005297	0.001291	0.004006	0.005781	0.00439	0.010171	1.003678	Normal
1.003227	1.001527	0.961277	0.975831	1.024663	1.088936	0.988867	1.030851	0.001701	0.040249	0.04195	0.048832	0.064273	0.113105	0.959272	Normal
1.005915	0.976149	1.029052	0.947875	0.986592	1.067085	1.003939	1.001749	0.029766	0.052903	0.023137	0.038717	0.080492	0.11921	1.002187	Normal
0.986889	0.997163	0.983886	1.072641	0.944353	0.994031	0.989329	1.005063	0.010274	0.013277	0.003003	0.128288	0.049678	0.07861	0.984345	Normal
1.007703	1.019101	0.968461	0.965031	1.02152	0.994623	0.998657	0.993992	0.011398	0.05064	0.039242	0.056488	0.026897	0.029591	1.004693	Normal
1.048901	1.00127	0.985808	1.027952	0.987201	1.061548	1.012349	1.026018	0.047632	0.015462	0.063094	0.040751	0.074348	0.033596	0.986678	Normal
0.997859	0.993197	0.98234	1.004596	0.992394	1.064638	0.991153	1.021031	0.004662	0.010857	0.015519	0.012203	0.072244	0.060042	0.970738	Normal
1.017246	0.982129	0.970114	0.984177	0.984366	1.038548	0.990031	1.00269	0.035117	0.012014	0.047132	0.000189	0.054182	0.054371	0.987375	Normal
1.004864	0.963656	1.007041	1.022775	0.995145	1.025688	0.992054	1.014629	0.041208	0.043386	0.002178	0.02763	0.030542	0.002913	0.977751	Normal
0.978156	0.992599	1.019841	1.028441	1.009146	1.130761	0.997015	1.057464	0.014444	0.027242	0.041685	0.019295	0.121615	0.10232	0.942836	Normal
0.991406	0.979136	0.985573	1.024235	1.070725	0.922427	0.985385	1.007701	0.01227	0.006437	0.005833	0.04649	0.148298	0.101808	0.977855	Normal

Figure 2: DATASET USED FOR THE AI-BASED FAULT DETECTION AND CLASSIFICATION IN TRANSMISSION LINES

The dataset comprises three-phase voltage and current measurements that characterize the electrical conditions at a selected monitoring point in the transmission system. These measurements provide the primary information required to distinguish normal operation from fault conditions. During normal operation, the electrical quantities generally remain within their expected ranges and exhibit an approximately balanced three-phase response. Fault occurrence alters the magnitude and balance of these quantities, producing patterns that support automated fault detection and classification.[17]

Five operating conditions are considered: normal operation, single-line-to-ground faults, line-to-line faults, double-line-to-ground faults, and three-phase faults. Single-line-to-ground faults involve an electrical connection between one phase and ground and commonly produce voltage depression and increased current in the affected phase. Line-to-line faults involve two phases and introduce disturbances in their voltage and current characteristics. Double-line-to-ground faults involve two phases connected to ground, resulting in changes in phase magnitudes and system balance. Three-phase faults affect all phases simultaneously and typically produce substantial voltage depression and elevated fault currents. The severity of these disturbances depends on the fault location, fault resistance, system configuration, and operating conditions.[18]

In addition to the directly measured quantities, the dataset includes derived electrical features describing root-mean-square voltage and current, differences between phase voltages and currents, and apparent impedance. The root-mean-square quantities characterize the effective magnitudes of the measured signals, while the differences between phases provide information about their relative electrical behaviour. Apparent impedance contributes an additional representation of the transmission-line operating condition and complements the voltage and current features. Together, the measured and derived quantities form a set of fifteen input features, with each observation assigned to its corresponding operating condition.[19]

The dataset contains an equal number of observations for each of the five classes. This distribution provides comparable representation of normal and faulty conditions during model development and reduces the influence of class-frequency imbalance on classifier training. However, balanced class representation alone does not ensure unbiased predictions or equivalent performance across fault categories. A stratified division into training and testing subsets preserves the class proportions and supports evaluation across all operating conditions.[20]

Feature standardization is incorporated into the preprocessing procedure to place electrical quantities with different numerical ranges on a common scale. The standardization parameters are determined exclusively from the training subset and subsequently applied to the testing subset to prevent information leakage. Although Random Forest classification does not generally require feature scaling, this step provides a consistent preprocessing framework for the selected inputs. The training subset is used to establish the relationship between electrical characteristics and operating conditions, while the testing subset provides previously unseen observations for performance assessment.[21]

The Random Forest classifier combines predictions from multiple decision trees to identify patterns associated with normal operation and the different fault categories. By jointly examining measured and derived electrical features, the model can capture nonlinear relationships relevant to fault discrimination. Classification performance is evaluated using accuracy, precision, recall, and the harmonic mean of precision and recall. A confusion matrix provides a detailed account of correctly classified observations and misclassifications between operating conditions, supporting the identification of fault categories that are difficult to distinguish.

The selected feature set provides a basis for investigating machine-learning-based transmission-line fault detection and classification. Its practical relevance depends on the representativeness of the underlying measurements and the range of operating

and fault conditions included. Journal reporting should therefore identify the actual dataset source, acquisition procedure, and conditions represented. Further validation under variations in loading, fault resistance, fault location, and measurement noise is necessary to establish the suitability of the trained classifier for practical protection applications.[22]

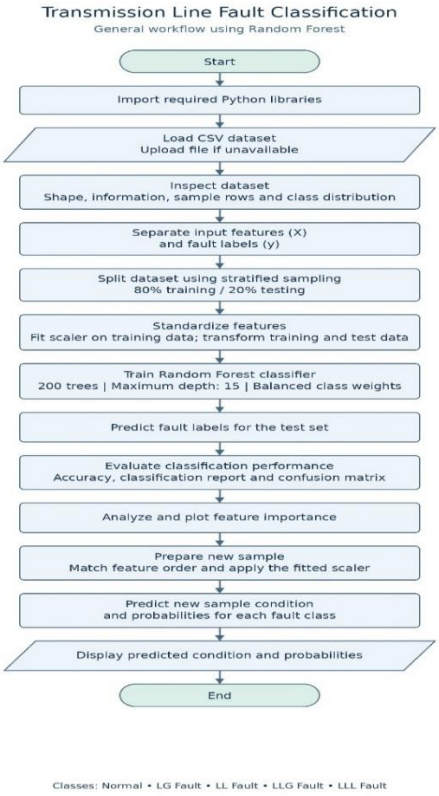


Figure 3: PYTHON PROGRAM IMPLEMENTING A MACHINE-LEARNING-BASED TRANSMISSION-LINE FAULT CLASSIFICATION SYSTEM USING A RANDOM FOREST CLASSIFIER

The proposed transmission-line fault classification framework employs a Random Forest classifier to distinguish normal operation from different fault conditions using electrical measurements. The implementation follows a structured workflow comprising dataset inspection, feature preparation, training and testing, performance evaluation, feature-importance analysis, and classification of new observations. Figure 3 illustrates the principal stages of the framework. The implementation is compatible with Google Colab and uses established Python libraries for numerical processing, data management, machine learning, and visualization.[23]

The dataset is initially examined to verify its dimensions, feature types, and class distribution. Each observation contains measured and derived electrical quantities describing the operating condition of the transmission line. These include three-phase voltages and currents, root-mean-square quantities, differences between phase measurements, and apparent impedance. The electrical features constitute the classifier inputs, while the corresponding operating-condition labels provide the reference categories for supervised learning. The classification task considers normal operation, single-line-to-ground faults, line-to-line faults, double-line-to-ground faults, and three-phase faults.[24]

The observations are divided into training and testing subsets, comprising eighty percent and twenty percent of the dataset, respectively. Stratified sampling preserves the relative representation of the operating conditions in both subsets. A fixed random seed ensures that the partition can be reproduced in subsequent executions. The training subset is used for model development, while the testing subset is reserved for evaluating classification performance on observations excluded from training.[25]

Feature standardization is incorporated into the preprocessing workflow to express the electrical measurements on a common numerical scale. The transformation is fitted exclusively to the training subset and subsequently applied to the testing subset, preventing information from the test observations from influencing preprocessing. Although feature scaling is generally unnecessary for decision-tree-based classifiers, its inclusion maintains a consistent input-processing procedure and supports potential comparisons with algorithms that are sensitive to differences in feature magnitude.[26]

The Random Forest model consists of two hundred decision trees, each restricted to a maximum depth of fifteen. Combining multiple trees enables the classifier to learn nonlinear relationships between electrical measurements and fault categories. The

depth restriction controls tree complexity and is intended to limit excessive fitting to the training observations. Balanced class weighting is also enabled, assigning weights according to class frequencies in the training subset. For an equally represented dataset, these weights are equivalent across classes. A fixed random seed supports reproducibility of model training.[27]

Following training, the classifier predicts the operating condition of each observation in the testing subset. The predicted categories are compared with their reference labels using accuracy, precision, recall, and the harmonic mean of precision and recall. Accuracy describes the overall proportion of correctly classified observations, while the class-specific measures provide a more detailed assessment of fault discrimination. Precision reflects the reliability of predictions assigned to a category, whereas recall indicates the proportion of observations within that category that are correctly identified. Reporting these measures for each operating condition enables assessment of performance differences that may be concealed by overall accuracy.[28]

A confusion matrix provides a complementary evaluation by showing the distribution of correct predictions and classification errors across the five operating conditions. Diagonal entries represent correctly identified observations, while off-diagonal entries reveal confusion between categories. This representation supports examination of whether errors occur between fault types or between normal and faulty operation. Such distinctions are relevant when assessing the potential application of the classifier to transmission-line monitoring.[29]

Feature-importance analysis is performed to examine the relative contribution of the electrical inputs to the trained model. The importance scores are ranked and displayed graphically to identify the measurements most frequently associated with reductions in classification impurity across the decision trees. This analysis provides insight into the features used by the classifier to distinguish operating conditions. However, these scores describe the behaviour of the fitted model and should not be interpreted as direct evidence of physical causation or the independent relevance of strongly correlated measurements.[30]

The trained framework is subsequently applied to a new observation representing an unclassified transmission-line condition. Before prediction, the measurements are arranged in the same feature order used during training and transformed using the previously fitted standardization procedure. Maintaining this consistency ensures that each measurement is interpreted according to its intended electrical meaning. The processed observation is then supplied to the classifier, which assigns it to one of the five operating categories.[31]

In addition to the predicted category, the framework produces estimated class probabilities that describe the model's relative preference among the available operating conditions. These estimates provide supplementary information about prediction ambiguity, particularly when several categories receive comparable scores. They should be interpreted as model outputs rather than calibrated measures of certainty unless probability calibration has been separately evaluated.[32]

The implementation establishes an integrated workflow for transmission-line fault classification, linking electrical feature processing with supervised learning, quantitative evaluation, and prediction of new observations. Its effectiveness must be established from the reported test results, while practical suitability requires further assessment using representative operating conditions and independently acquired measurements.

VI. RESULTS AND DISCUSSION

The developed transmission-line fault classification system was evaluated using a dataset containing 5,000 observations and sixteen columns, comprising fifteen electrical input features and one operating-condition label, as presented in Figure 4. The inputs include three-phase voltages and currents, root-mean-square voltage and current, differences between phase measurements, and apparent impedance. Five operating conditions were considered: normal operation, single-line-to-ground faults, line-to-line faults, double-line-to-ground faults, and three-phase faults. Each category contained approximately 1,000 observations, providing comparable class representation during model development. The testing subset comprised 1,000 observations, with 200 observations representing each operating condition.[33]

The Random Forest classifier achieved an overall testing accuracy of 97.90%, correctly identifying 979 of the 1,000 observations. Normal operation, single-line-to-ground faults, and three-phase faults each achieved precision, recall, and an F1-score of 1.00. These results indicate that all testing observations belonging to these categories were correctly identified, with no observations from other categories incorrectly assigned to them. For line-to-line faults, precision, recall, and the F1-score were approximately 0.92, 0.98, and 0.95, respectively. The corresponding values for double-line-to-ground faults were approximately 0.98, 0.92, and 0.95. Both the macro-averaged and weighted-averaged F1-scores were approximately 0.98, indicating strong aggregate performance across the evaluated categories.[34]

The confusion matrix reveals that classification errors were confined to line-to-line and double-line-to-ground faults. Of the 200 line-to-line fault observations, 196 were correctly classified, while four were identified as double-line-to-ground faults. Conversely, 183 double-line-to-ground fault observations were correctly classified, while seventeen were assigned to the line-to-line category. This asymmetric error pattern indicates that the classifier had greater difficulty recognizing double-line-to-ground

faults. Overlap in the electrical features associated with these two categories offers a plausible explanation, although confirming the cause would require closer examination of the misclassified observations and their underlying fault conditions.[35]

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 16 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Va      5000 non-null   float64
 1   Vb      5000 non-null   float64
 2   Vc      5000 non-null   float64
 3   Ia      5000 non-null   float64
 4   Ib      5000 non-null   float64
 5   Ic      5000 non-null   float64
 6   V_RMS   5000 non-null   float64
 7   I_RMS   5000 non-null   float64
 8   Vab     5000 non-null   float64
 9   Vbc     5000 non-null   float64
 10  Vca     5000 non-null   float64
 11  Iab     5000 non-null   float64
 12  Ibc     5000 non-null   float64
 13  Ica     5000 non-null   float64
 14  Z_App   5000 non-null   float64
 15  Fault   5000 non-null   object
dtypes: float64(15), object(1)
memory usage: 625.1+ KB
None
```

```
First Five Rows:
         Va        Vb        Vc        Ia        Ib        Ic     V_RMS  \
0  1.025726  1.004909  1.008259  1.056542  0.983707  0.983661  1.013006
1  0.995759  0.947611  0.980158  0.978616  0.970966  0.980932  0.974716
2  0.983095  1.020741  0.976845  0.990117  0.998432  0.966832  0.993750
3  1.003021  0.957796  0.958651  0.991509  1.012941  1.026283  0.973385
4  1.017182  0.974052  0.998090  0.982263  0.986318  1.037861  0.996598

      I_RMS       Vab       Vbc       Vca       Iab       Ibc       Ica  \
0  1.008555  0.020817  0.003350  0.017467  0.072835  0.000046  0.072881
1  0.976847  0.048148  0.032547  0.015601  0.007650  0.009967  0.002317
2  0.985218  0.037646  0.043896  0.006250  0.008315  0.031599  0.023284
3  1.010346  0.045225  0.000855  0.044370  0.021432  0.013342  0.034775
4  1.002467  0.043130  0.024038  0.019093  0.004054  0.051544  0.055598

      Z_App   Fault
0  1.004413  Normal
1  0.997818  Normal
2  1.008660  Normal
3  0.963417  Normal
4  0.994145  Normal

Class Distribution:
Fault
Normal       1000
LG Fault     1000
LL Fault     1000
LLG Fault    1000
LLL Fault    1000
Name: count, dtype: int64
```

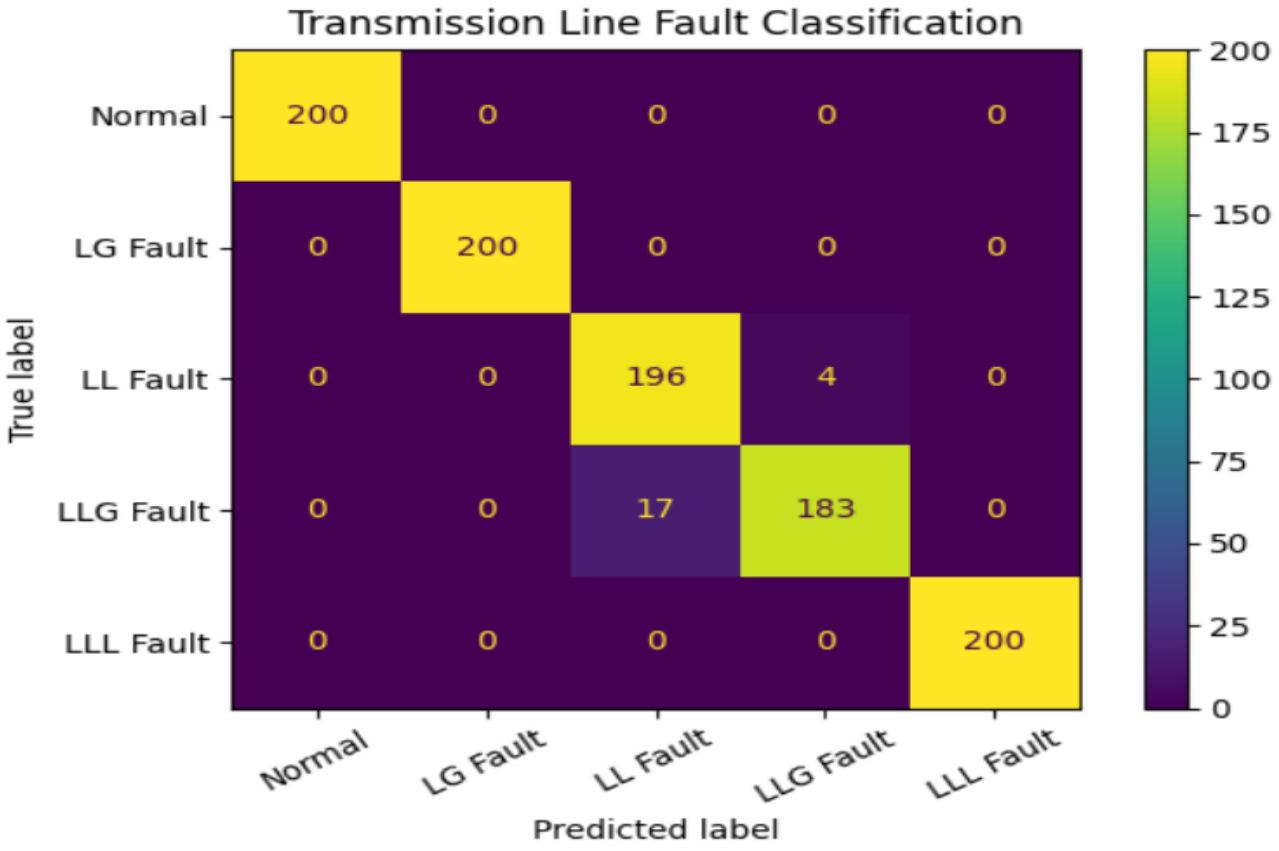
```
Model Accuracy: 97.90%

Classification Report:
              precision    recall  f1-score   support

    LG Fault       1.00      1.00      1.00       200
    LL Fault       0.92      0.98      0.95       200
   LLG Fault       0.98      0.92      0.95       200
   LLL Fault       1.00      1.00      1.00       200
      Normal       1.00      1.00      1.00       200

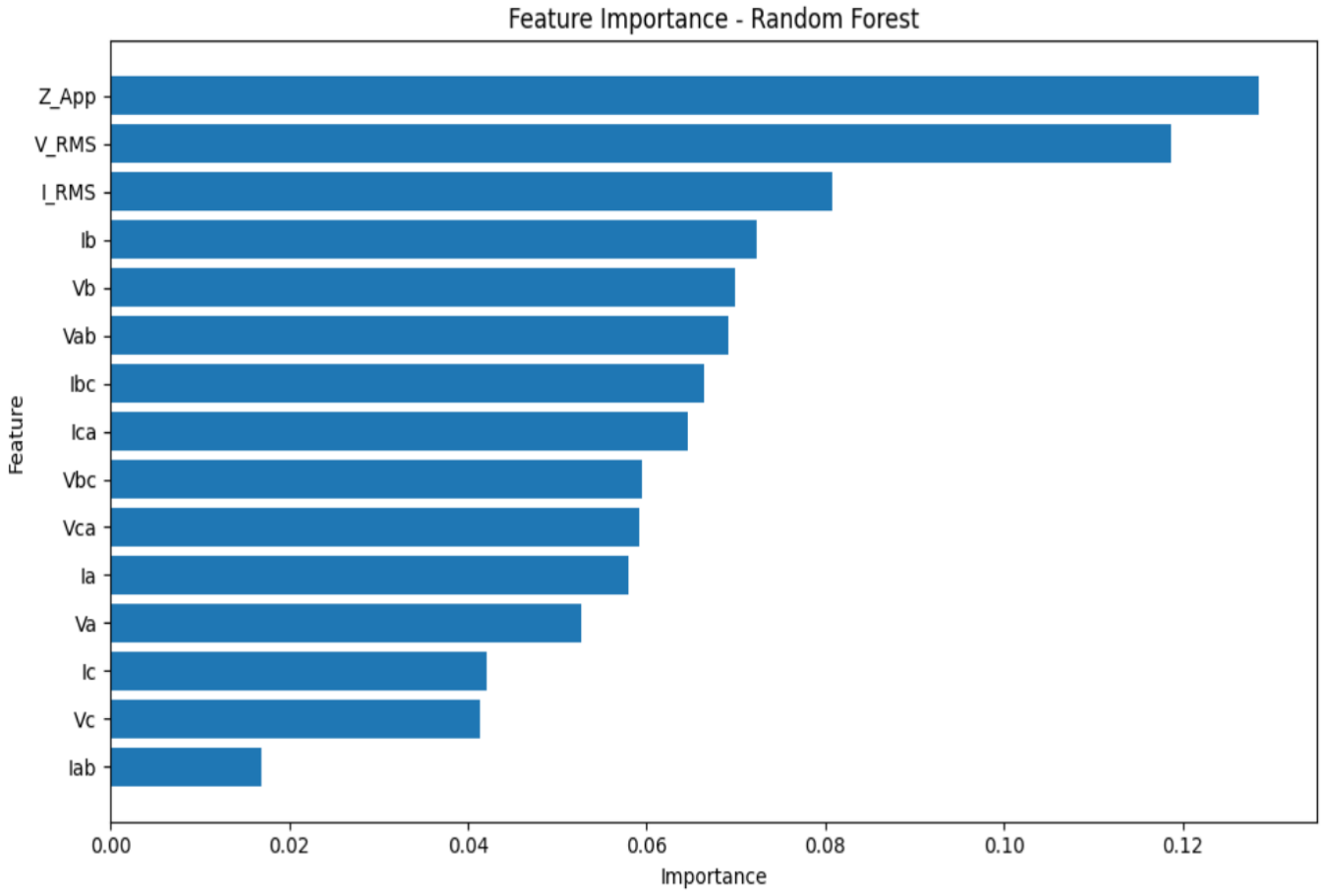
    accuracy                           0.98      1000
   macro avg       0.98      0.98      0.98      1000
weighted avg       0.98      0.98      0.98      1000


Confusion Matrix:
[[200   0   0   0   0]
 [  0 200   0   0   0]
 [  0   0 196   4   0]
 [  0   0  17 183   0]
 [  0   0   0   0 200]]
```



Feature Importance:

	Feature	Importance
14	Z_App	0.128563
6	V_RMS	0.118698
7	I_RMS	0.080737
4	Ib	0.072252
1	Vb	0.069845
8	Vab	0.069165
12	Ibc	0.066408
13	Ica	0.064595
9	Vbc	0.059522
10	Vca	0.059110
3	Ia	0.058043
0	Va	0.052623
5	Ic	0.042145
2	Vc	0.041366
11	Iab	0.016930



```
Predicted Condition:
LG Fault

Prediction Probabilities:
LG Fault: 88.00%
LL Fault: 11.00%
LLG Fault: 0.00%
LLL Fault: 1.00%
Normal: 0.00%
```

Figure 4: RANDOM FOREST–BASED TRANSMISSION LINE FAULT CLASSIFICATION: DATASET CHARACTERISTICS, CLASSIFICATION ACCURACY, PERFORMANCE METRICS, CONFUSION MATRIX, FEATURE IMPORTANCE, AND PREDICTION OF FAULT CONDITIONS WITH ASSOCIATED CLASS PROBABILITIES.

Feature-importance analysis identified apparent impedance as the highest-ranked input, with an importance score of approximately 0.129. Root-mean-square voltage and root-mean-square current followed, with scores of approximately 0.119 and 0.080, respectively. Additional contributions were provided by individual phase measurements and differences between phase voltages and currents. The ranking suggests that impedance-related behaviour and effective electrical magnitudes contributed substantially to the model's classification decisions. However, these importance scores describe how the trained ensemble used the available features; they do not establish physical causation or independently quantify the diagnostic value of correlated inputs.[36]

The framework was also applied to an individual observation, which was classified as a single-line-to-ground fault. The estimated class probabilities were 88% for a single-line-to-ground fault, 11% for a line-to-line fault, and 1% for a three-phase fault, with the remaining categories receiving zero at the reported precision. This distribution indicates a clear model preference for the predicted category while retaining some ambiguity with the line-to-line condition. The probability estimates supplement the predicted label but should not be interpreted as calibrated certainty without a separate calibration assessment. Moreover, the correctness of this individual prediction cannot be established without its reference label.[37][38]

The results demonstrate that the Random Forest classifier effectively distinguished the five operating conditions within the evaluated testing subset. Its principal limitation was the confusion between line-to-line and double-line-to-ground faults, particularly the reduced recall for the latter category. Apparent impedance and root-mean-square measurements were prominent contributors to the learned classification patterns. These findings support the approach as a candidate for transmission-line fault diagnosis, while broader validation is required to establish performance under independent operating conditions. Practical protection applications would additionally require evaluation of measurement noise, fault-location and fault-resistance variations, and processing latency before claims of real-time reliability can be substantiated.[39][40]

VII. CONCLUSION

The proposed AI-based transmission-line fault detection and classification system successfully demonstrates the application of machine learning for identifying and classifying different operating conditions of a transmission line. The developed system uses electrical parameters such as three-phase voltages, three-phase currents, RMS quantities, line-to-line voltages and currents, and apparent impedance as input features. A Random Forest classifier was implemented to distinguish between Normal, LG, LL, LLG, and LLL fault conditions. The model achieved an overall classification accuracy of 97.90%, demonstrating its high effectiveness and reliability in identifying transmission-line faults. The classification results show that the model achieved perfect classification for Normal, LG Fault, and LLL Fault conditions, while only minor misclassification was observed between LL and LLG faults. The confusion matrix confirms that most of the test samples were correctly classified, with the main errors occurring because of the similarity between LL and LLG fault characteristics. The feature-importance analysis also demonstrated that apparent impedance, RMS voltage, and RMS current are among the most influential parameters for fault classification. In addition, the model successfully predicted a new operating condition as an LG Fault with an 88% probability, demonstrating its ability to provide both fault classification and prediction confidence. Overall, the study demonstrates that machine-learning

techniques can provide a fast, accurate, and intelligent approach to transmission-line fault detection and classification. The proposed system has the potential to support conventional protection methods and improve the reliability, security, and monitoring capability of modern power transmission networks. However, further validation using larger and more diverse real-world datasets, different fault conditions, noise levels, fault resistances, and varying system operating conditions is necessary before practical implementation. The integration of such AI-based techniques with conventional protection systems can contribute significantly to the development of intelligent, adaptive, and reliable power-system protection and smart-grid applications.

REFERENCES

[1] S. H. Horowitz and A. G. Phadke, *Power System Relaying*, 4th ed. Chichester, U.K.: Wiley, 2014.

[2] G. Ziegler, *Numerical Distance Protection: Principles and Applications*, 4th ed. Erlangen, Germany: Publicis Publishing, 2011.

[3] A. T. Johns and S. K. Salman, *Digital Protection for Power Systems*. London, U.K.: Institution of Electrical Engineers, 1995.

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

[5] A. Prasad, J. B. Edward, and K. Ravi, "A review on fault classification methodologies in power transmission systems: Part—I," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 1, pp. 48–60, 2018, doi: 10.1016/j.jesit.2017.01.004.

[6] A. Prasad, J. B. Edward, and K. Ravi, "A review on fault classification methodologies in power transmission systems: Part—II," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 1, pp. 61–67, 2018, doi: 10.1016/j.jesit.2016.10.003.

[7] A. Mukherjee, P. K. Kundu, and A. Das, "Transmission line faults in power system and the different algorithms for identification, classification and localization: A brief review of methods," *Journal of The Institution of Engineers (India): Series B*, vol. 102, no. 4, pp. 855–877, 2021, doi: 10.1007/s40031-020-00530-0.

[8] R. N. Mahanty and P. B. Dutta Gupta, "A fuzzy logic based fault classification approach using current samples only," *Electric Power Systems Research*, vol. 77, nos. 5–6, pp. 501–507, 2007, doi: 10.1016/j.epsr.2006.04.009.

[9] P. K. Dash, S. R. Samantaray, and G. Panda, "Fault classification and section identification of an advanced series-compensated transmission line using support vector machine," *IEEE Transactions on Power Delivery*, vol. 22, no. 1, pp. 67–73, 2007, doi: 10.1109/TPWRD.2006.876695.

[10] S. R. Samantaray, "Decision tree-based fault zone identification and fault classification in flexible AC transmissions-based transmission line," *IET Generation, Transmission & Distribution*, vol. 3, no. 5, pp. 425–436, 2009, doi: 10.1049/iet-gtd.2008.0316.

[11] J. Upendar, C. P. Gupta, and G. K. Singh, "Statistical decision-tree based fault classification scheme for protection of power transmission lines," *International Journal of Electrical Power & Energy Systems*, vol. 36, no. 1, pp. 1–12, 2012, doi: 10.1016/j.ijepes.2011.08.005.

[12] P. Ray and D. P. Mishra, "Support vector machine based fault classification and location of a long transmission line," *Engineering Science and Technology, an International Journal*, vol. 19, no. 3, pp. 1368–1380, 2016, doi: 10.1016/j.jestch.2016.04.001.

[13] V. Malathi, N. S. Marimuthu, and S. Baskar, "Intelligent approaches using support vector machine and extreme learning machine for transmission line protection," *Neurocomputing*, vol. 73, nos. 10–12, pp. 2160–2167, 2010, doi: 10.1016/j.neucom.2010.02.001.

[14] S. Belagoune, N. Bali, A. Bakdi, B. Baadji, and K. Atif, "Deep learning through LSTM classification and regression for transmission line fault detection, diagnosis and location in large-scale multi-machine power systems," *Measurement*, vol. 177, Art. no. 109330, 2021, doi: 10.1016/j.measurement.2021.109330.

[15] A. Moradzadeh, H. Teimourzadeh, B. Mohammadi-Ivatloo, and K. Pourhossein, "Hybrid CNN-LSTM approaches for identification of type and locations of transmission line faults," *International Journal of Electrical Power & Energy Systems*, vol. 135, Art. no. 107563, 2022, doi: 10.1016/j.ijepes.2021.107563.

[16] S. Bin Akter, T. Sarkar Pias, S. Rahman Deeba, J. Hossain, and H. Abdur Rahman, "Ensemble learning based transmission line fault classification using phasor measurement unit (PMU) data with explainable AI (XAI)," *PLOS ONE*, vol. 19, no. 2, Art. no. e0295144, 2024, doi: 10.1371/journal.pone.0295144.

[17] A. Abdullah, "Ultrafast transmission line fault detection using a DWT-based ANN," *IEEE Transactions on Industry Applications*, vol. 54, no. 2, pp. 1182–1193, 2018, doi: 10.1109/TIA.2017.2774202.

[18] A. Prasad and J. B. Edward, "Application of wavelet technique for fault classification in transmission systems," *Procedia Computer Science*, vol. 92, pp. 78–83, 2016, doi: 10.1016/j.procs.2016.07.326.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[20] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.

[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[22] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998, doi: 10.1109/34.709601.

[23] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, vol. 1857, Lecture Notes in Computer Science. Berlin, Germany: Springer, 2000, pp. 1–15, doi: 10.1007/3-540-45014-9_1.

[24] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA, USA: Wadsworth International Group, 1984.

[25] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.
[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995, doi: 10.1007/BF00994018.
[27] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.
[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
[29] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, Art. no. 25, 2007, doi: 10.1186/1471-2105-8-25.
[30] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, Art. no. 307, 2008, doi: 10.1186/1471-2105-9-307.
[31] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in Neural Information Processing Systems*, vol. 26, pp. 431–439, 2013.
[32] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
[33] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
[34] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[35] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, pp. 3133–3181, 2014.
[36] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[37] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[38] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[39] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020, doi: 10.1038/s41586-020-2649-2.
[40] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.