

IDENTIFICATION OF LINE FAULTS IN POWER DISTRIBUTION NETWORKS SUPPORTED BY MICRO PHASOR MEASUREMENT UNITS

Sukalyan Maji^{1,2}, Deepak Kumar Singh¹, Emon Das¹, Partha Kayal^{1,}*

¹Department of Electrical Engineering, National Institute of Technology Silchar, Assam-788010, India

²Department of Electrical Engineering, Veermata Jijabai Technological Institute, Matunga East, Maharashtra-400019, India

Corresponding Author: ^{1,}Partha Kayal (Email: partha@ee.nits.ac.in)

Co-Authors: ^{1,2}Sukalyan Maji (Email: connectsukalyanmaji@gmail.com), ¹Deepak Kumar Singh (Email: deepakk21_ug@ee.nits.ac.in), ¹Emon Das (Email: emon21_ug@ee.nits.ac.in)
 (Received: 23-February-2026; accepted: 25-August-2026; published: 30-September-2026)

<http://dx.doi.org/10.55579/jaec.2026103.540>

Abstract. Power distribution systems are becoming more complex and vulnerable day by day due to the increasing penetration of non-linear loads, renewable sources, and critical expansions of the overhead supply networks. Thus, the chances of distribution line failures have increased. On this occasion, effective fault detection aids in developing predictive maintenance strategies, helping utilities reduce downtime and utilize resources more effectively. In this paper, a machine learning (ML)-based methodology for fault detection on overhead distribution lines has been presented. The study has demonstrated the application of the two typical tree-based ML classifiers, i.e., random forest (RF) and extreme gradient boosting (XGBoost), to diagnose distribution line faults with the help of a micro phasor measurement unit (μ PMU) supported data set. The study reveals a μ PMU allocation strategy and data collection method to aid in the preparation of the data set. The classification models are structured to classify the location of the line faults, followed by the faults' type, based on the prepared data set for a test network. The efficiency of the RF and XGBoost has been verified through a quantitative comparison with the existing approaches.

Keywords: Classification, Line faults, Micro phasor measurement unit, Power distribution network.

1. Introduction

Faults within the distribution systems can significantly impair the continuous and safe delivery of electricity. The most frequent types include short circuits, open circuits, and ground faults. A short circuit occurs when a phase conductor unintentionally contacts another phase, neutral wire, or the ground, forming a low-resistance path that causes a surge in current. This excessive current can result in overheating, malfunctioning equipment, or large-scale outages if not promptly managed. Ground faults happen when an energized conductor touches the ground or a grounded object, producing a fault current that can lead to serious safety concerns. Fault detection and localization in power systems play a critical role in ensuring system reliability, minimizing downtime, and maintaining efficient operations. With the increasing com-

plexity of modern distributed power systems, accurate identification of fault type and fault location has become even more essential. Therefore, efficient fault detection and isolation are essential to reduce downtime and maintain the safety and reliability of the supply system.

The recent application of phasor measurement units in power distribution networks facilitates fast and efficient fault detection by analyzing the collected micro phasor measurement unit (μ PMU) data at the control center. Some authors have already devoted some efforts towards the development of a methodology for efficient fault identification. In [1], the authors have proposed a graph-based fault identification method. The authors first focused on allocating the μ PMUs to reduce the measurement redundancy, followed by the distribution system state estimation (DSSE) to identify the prospective fault location region. With the demonstrated approach, the authors have sought to determine the precise location of a faulted feeder using graph analysis techniques. Recent PMU-based fault localization studies have also reported improved observability and fault discrimination in distribution networks [2, 3]. In [3], Mirshekali et al. presented an SVM-based fault classification approach to identify fault locations and types in a power distribution network. The authors in [4] have used bus voltage phasors measured with μ PMUs as input data for their proposed model. In [5], the authors employed a graph convolutional network (GCN)-based classification model to identify the faulted condition and its location within a test distribution network. The authors have organized the voltage magnitude and their phase angle data fed by the μ PMUs to use in their CGN-based classification model. In [6, 7], the authors have developed a novel artificial neural network (ANN)-based system for fault detection. This approach addresses the challenges posed by communication delays in Wide Area Monitoring Systems, ensuring timely and accurate detection. The authors have emphasized the need for further testing to explore additional features that could enhance its adaptability to diverse network scenarios. Deep learning models based on PMU measurements have also been explored for multi-class fault diagnosis in active distribution systems [8].

In [9], the authors have firstly developed a spatial-temporal relationship of the measured data from the μ PMU installed in their test system. The authors have used a multi-residual generation learning model on the augmented spatio-temporal data set for the identification of the prospective phase to ground faults at the distribution lines. In [10], Dashtdar et al. have proposed an optimization-based model to identify the fault-location region and the faulted line using voltage variation information across both sides of the distribution lines. The authors have realized the minimization of the power system state estimation error for obtaining suitable positions of PMUs in a test distribution network and collected the measurement and estimated voltage magnitudes for all the buses thereafter. Though the authors have claimed about its efficacy over the existing classification-based fault detection approaches, the time complexity of the approach can not be avoided. In [11], Medattil et al. have addressed the challenge of classifying faults in a dataset with the imbalanced classes. The study utilized SMOTE-based oversampling techniques combined with the K-Nearest Neighbor (KNN) algorithm to achieve impressive fault classification accuracy. In [12], Xiang has proposed a deep learning-based approach for the detection of high impedance faults in distribution systems. The authors in [13, 14] proposed applying current signal data collected by μ PMUs to a recurrent neural network to determine the faulty zone in the test distribution system. The author has concluded that higher accuracy in fault type detection was possible in comparison to fault location detection by his developed model. In [15, 16], the authors have introduced a cutting-edge framework for addressing fault management in distribution networks using data from the PMUs. Leveraging the capabilities of deep convolutional neural networks (CNNs), this study focuses on classifying faults, identifying affected sections, and pinpointing fault locations.

It has been learned from the existing state-of-the-art that most of the recent investigations were based on the application of machine learning (ML) methods to diagnose the distribution line faults. Though some of the authors have applied graph-based algorithms for the classification of distribution line faults, i.e., [1, 4, 12].

Though an increasing interest has been observed in recent days in applying neural network-inspired deep learning techniques, sometimes deep learning techniques underperform with respect to the augmented tree-based ML techniques, i.e., extreme gradient boosting (XGBoost) [17, 18]. The authors in [17] have identified that the non-linearity present in the structure of the neural network-based models is one of the causes for their underperformance with respect to the XGBoost model. Therefore, this study delivers the application of recently developed machine learning algorithms, i.e., XGBoost and the random forest (RF) algorithm, for the identification of distribution line faults, classifying the data from μ PMU measurements. The study further reveals a data pre-processing technique and demonstrates its significance in improving the classification accuracy. The study has dealt with all the possible types of shunt faults for an unbalanced distribution test network and found the type of line fault followed by the location of the faulty line. This study also highlights a quantitative comparative study of the proposed machine learning algorithm with respect to neural network-based models.

The remainder of the paper is structured as follows. Section 2 provides a concise description of the μ PMU allocation technique and the framework for the data set preparation with the measurements obtained from the μ PMUs, followed by Section 3, where the structure of the tree-based ML techniques is discussed. Section 4 demonstrates the data pre-processing technique to feed the classification models, followed by the classification framework. Section 5 highlights the simulation results and the analysis of the results. Finally, the significant concluding remarks are summarized in Section 6.

2. μ PMU Allocation and Data Collection

The data collected from the multiple bus μ PMU installed locations aids in identifying fault location and the type of fault on the defective line. However, it is to be noted that the μ PMUs should be placed at the strategic loca-

tions so that they can decipher full topological observability over the test distribution network. Whereas redundant installation of μ PMUs creates unnecessary cost burden, the low number of μ PMUs may fail to deliver the full information about the system in a time-synchronized manner. Thus, a certain strategy is needed for μ PMU allocation.

2.1. μ PMU Allocation

The allocation strategy employed in this study focuses on minimizing the total installation cost associated with deploying μ PMU devices in the distribution network. The objective function is formulated to achieve cost-effective placement of μ PMUs while guaranteeing complete systems observability. To this end, the number of measurement channels is tailored based on the incident branch count and phase configurations at the candidate bus location. The optimized objective is fulfilled through the resolution of the function F , as defined in Equation (1).

$$F = \min \sum_{i=1}^{N_B} C_i^{\mu PMU} x_i \quad (1)$$

where, N_B represents the number of buses in the test system. The following equation defines the characteristic of the variable x_i :

$$x_i = \begin{cases} 1, & \text{if a } \mu\text{PMU is installed at bus } i, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where, $C_i^{\mu PMU}$ is the cost of μ PMU for deploying at the i^{th} bus location.

The important criteria about the observability on the distribution network must be fulfilled while minimizing the mentioned objective function. The observability criteria can be realized as follows:

$$A \times X > B \quad (3)$$

where,

$$[A] = a_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 1, & \text{if buses } i \text{ and } j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

$$X = [x_1 \ x_2 \ \cdots \ x_{N_B}]^T \quad (5)$$

$$B = [1 \ 1 \ 1 \ \cdots \ 1]_{N_B \times 1}^T \quad (6)$$

where, $[A]$ is called the connectivity matrix. Following the above constraint criteria, it is possible to obtain full observability over the system either by direct measurement from the buses or indirect measurements with a direct connection of μ PMU installed buses. The formulated μ PMU allocation problem is a constrained binary optimization problem. Thus, the result for the location of the μ PMUs in the test network should be evaluated with the application of an optimization technique. Though a variety of optimization techniques are available, the proposed study has adopted the integer linear programming (ILP) method [19] to solve the μ PMU allocation problem. The authors have verified their results with the existing literature and affirm that ILP can work well for the μ PMU allocation problem. Recent work on PMU placement has further confirmed that observability and measurement placement strongly influence downstream fault localization performance [20]. Moreover, the issue of local convergence, which may occur due to the application of heuristic optimization approaches, can be avoided by employing ILP. However, the present placement strategy is validated for the considered 19-bus network topology, and the resulting observability guarantee applies to that modeled configuration only.

2.2. Data Collection

The test system considered in this study is a 19-bus unbalanced power distribution network [21]. The test distribution network is a low-voltage Indian distribution network with a nominal voltage level of 11 kV, and it is radial in nature. The application of ILP yields the desired number of locations to install μ PMUs in the 19-bus test system is eight. Thus, it can be observed from Figure 1 that a minimum of eight μ PMUs are required to obtain the full observability over the 19-bus test network. After allocation of μ PMUs, the voltage and current

phasor data from each μ PMU are collected and organized together at a central station, namely the control center, for further processing. In this study, the described μ PMU-installed test system is simulated in MATLAB software, and sample data from μ PMUs are generated, creating a variety of scenarios in the simulation workspace. The developed scenarios include the different types of fault situations at the different distribution lines with nominal loads and load change scenarios without faulty conditions. Further, the different faulty condition scenarios include all the possible shunt faults that may occur among the different phases of the overhead distribution lines. Thus, line-to-ground (LG) faults for the three phases, line-to-line (LL) faults among the three phases, a combination of double line-to-ground (LLG) faults among the phases to ground, and three-phase (LLL) faults are considered in this study. The different operational scenarios are simulated with the test distribution network, and the voltage magnitude along with its phase angle, current magnitude along with its phase angle from the eight buses fetched by the μ PMUs for each scenario are recorded to create a data set. It is to be noted that different types of faults result in different voltage and current phasors for different phases of the distribution system. Thus, the data collected by the μ PMUs can be suitably synthesized to identify the type of faults. Further, the values of the phasors are influenced by the location of the faults. A data set corresponding to the different faults is organized using data collected at the end of three cycles to avoid transient value collection. The nature of the fault current and fault voltage for different types of faults happened at the preceding branch of bus-9 is depicted in Figures 2 to 17.

3. Tree-Based ML Models

The ML models inspired by the decision tree structures are quite powerful in classification and regression, and these models work with a supervised learning strategy. The tree-based ML models include SVM, classification and regression tree (CART), RF, and XGBoost. These models are different in their tree-building mech-

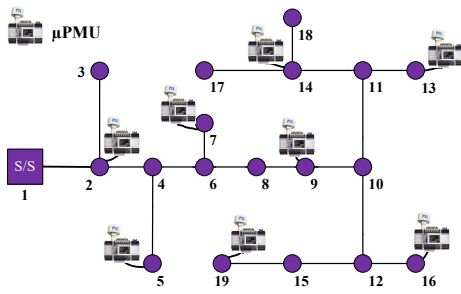


Figure 1: The identified bus locations for μ PMU allocation in the test network.

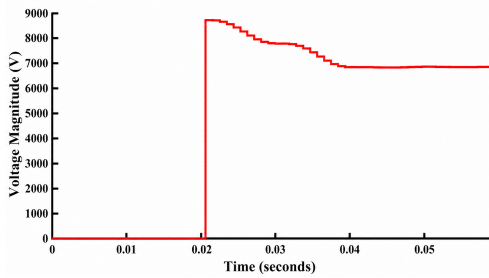


Figure 2: Voltage magnitude at bus 9 during phase-A to ground fault at the preceding branch of bus 9.

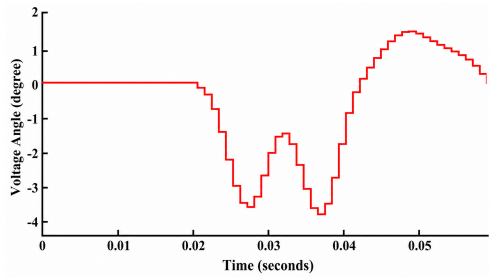


Figure 3: Voltage angle at bus 9 during phase-A to ground fault at the preceding branch of bus 9.

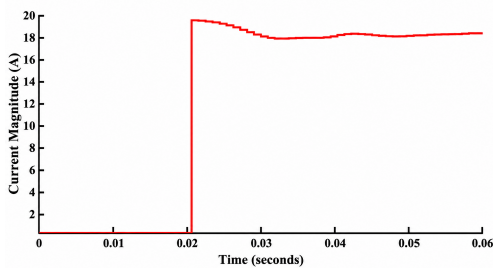


Figure 4: Current magnitude at bus 9 during phase-A to ground fault at the preceding branch of bus 9.

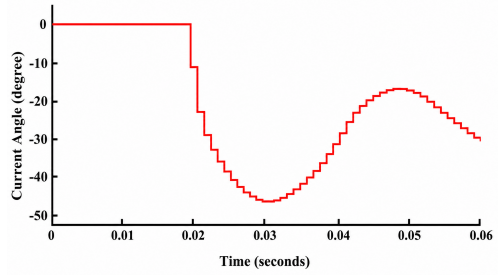


Figure 5: Current angle at bus 9 during phase-A to ground fault at the preceding branch of bus 9.

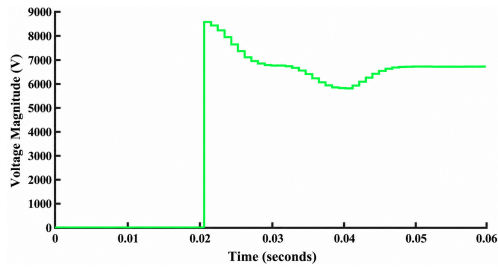


Figure 6: Voltage magnitude at bus 9 during phase-A-B to ground fault at the preceding branch of bus 9.

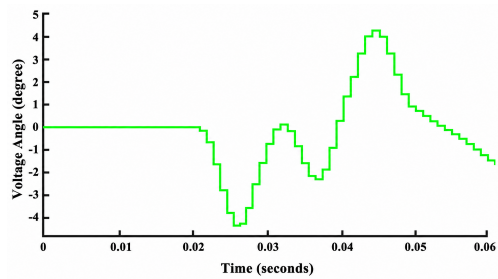


Figure 7: Voltage angle at bus 9 during phase-A-B to ground fault at the preceding branch of bus 9.

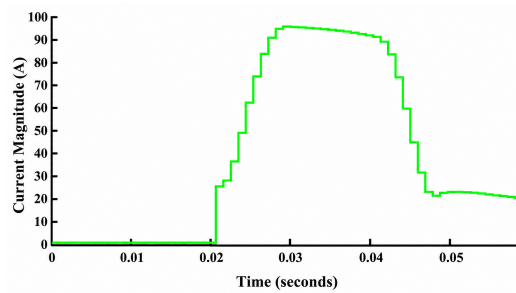


Figure 8: Current magnitude at bus 9 during phase-A-B to ground fault at the preceding branch of bus 9.

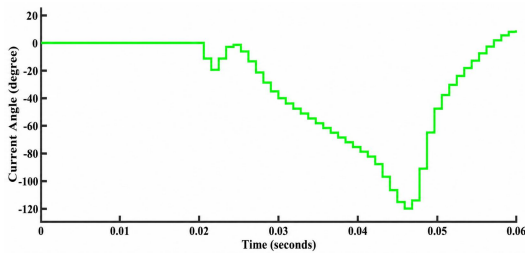


Figure 9: Current angle at bus 9 during phase-A-B to ground fault at the preceding branch of bus 9.

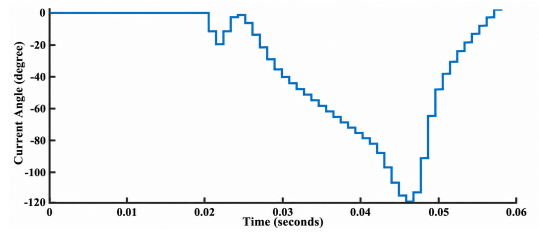


Figure 13: Current angle at bus 9 during phase-A to phase-B fault at the preceding branch of bus 9.

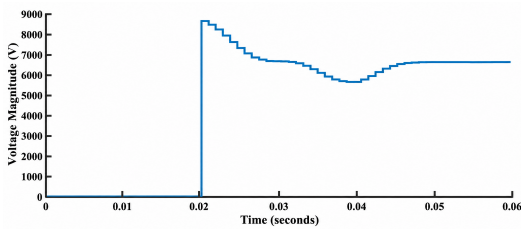


Figure 10: Voltage magnitude at bus 9 during phase-A to phase-B fault at the preceding branch of bus 9.

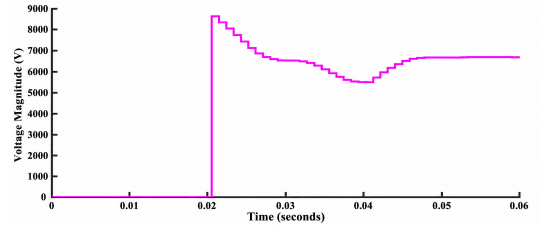


Figure 14: Voltage magnitude at bus 9 during phase-A-B-C to ground fault at the preceding branch of bus 9.

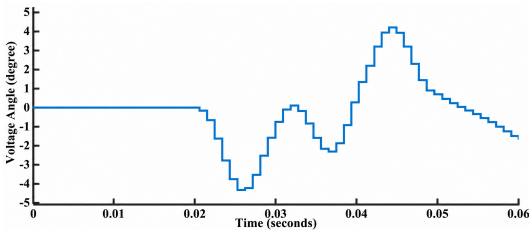


Figure 11: Voltage angle at bus 9 during phase-A to phase-B fault at the preceding branch of bus 9.

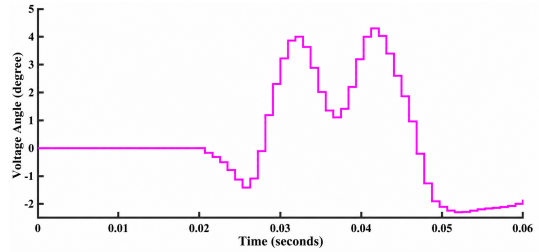


Figure 15: Voltage angle at bus 9 during phase-A-B-C to ground fault at the preceding branch of bus 9.

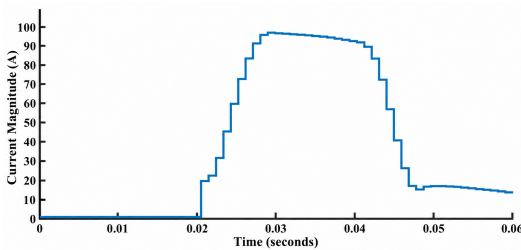


Figure 12: Current magnitude at bus 9 during phase-A to phase-B fault at the preceding branch of bus 9.

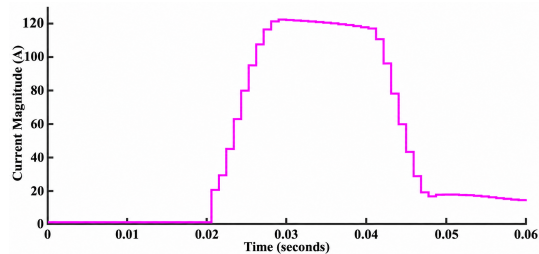


Figure 16: Current magnitude at bus 9 during phase-A-B-C to ground fault at the preceding branch of bus 9.

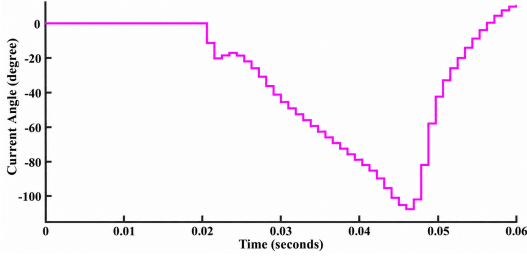


Figure 17: Current angle at bus 9 during phase-A-B-C to ground fault at the preceding branch of bus 9.

anisms and decision rules. The deeper the tree, the more complex the decision rules and the fitter the model. As RF and XGBoost are comparatively advanced tree-based models, they are considered as potential ML models to classify the distribution line faults. Recent studies have also shown that ensemble tree-based models such as Random Forest and XGBoost can achieve strong fault classification performance in power-system applications [22,23]. Moreover, the RF and XGBoost have not been applied so far to classify fault type and fault location in unbalanced power distribution networks to the best of the authors' knowledge.

3.1. RF Model

The RF algorithm builds a group of decision trees during the training process and combines their individual outputs to produce a final prediction [24]. The trees in RF are generated randomly through a training phase on a random sample with a replacement of the training set (bagging) and a restriction of the features. Through this mechanism, random forest is also used to identify the most important features by minimizing the out-of-bag error (OOB). Each tree has its own OOB data set, which is used for error estimation of the individual tree in the forest. It is called OOB error estimation [14]. The generalization error in RF is represented as:

$$PE^* = P_{X,Y}(\text{Marg}(X,Y) < 0) \quad (7)$$

where, $\text{Marg}(X,Y)$ is a Margin function. The Margin function measures the extent to which the average number of votes at (X,Y) for the right class exceeds the average vote for any other

class. Here, X is the predictor vector and Y is the classification. P_{XY} signifies the probability taken for all the possible combinations of X and Y .

3.2. XGBoost Model

XGBoost is an advanced ensemble machine learning algorithm based on the gradient boost framework [15]. It builds a group of decision trees sequentially, where each tree corrects the errors made by the previous ones. The key strength of XGBoost lies in its ability to efficiently handle large, complex datasets and its regularization techniques that prevent overfitting, making it highly effective for fault detection and classification tasks in power systems [25]. The mathematical basis of XGBoost revolves around optimizing a loss function while adding weak learners (decision trees) sequentially. Thus, the objective is to minimize the following function:

$$Obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \Omega(f) \quad (8)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^k w_j^2 \quad (9)$$

where, L represents the loss function, and Ω is the regularization term ensuring simplicity of the model; f represents the prediction function; n indicates the number of training samples in the dataset; y_i and $\hat{y}_i$ are the actual and predicted values corresponding to the i^{th} training sample; T is the number of leaves whose weight is represented by ω , γ is a learning rate used for pruning, and λ is a regularization parameter. Each iteration of boosting tries to fit the residual errors optimally, improving predictions iteratively.

4. Data Preprocessing and Classification Framework

Preprocessing the raw data in ML is a significant step done in the initial stages after data

extraction. The data preprocessing helps improve the quality of data and the extraction of substantial information.

4.1. Data Preprocessing

The dataset created by the information about the voltage and current phasors from the μ PMU measurements is further processed to obtain a synthetic dataset. The procedure of developing the synthetic data set includes interpolation and noise addition [16]. By utilizing noise and interpolation, the dataset introduces a wide range of realistic operation conditions.

With additional diversified data, the original data set is smoothened and it generalizes the cause-effect relationship in the data more effectively. This enhanced dataset is crucial for training the ML models to improve the fault classification accuracy. By the interpolation operation, the intermediate data points are inferred. For two consecutive points (x_i, y_i) and (x_{i+1}, y_{i+1}) , the interpolated value $y(x)$ at any point x between x_i and x_{i+1} can be calculated as:

$$y(x) = y_i + \left\{ \frac{(x - x_i)}{(x_{i+1} - x_i)} \times (y_{i+1} - y_i) \right\} \quad (10)$$

Noise has been introduced to the dataset to simulate real world measurements inaccuracies and uncertainties, enhancing the robustness of the model.

$$Y_{noisy} = y_i + \mathcal{N}(0, \sigma \cdot \text{range}(y)) \quad (11)$$

where, σ is the standard deviation of the Gaussian noise.

The overall synthetic dataset can be described as the combination of interpolated, noisy, and original data samples.

$$Y_{synthetic} = Y_{original} \cup Y_{interpolated} \cup Y_{noisy} \quad (12)$$

The interpolation and Gaussian-noise steps are introduced solely to enrich the training dataset and improve model robustness against measurement uncertainty. Because the input samples are phasor measurements recorded after three cycles, the synthetic augmentation

is not intended to emulate transient waveform behavior or HIF-specific dynamics. $Y_{synthetic}$, $Y_{interpolated}$, Y_{noisy} , and $Y_{original}$ represent synthetic, interpolated, noisy, and original data samples, respectively.

4.2. Classification Framework

The complete data sets are always divided into two sub-sets by the classification models. The sub-sets are created with randomly chosen data points from the complete set. The sub-sets are called the training sub-set and the testing sub-set. The data subset that is used to train the classification model is designated as the training subset. The subset that is kept to measure the performance of the classification models is known as the test dataset. Usually, 70% of the data from the complete data sets is kept for training, and 30% of the data is kept for testing purposes. After obtaining the prediction results with the test data sets, some performance indices can be integrated to quantify the performance of the classification models. In this study, mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the R^2 coefficient have been utilized to reveal the performance of the classification models. The workflow process for the proposed classification methodology, along with the performance measurement, is illustrated in Figure 18.

4.3. Classification Performance Indices

MSE: MSE measures the average squared difference between the predicted fault locations and actual locations. It provides an indication of how far off the predictions are from the actual values.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_{\text{predicted},i} - y_{\text{actual},i})^2 \quad (13)$$

where, $y_{\text{predicted},i}$ and $y_{\text{actual},i}$ indicate the predicted and the actual values for the i^{th} sample, respectively. N is the total number of samples.

RMSE: RMSE is the square root of the MSE, providing a metric in the same units as the orig-

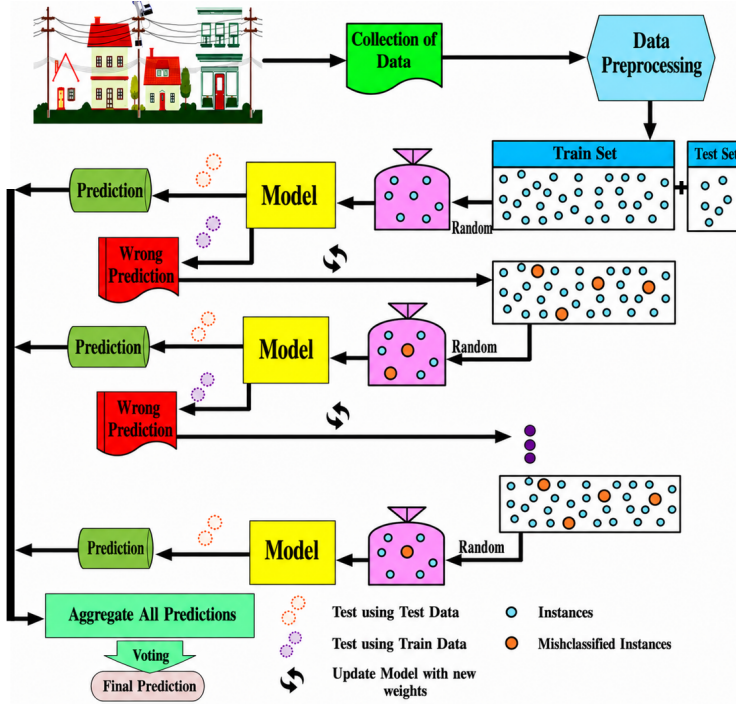


Figure 18: Workflow for fault classification and performance quantification of the proposed ML models.

inal fault location. It is more interpretable than MSE as it represents the average magnitude of error in the predicted fault locations.

$$RMSE = \sqrt{MSE} \quad (14)$$

MAE: MAE calculates the average of the absolute differences between predicted and actual values. Unlike RMSE, MAE treats all errors equally, regardless of their direction or magnitude.

$$MAE = \frac{\sum_{i=1}^n |y_{\text{predicted},i} - y_{\text{actual},i}|}{n} \quad (15)$$

R^2 coefficient: The R^2 coefficient is a statistical measure used to assess the goodness of fit of the predicted data points with respect to the actual values.

In this study, the RMSE and MAE values are reported in the same numerical scale as the encoded target labels; therefore, for fault type classification, they may be interpreted as fault-type index errors, while for fault location classifica-

tion, they may be interpreted as bus-index errors.

5. Results and Discussion

The data set generated by composing the fetched information from the 8 μ PMUs for the 19-bus test network is augmented, and a synthetic data set has been prepared. The classification accuracy has been tested for the fault location and fault types utilizing the synthetic data set. It is to mention here that the different types of faults are realized with numerical values. The $L_A - G$, $L_B - G$, $L_C - G$, $L_A - L_B$, $L_B - L_C$, $L_A - L_C$, $L_A - L_B - G$, $L_B - L_C - G$, $L_A - L_C - G$, and $L_A - L_B - L_C - G$ faults are designated with 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 respectively. A numerical value of 0 has been forwarded to indicate a non-faulty situation. The application of RF reveals that 99.7% and 88.9% testing classification accuracy is possible to be achieved for the identification of fault type and fault location, respectively. The calculated performance index values for the classification of

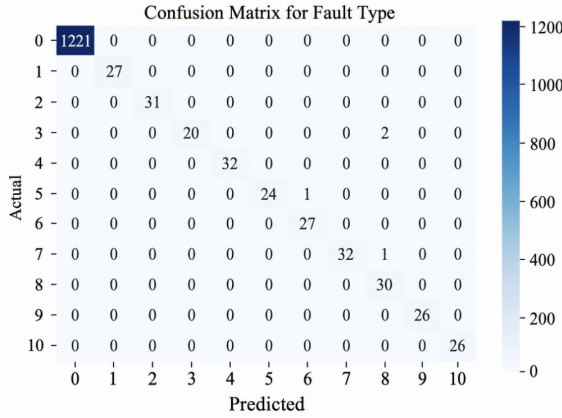


Figure 19: Confusion matrix obtained by application of RF for fault type classification.

fault type and fault location are summarized in Table 1.

Table 1: Performance indices obtained with application of RF for fault classification.

Classification	RMSE	MSE	MAE	R^2
Fault type	0.18	0.03	0.01	0.99
Fault location	1.86	3.48	0.45	0.80

The confusion matrix obtained by the application of RF while classifying the fault types and fault locations is furnished in Figure 19 and Figure 20, respectively.

When the model has been redesigned with an XGBoost classifier and tested utilizing the same synthetic data set, an accuracy of 99.7% and 89.5% has been achieved for the classification of fault type and fault location, respectively. The difference between the two accuracies reflects the intrinsic difference in task complexity. In general, fault type classification is associated with more distinguishable variations in the measured voltage and current phasors, whereas fault location identification requires finer spatial discrimination among network sections that may exhibit comparatively similar phasor patterns. Since the present framework utilizes μ PMU voltage and current phasor measurements collected after three cycles, the classification is primarily based on steady-state feature variations rather than on explicit mutual-impedance feedback parameters.

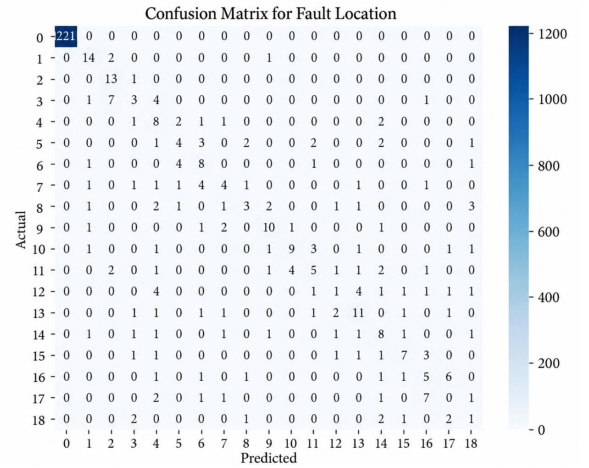


Figure 20: Confusion matrix obtained by application of RF for fault location classification.

The evaluated performance indices applying the XGBoost classifier are summarized in Table 2.

Table 2: Performance indices obtained with the application of XGBoost for fault classification.

Classification	RMSE	MSE	MAE	R^2
Fault type	0.139	0.019	0.00	0.99
Fault location	1.56	2.43	0.37	0.86

The confusion matrices generated by the XGBoost-based classification model for fault types and fault locations are shown in Figure 21 and Figure 22, respectively. For the fault-type classification task, labels 1 to 10 represent the considered fault categories, whereas label 0 indicates the non-faulty condition. For the fault-location classification task, label 0 represents the non-faulty condition, and labels 1 to 18 represent the 18 faulted line locations in the 19-bus system. These numeric labels are used only for classification and confusion-matrix evaluation.

The simulation test results suggest that the XGBoost-based classification model can perform better with respect to the RF-based classification model for fault classification.

The analysis has been further extended with a comparison of the proposed tree-based classification model with respect to the existing neural network-based classifier models. In this re-

Table 3: Comparative study of the proposed models with respect to the existing models for fault classification.

Classification	Model	RMSE	MSE	MAE	R^2
Fault type	RF-based model	0.186	0.035	0.008	0.994
	XGBoost-based model	0.139	0.019	0.00	0.997
	ANN-based model	0.045	0.002	0.01	0.998
Fault location	RF-based model	1.867	3.485	0.451	0.805
	XGBoost-based model	1.56	2.435	0.375	0.864
	ANN-based model	1.795	3.22	0.449	0.820

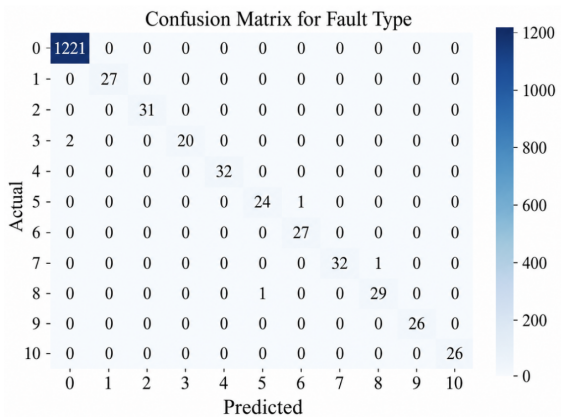


Figure 21: Confusion matrix obtained by XGBoost-based model for fault type classification.

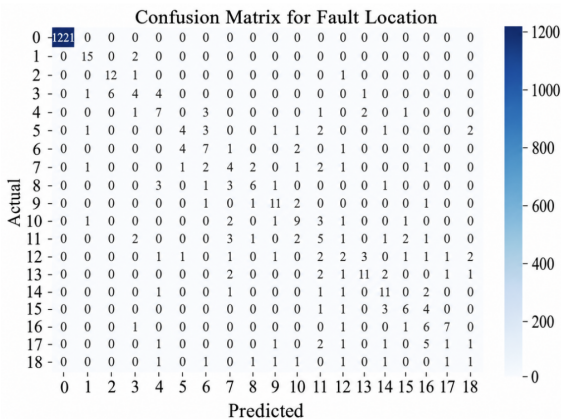


Figure 22: Confusion matrix obtained by application of XGBoost-based model for fault location classification.

gard, the existing neural network-based classifier models are further trained and tested with the same synthetic data set. The results of the classification quality are measured in terms of the performance indices. A comparative study with respect to the proposed RF-based and XGBoost-based classification models is summarized in Table 3. It can be comprehended from Table 3 that the proposed tree-based classification models can perform better with respect to the neural network-based classification models. It is also observed that, across all compared models, fault location remains more challenging than fault type classification, indicating that the lower localization accuracy is primarily due to the inherent complexity of spatial identification in the considered unbalanced network rather than a specific weakness of the XGBoost formulation. Here, the ANN is included as a baseline neural comparator to provide a fair reference against the proposed tree-based classifiers. The present study does not provide an exhaustive benchmark across all possible deep-learning architectures; rather, it focuses on structured μ PMU phasor-based fault classification and localization using RF and XGBoost. Since the available inputs are phasor features rather than sequential waveform data, architectures such as CNNs and LSTMs were not considered in the present comparison.

An analysis of the performance of the proposed models on the basis of raw data reveals that synthetic data has the capability to improve the classification accuracy for fault type identification; however, for fault location classification, synthetic data shows a marginal trade-off in RMSE despite improved classification accuracy and MAE, which is elaborated further in the following discussion. As an example, the comparative classification accuracy of the XGBoost-

Table 4: Comparative performance indices of XGBoost-based fault classification using raw and synthetic datasets.

Classification	Model	RMSE	MSE	MAE	R^2
Fault type	XGBoost-based model by raw data	0.603	0.363	0.043	0.938
	XGBoost-based model by synthetic data	0.139	0.019	0.00	0.997
Fault location	XGBoost-based model by raw data	1.511	2.283	0.443	0.874
	XGBoost-based model by synthetic data	1.56	2.435	0.375	0.864

based model simulated on raw and synthetic data are summarized in Table 4.

penetration is therefore a significant future direction for strengthening the generality of the method.

6. Conclusions

This study has revealed the importance of applying tree-based ML models to classify distribution line faults over neural network-based ML models. The performance of the classification models has been assessed using multiple performance indices. It has been observed that the XGBoost-based model performs better in classification than the RF and ANN-based models. The study also found that preprocessing-based dataset augmentation improves classification accuracy. While fault type classification accuracy has been very high for all the models, fault location classification accuracy has been slightly lower. Nevertheless, the 89.5% accuracy achieved by XGBoost is noteworthy and compares favorably with the existing approaches. The proposed classification model is designed to be robust and deliver efficient fault-type classification for all possible line-fault types across multiple phases.

Future research may focus on identifying the influencing factors in synthetic dataset generation and their impact on classification accuracy. Moreover, it may be useful to investigate whether a larger collection of μ PMU measurements can further improve performance. Since the proposed RF- and XGBoost-based classifiers learn from μ PMU phasor signals, substantial changes in short-circuit current due to higher distributed generation penetration may alter the feature pattern and therefore affect classification behavior if such cases are not included in training. Extending the framework to reconfigurable network topologies, broader short-circuit-current regimes, and higher distributed-source

Acknowledgement

The authors thank the executive editor and the reviewers of the Journal of Advanced Engineering and Computation for enriching the quality of the work presented in this paper.

References

- [1] Y. Zhang, J. Wang, and M. E. Khodayar. Graph-based faulted line identification using micro-PMU data in distribution systems. *IEEE Transactions on Smart Grid*, 11(5):3982–3992, 2020. DOI: [10.1109/TSG.2020.2988349](https://doi.org/10.1109/TSG.2020.2988349)
- [2] H. A. El-Ghany, I. A. Soliman, E. S. Ahmed, and A. E. ElGebaly. Generalized wide-area fault detection and location algorithm for transmission system based on optimal PMUs allocation. *International Journal of Electrical Power & Energy Systems*, 155(Part B):109634, 2024. DOI: [10.1016/j.ijepes.2023.109634](https://doi.org/10.1016/j.ijepes.2023.109634).
- [3] M. Jamei, R. Ramakrishna, T. Tesfay, R. Gentz, C. Roberts, A. Scaglione, and S. Peisert. Phasor Measurement Units Optimal Placement and Performance Limits for Fault Localization. *IEEE Journal on Selected Areas in Communications*, 38(1):180–192, 2020. DOI: [10.1109/JSAC.2019.2951971](https://doi.org/10.1109/JSAC.2019.2951971).
- [4] H. Mirshekali, R. Dashti, A. Keshavarz, and H. R. Shaker. Machine learning-based

- fault location for smart distribution networks equipped with micro-PMU. *Sensors*, 22(3):945, 2022. DOI: [10.3390/s22030945](https://doi.org/10.3390/s22030945).
- [5] K. Chen, J. Hu, Y. Zhang, Z. Yu, and J. He. Fault location in power distribution systems via deep graph convolutional networks. *IEEE Journal on Selected Areas in Communications*, 38(1):119–131, 2020. DOI: [10.1109/JSAC.2019.2951964](https://doi.org/10.1109/JSAC.2019.2951964).
 - [6] M. Mansour Lakouraj, M. Gautam, H. Livani, and M. Benidris. A multi-rate sampling PMU-based event classification in active distribution grids with spectral graph neural network. *Electric Power Systems Research*, 211:108145, 2022. DOI: [10.1016/j.epsr.2022.108145](https://doi.org/10.1016/j.epsr.2022.108145).
 - [7] H. Mirshekali, R. Dashti, A. Keshavarz, A. J. Torabi, and H. R. Shaker. A novel fault location methodology for smart distribution networks. *IEEE Transactions on Smart Grid*, 12(2):1277–1288, 2021. DOI: [10.1109/TSG.2020.3031400](https://doi.org/10.1109/TSG.2020.3031400).
 - [8] X. Lei, Z. Li, H. Jiang, S. S. Yu, Y. Chen, B. Liu, and P. Shi. Deep-learning based optimal PMU placement and fault classification for power system. *Expert Systems with Applications*, 292:128586, 2025. DOI: [10.1016/j.eswa.2025.128586](https://doi.org/10.1016/j.eswa.2025.128586).
 - [9] T. Malini, P. Thirumoorathi, and K. Lakshmi. ANN-based fault classification and localization with optimized PMU deployment for transmission systems. *Scientific Reports*, 15:39165, 2025. DOI: [10.1038/s41598-025-24955-z](https://doi.org/10.1038/s41598-025-24955-z).
 - [10] T. Zhang, J. Liu, H. Wang, Y. Li, N. Wang, and C. Kang. Fault diagnosis and protection strategy based on spatio-temporal multi-agent reinforcement learning for active distribution system using phasor measurement units. *Measurement*, 220:113291, 2023. DOI: [10.1016/j.measurement.2023.113291](https://doi.org/10.1016/j.measurement.2023.113291).
 - [11] M. Dashtdar, A. Hussain, H. Z. Al Garni, A. A. Mas’ud, W. Haider, K. M. AboRas, and H. Kotb. Fault location in distribution network by solving the optimization problem based on power system status estimation using the PMU. *Machines*, 11(1):109, 2023. DOI: [10.3390/machines11010109](https://doi.org/10.3390/machines11010109).
 - [12] A. H. Medattil Ibrahim, M. Sharma, and V. S. Rajkumar. Integrated fault detection, classification and section identification (I-FDCSI) method for real distribution networks using μ PMUs. *Energies*, 16(11):462, 2023. DOI: [10.3390/en16114262](https://doi.org/10.3390/en16114262).
 - [13] X. Li. High impedance fault location in power distribution systems using deep learning and micro phasor measurements units (μ PMU). *Multiscale and Multidisciplinary Modeling, Experiments and Design*, 7:1045–1056, 2023. DOI: [10.1007/s41939-023-00257-7](https://doi.org/10.1007/s41939-023-00257-7).
 - [14] D. Sodin, M. Smolnikar, U. Rudež, and A. Čampa. Precise PMU-based localization and classification of short-circuit faults in power distribution systems. *IEEE Transactions on Power Delivery*, 38(5):3262–3273, 2023. DOI: [10.1109/TPWRD.2023.3268767](https://doi.org/10.1109/TPWRD.2023.3268767).
 - [15] M. S. Attar and M. R. Miveh. High-impedance fault detection in distribution networks based on support vector machine and wavelet transform approach (case study: Markazi Province of Iran). *Energy Science & Engineering*, 13(3):1171–1183, 2025. DOI: [10.1002/ese3.2056](https://doi.org/10.1002/ese3.2056).
 - [16] M. N. I. Siddique, M. Shafiullah, S. Mekhilef, H. Pota, and M. A. Abido. Fault classification and location of a PMU-equipped active distribution network using deep convolution neural network (CNN). *Electric Power Systems Research*, 229:110178, 2024. DOI: [10.1016/j.epsr.2024.110178](https://doi.org/10.1016/j.epsr.2024.110178).
 - [17] D. Santoro, T. Ciano, and M. Ferrara. A comparison between machine and deep learning models on high stationarity data. *Scientific Reports*, 14:19409, 2024. DOI: [10.1038/s41598-024-70341-6](https://doi.org/10.1038/s41598-024-70341-6).
 - [18] I. Ayus, N. Natarajan, and D. Gupta. Comparison of machine learning and deep

- learning techniques for the prediction of air pollution: a case study from China. *Asian Journal of Atmospheric Environment*, 17(1):4, 2023. DOI: [10.1007/s44273-023-00005-w](https://doi.org/10.1007/s44273-023-00005-w).
- [19] N. Khanjani and S. M. Moghaddas-Tafreshi. An ILP model for stochastic placement of μ PMUs with limited voltage and current channels in a re-configurable distribution network. *International Journal of Electrical Power & Energy Systems*, 148:108951, 2023. DOI: [10.1016/j.ijepes.2023.108951](https://doi.org/10.1016/j.ijepes.2023.108951).
- [20] M. Mukherjee and B. K. Saha Roy. Optimal μ PMU placement in radial distribution networks using novel zero injection bus modelling. *SN Computer Science*, 4(5):575, 2023. DOI: [10.1007/s42979-023-02004-8](https://doi.org/10.1007/s42979-023-02004-8).
- [21] S. Maji, P. Kayal, and S. Ray. A fuzzy decision-based multiobjective planning for microphaser measurement unit placement in power distribution infrastructure. *IEEE Transactions on Industrial Informatics*, 20(5):7103–7111, 2024. DOI: [10.1109/TII.2024.3354328](https://doi.org/10.1109/TII.2024.3354328).
- [22] K. Redwan, M. M. R. Shamim, A. Sultana, D. H. Tuhin, N. Hannan, M. Ahmed, and A. Shufian. Intelligent fault diagnosis in power systems using Random Forest and ensemble learning. *Measurement: Energy*, 10:100098, 2026. DOI: [10.1016/j.meae.2026.100098](https://doi.org/10.1016/j.meae.2026.100098).
- [23] D. Kulshrestha and V. Chaudhary. Enhancing electrical fault detection in transmission lines through machine learning. *2024 IEEE 3rd International Conference on Electrical Power and Energy Systems (ICEPES)*, Bhopal, India, 2024. DOI: [10.1109/ICEPES60647.2024.10653619](https://doi.org/10.1109/ICEPES60647.2024.10653619).
- [24] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. DOI: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).
- [25] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).

About Authors

Sukalyan Maji received the B.Tech degree in Electrical Engineering from West Bengal University of Technology in 2012. He earned the M.Tech degree in Electrical Devices and Power Systems from Maulana Abul Kalam Azad University of Technology in 2016. He completed his Ph.D. in Electrical Engineering at the National Institute of Technology Silchar. A Ministry of Education fellowship supported his doctoral research. He is currently an Assistant Professor and tenured faculty member in the Department of Electrical Engineering at Veermata Jijabai Technological Institute in Mumbai. His research focuses on smart grid automation and unbalanced power distribution network reconfiguration. He actively investigates microgrid topology identification and networked microgrids. His studies address power system observability utilizing microphasor measurement units. He explores fault type and location identification in distribution systems. Additionally, he researches electric vehicle charging integration and applies machine learning models to enhance power system security. He can be contacted at email: connectsukalyanmaji@gmail.com.

Deepak Kumar Singh completed his B.Tech. degree in Electrical Engineering from the Department of Electrical Engineering, National Institute of Technology Silchar, India in 2025. His research interests include monitoring of power systems, fault analysis and diagnosis strategies, application of machine learning techniques and reliability of power supply. He can be reached via email: deepakk21_ug@ee.nits.ac.in.

Emon Das completed his bachelor degree in Electrical Engineering from the Department of Electrical Engineering, National Institute of Technology Silchar, India in 2025. His research interests include application phasor measurement units in power systems, fault analysis and diagnosis strategies, and application of machine learning techniques. He can be contacted at email: emon21_ug@ee.nits.ac.in.

Partha Kayal received the Ph.D. degree in Electrical Engineering from the Indian Institute of Engineering Science and Technology Shibpur in 2016. Presently he is working as an Assistant Professor at the Department of Electrical Engineering, National Institute of Technology Silchar, India. He has published more than fifty papers in peer reviewed journals and conferences. He is an Editorial Board Member of the Scientific Reports journal. His research interests include optimization of distributed energy resources, grid integration of renewable energies, grid modernization considering energy storage and electric vehicles, reliability and power quality issues, statistical modelling and forecasting of energy, power distribution system monitoring and application of micro phasor measurement units, resiliency enabling approaches, and networked microgrids. He can be contacted at email: partha@ee.nits.ac.in.