Leakage-Aware Fault Detection in Three-Phase Power Systems Using an Artificial Neural Network and Scenario-Based Data Splitting

Verdi Yasin[1], Sri Mardiyati[2], Ifan Junaedi[3], Rumadi Hartawan[1], Ito Riris Immasari[4], Ruben Cornelius Siagian[5*]

[1]*Department of Informatics Engineering, Sekolah Tinggi Manajemen Informatika dan Komputer Jayakarta, Jakarta*

[2]*Department of Informatics Engineering, Universitas Indraprasta PGRI*

[3]*Department of Information System, Sekolah Tinggi Manajemen Informatika dan Komputer Jayakarta*

[4]*Department of Informatics Managemen, Sekolah Tinggi Manajemen Informatika dan Komputer Jayakarta, Jakarta*

[5]*PT Siagian Global Research, Tugumulya Area, Margamulya Village, Pangalengan District, Bandung Regency, West Java 40378, Indonesia*

[*]*Corresponding author. Email: rubensiagian17april@gmail.com*

Abstract— The reliability of three-phase power systems depends heavily on the ability to detect faults quickly and accurately. However, much previous research has faced challenges such as data leakage, biased evaluations due to random data splitting, and limitations in the utilisation of system dynamics-based features. This study aims to develop a machine learning-based fault detection model that is accurate, robust, and possesses high generalisation capabilities. The method employed is a quantitative approach using an Artificial Neural Network (ANN) model as a binary classifier (normal and fault). The research process includes data preprocessing, elimination of features that could potentially cause data leakage, feature engineering based on current, voltage, energy, and system imbalance, as well as scenario-based data splitting to ensure the validity of the evaluation. The model was trained using repeated cross-validation and optimised based on the Area Under the Curve (AUC) metric. The results show that the ANN model achieved high performance with an AUC of 0.987 and an accuracy of approximately 98.5% and remained stable under varying conditions such as noise and load changes. Compared to other models, Random Forest delivered the best performance, followed by Support Vector Machine (SVM) and ANN. This study addresses the gap regarding more realistic evaluation methods that are free from data leakage. Its novelty lies in the use of scenario-based splitting, comprehensive evaluation, and the finding that raw signal features make a dominant contribution to fault detection.

Keywords— Artificial neural network; Energy metrics; Machine learning; Phase space analysis; Power system disturbances; Three-phase systems

I. INTRODUCTION

The three-phase electrical power system is a vital piece of infrastructure that underpins various industrial sectors and everyday life; therefore, its reliability must be always maintained [2]. In practice, faults in three-phase systems can have serious consequences, ranging from equipment damage and disruption to operational processes to significant financial losses if they are not detected quickly and accurately [3], [4]. However, conventional fault detection methods generally still rely on threshold-based approaches or specific rules, which have limitations when it comes to capturing the complex dynamics of modern power systems [5]. The characteristics of electrical signals such as current, voltage and energy in three-phase systems are non-linear and dynamic, making it difficult to model them optimally using simple analytical approaches [6]. With the advancement of technology, machine learning offers the opportunity to build fault detection models capable of learning directly from data and recognising complex patterns that are difficult to identify manually. However, the application of machine learning in power systems is not without its challenges, such as the potential for data leakage, data quality, and the need for appropriate feature representation to ensure that the model can generalise effectively [7]. Furthermore, much previous research has not considered the validity of evaluation, particularly about data splitting, which has the potential to introduce bias or lead to an overestimation of model performance. The selection and engineering of features capable of accurately representing the physical conditions of the system also remain significant challenges, particularly as the use of certain derived features may contain implicit information regarding the label, which risks causing the model to learn from leaked information rather than from the system's actual characteristics. Therefore, an approach is required that can integrate current, voltage, and energy data with appropriate processing techniques and feature engineering to improve the quality of data representation. On the other hand, robust and bias-free model validation is a crucial aspect to ensure that the developed model can be applied to real-world conditions with a high level of reliability. Performance evaluation must also be conducted comprehensively using various classification metrics and probability calibration analysis to ensure the quality of the predictions produced. Furthermore, testing the model's resilience to variations in system conditions such as noise, load changes and fault variations, remains a crucial aspect that has yet to be explored in depth. Considering this, this research is essential to develop a machine learning-based fault detection model that is not only accurate but also robust,

Received 7 January 2026, Revised 19 March 2026, Accepted 5 April 2026.
DOI: https://doi.org/10.15294/jte.v18i1.40675

unbiased, and capable of representing the conditions of the power system more realistically. Consequently, this research is expected to contribute to the development of more reliable and practical fault detection systems for modern power systems, which are becoming increasingly complex.

The hypothesis of this study states that a machine learning model based on an Artificial Neural Network (ANN) is capable of accurately classifying the conditions of a three-phase power system (normal and fault) by utilising current, voltage and energy features that possess high discriminatory power. By eliminating variables that could potentially cause data leakage, and by applying a scenario-based data splitting approach, it is hoped that the model will demonstrate better generalisation capabilities and more realistic evaluation. Preprocessing steps such as median imputation and standardisation, along with the application of feature engineering, are believed to enhance data representation quality and classification performance. Furthermore, the use of regularisation in the ANN can reduce overfitting and maintain model stability. Area Under the Curve (AUC)-based evaluation is expected to provide a more robust measure of performance, whilst determining the optimal threshold using the Youden Index can balance sensitivity and specificity. The developed model is also expected to perform competitively compared to the baseline model, as well as demonstrate a high degree of robustness against noise and variations in system conditions. Furthermore, the model is expected to exhibit good generalisation capabilities on new data, produce well-calibrated prediction probabilities, and demonstrate statistically significantly superior performance compared to the benchmark method.

This study aims to develop a machine learning-based fault detection model for three-phase power systems, specifically using an ANN, capable of classifying normal and fault conditions based on the characteristics of current, voltage, and energy signals. Furthermore, this study also aims to produce a model with high accuracy and generalisation capabilities through the application of pre-processing, feature engineering, and appropriate validation, whilst avoiding data leakage so that the model's performance reflects real-world conditions. The optimisation process is carried out through cross-validation and AUC-based hyperparameter tuning, as well as evaluation using various metrics such as accuracy, precision, recall, F1-score, and Kappa. This study also compares the performance of ANNs with other methods such as Logistic Regression, Support Vector Machine (SVM), and Random Forest, and tests the model's robustness against noise, load variations, and changes in system conditions. The benefits of this study include providing an artificial intelligence-based solution for the early detection of faults in three-phase power systems, thereby enhancing system reliability and safety. Furthermore, this research can serve as a reference in the development of smart monitoring and predictive maintenance systems, as well as providing methodological contributions to preventing data leakage in machine learning research. The results of this study also highlight the importance of domain-based feature engineering in improving model performance, offer insights into the comparison of machine learning algorithms for fault detection, and support the development of systems that are robust against various operating conditions. Overall, this research contributes to the advancement of knowledge and the application of machine learning within the field of power system engineering.

This study has several limitations that should be noted. Firstly, the use of a dataset derived from a single publication means that the range of power system conditions represented remains limited and thus does not fully reflect the complexity of real-world conditions. This also affects the model's generalisability, as although data splitting and robustness tests were carried out, all data still originate from a relatively controlled environment. Furthermore, the assumption of scenario division based on 100 observations per block may not always align with the actual dynamics of power systems, which can change at a faster or slower rate. On the other hand, the removal of certain variables to prevent data leakage risks reducing the information that is relevant to the model. The median-based imputation approach is also relatively simple and may be less than optimal for complex data patterns. The model used an ANN with a simple architecture has limitations in capturing more complex non-linear relationships compared to deeper deep learning models, particularly given the still-limited hyperparameter search space. Furthermore, this study focuses solely on binary classification between normal and fault conditions without distinguishing specific types of faults, meaning diagnostic information remains limited. The reliance on current, voltage, and energy-based features also does not encompass all characteristics of the power system, such as frequency or harmonic aspects. The robustness evaluation conducted is also limited to specific scenarios, such as Gaussian noise and load variations, and thus does not reflect all possible real-world disturbances such as data drift or sensor failure. Furthermore, the relatively balanced data distribution means the model's performance under unbalanced data conditions has not been tested. Although probability calibration has been analysed using metrics such as the Brier Score and ECE, its implications for operational decision-making have not been evaluated in depth. Finally, although the data is time-series in nature, the models used do not explicitly capture temporal dependencies, meaning that the temporal dynamics within the power system have not been optimally utilised.

Recent research into fault detection in three-phase power systems indicates that non-linear machine learning-based approaches such as Random Forest, Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs) have become the dominant methods due to their ability to capture non-linear relationships in electrical signals. Furthermore, the majority of studies utilise domain-specific features, such as three-phase current and voltage, energy-based and imbalance features, as well as various signal transformations such as Root Mean Square (RMS) or wavelet, to enhance the model's discriminative capability. However, in terms of evaluation, many studies still employ conventional approaches such as random train-test splits and standard cross-validation, with accuracy and AUC as the primary metrics. This leads to research focusing primarily on achieving high performance, without considering the overall validity of the evaluation. Consequently, although many studies report very high AUC values, potential issues such as data leakage, temporal or data scenario dependencies, and a lack of analysis regarding the model's robustness and generalisation ability are often overlooked.

Given these circumstances, there are several significant Given these circumstances, there are several significant research gaps. Firstly, many studies still overlook the issue of data leakage, where derived features directly related to prediction outcomes, such as model probabilities or thresholds are used as inputs, resulting in seemingly high performance that does not reflect the model's true capability. Secondly, the use of random splits in data partitioning risks causing data from identical system conditions to appear in both the training and test sets, leading to unrealistic evaluations that tend to

overestimate performance. Thirdly, most research has not conducted adequate analysis of model robustness and generalisation, particularly regarding variations in conditions such as noise, load changes, and variations in disturbance levels. Fourthly, model evaluation generally focuses solely on classification performance without considering the quality of prediction probabilities, meaning that probability calibration metrics such as the Brier Score or Expected Calibration Error are still rarely analysed. Fifthly, although feature engineering is widely used, there is still little research that systematically evaluates the contribution of each feature group through an ablation study approach; consequently, it remains unclear whether complex features provide significant benefits compared to raw signals. Many studies lack robust statistical validation, such as the use of confidence intervals, significance tests and effect sizes, meaning that the superiority of the models is often not rigorously demonstrated, either statistically or in practice.

This research offers a novel approach through the development of a modelling framework that explicitly accounts for potential data leakage by eliminating features that implicitly contain prediction results, thereby ensuring that the model learns solely from the physical characteristics of the power system. Furthermore, this research introduces a scenario-based data splitting approach that replaces the conventional random split method, making model evaluation more realistic and reflecting actual operational conditions without information overlap between data subsets. In terms of data representation, this research integrates various domain-based features such as mean values, phase imbalance, RMS, phase space radius, and energy ratio, which are then systematically validated through an ablation study to ensure the actual contribution of each feature group to model performance. Furthermore, the novelty of this research lies in the evaluation, which not only focuses on accuracy but also includes an analysis of robustness against various disturbances such as noise, load variations and changes in disturbance levels, thereby demonstrating that the model remains stable under dynamic conditions. Moreover, the quality of the probability predictions is analysed using calibration metrics such as the Brier Score and Expected Calibration Error (ECE), which provide a more in-depth perspective compared to conventional classification evaluations. This research is also complemented by a comprehensive statistical validation framework through bootstrap analysis, confidence intervals, and significance tests and effect sizes, thereby ensuring that the performance improvements obtained are consistent and practically meaningful. Interestingly, the results of the ablation study show that raw signals make the most dominant contribution compared to engineered features, and indicate potential redundancy in phase-space features, thereby providing new insights that feature complexity does not always correlate directly with improved model performance.

II. METHOD

This research employs a quantitative approach based on machine learning to develop a fault detection model for three-phase power systems. The model developed is a binary classifier designed to distinguish between normal and fault conditions based on current and voltage signal characteristics, as well as energy features. The research process was conducted systematically, ranging from data acquisition, pre-processing, feature engineering, and scenario-based data splitting, through to model training and evaluation.

The data used is taken from publications [8], which provides a dataset of current and voltage measurements under various three-phase system fault conditions. This dataset serves as a crucial foundation for research as it accurately represents the dynamics of power systems, particularly through current components (I_a, I_b, I_c), voltage components (V_a, V_b, V_c), and energy features, thereby supporting the development of reliable fault detection models.

In the initial stage, data was read from a spreadsheet file and the variable structure was identified using data inspection functions. The target variable in this study was Output (S), which was then transformed into a categorical label: Normal for class 0 and Fault for class 1. The output label for each sample is defined in Equation (1):

$$Y_i = \begin{cases} 0, & \text{if the sample is -}i\text{ in a normal state} \\ 1, & \text{if the sample is -}i\text{ in a state of disruption} \end{cases} \quad (1)$$

where (i = 1, 2, 3, …, N), and (N) denotes the total number of samples in the dataset. As this study involves binary classification, each observation has only one output label.

Before modelling is carried out, the first crucial step is to identify variables that have the potential to cause data leakage. In this study, several columns such as ANN_Pred_Prob, Energy_Fault_dyn, Threshold_upper, and Threshold_lower are treated as candidates for information leakage because, conceptually, they contain derived information that is very closely related to the prediction results or final decision. If variables such as these are used directly as model inputs, the model's performance may appear very high but does not reflect its actual generalisation ability. Therefore, these variables are excluded from the set of predictors. The complete set of variables considered in this study is defined in equation (2):

$$\mathcal{X}_{all} = x_1, x_2, x_3, \dots, x_p \quad (2)$$

The set of features identified as potential sources of data leakage is defined in equation (3):

$$\mathcal{L} = \{\text{ANN\_Pred\_Prob, Energy\_Fault\_dyn, Threshold\_upper, Threshold\_lower}\} \quad (3)$$

Therefore, the final set of predictors used in training the model is defined in equation (4):

$$\mathcal{X}_{\text{strict}} = \mathcal{X}_{\text{all}} \setminus (\mathcal{L} \cup \{Y\}) \quad (4)$$

where (Y) is the target variable. This step is taken to ensure that the model truly learns from the physical characteristics of the system, rather than from variables that implicitly contain the answer. Next, a basic statistical analysis of the dataset is performed. The total number of samples, denoted by (N), is calculated using equation (5):

$$N = \sum_{i=1}^{N} 1 \quad (5)$$

where the number of predictor features is denoted by (p). The class distribution is calculated using the frequency of each label. If (N_0) represents the number of samples in the normal class and (N_1) represents the number of samples in the fault class, then the proportion of the normal class is defined in equation (6), while the proportion of the fault class is defined in equation (7):

$$P(Y = 0) = \frac{N_0}{N}, \quad (6)$$

$$P(Y = 1) = \frac{N_1}{N} \quad (7)$$

In addition, to describe the prevalence of disorder labels, the concepts of label cardinality and label density are used. As this is a binary classification problem, the label cardinality is

defined as the average number of positive label occurrences across the entire dataset, as expressed in equation (8):

$$\text{Label Cardinality} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad (8)$$

where $(y_i \in 0,1)$. Since there is only one possible label, the label density is calculated using equation (9):

$$\text{Label Density} = \frac{\text{Label Cardinality}}{1} \quad (9)$$

This value provides an indication of how prevalent missing values are across the entire dataset. The next step is to analyse missing values. For each feature (x_j), the number of missing values is calculated using equation (10):

$$M_j = \sum_{i=1}^{N} \mathbb{I}\,(x_{ij} \text{ missing}) \quad (10)$$

where $\{\mathbb{I}(\cdot)\}$ is an indicator function that takes the value 1 if an observation is missing and 0 otherwise. The percentage of missing values in the (j)-th feature is calculated using equation (11):

$$\text{MissingPct}_j = \frac{M_j}{N} \times 100 \quad (11)$$

This analysis is conducted to ensure data quality prior to the model training stage. Based on the pipeline used, most key features, such as current, voltage and fundamental energy, exhibit high completeness, whilst some derived features show a small, reasonable proportion of missing values due to the window-based feature engineering process.

To maintain the validity of the evaluation, this study does not use the standard random split per observation but instead employs a scenario-based split. The data is divided into blocks of observations representing a single operational scenario. In practice, every 100 observations are grouped into a single scenario. The scenario identifier for the i-th observation is calculated using equation (12):

$$\text{scenario\_id}_i = \left\lceil \frac{i}{100} \right\rceil \quad (12)$$

where $(\lceil\cdot\rceil)$ denotes the rounding-up function. If the total number of samples is (N), the number of scenarios is calculated using equation (13):

$$S = \left\lceil \frac{N}{100} \right\rceil \quad (13)$$

This approach was chosen because observations that are chronologically close together often originate from the same or very similar system conditions. If observations within a single scenario are randomly distributed across the training, validation and test sets, the model may derive information indirectly from nearly identical patterns. This situation would result in an overly optimistic evaluation. Therefore, all observations within a single scenario must be placed in only one of the data subsets. Once the scenarios have been determined, the proportion of fault classes for each scenario is calculated using equation (14):

$$r_s = \frac{1}{n_s}\sum_{i \in s} \mathbb{I}\,(Y_i = 1) \quad (14)$$

where (ns) is the number of observations in the (s)-th scenario. The majority class in each scenario was determined using (15), where scenarios with a fault ratio greater than or equal to 0.5 were classified as Fault, while the remaining scenarios were classified as Normal.

$$C_s = \begin{cases} \text{Fault}, & \text{If } r_s \geq 0.5 \\ \text{Normal}, & \text{If } r_s < 0.5 \end{cases} \quad (15)$$

A stratified partition was then performed on the scenario table so that the proportions of scenarios dominated by normal and fault conditions were maintained in the training, validation and test data. The dataset was divided into 70% training data, 15% validation data, and 15% test data. Let $(\mathcal{S})$ denote the set of all operating scenarios. The scenario-based data partitioning is defined in equation (16):

$$\mathcal{S} = \mathcal{S}train \cup \mathcal{S}val \cup \mathcal{S}_{test} \quad (16)$$

where $(\mathcal{S} * train)$, $(\mathcal{S} * val)$, and $(\mathcal{S}_{test})$ represent the training, validation, and testing scenario sets, respectively. Furthermore, the non-overlapping condition between subsets is defined in equation (17):

$$\mathcal{S}_{train} \cap \mathcal{S}_{val} = \emptyset, \mathcal{S}_{train} \cap \mathcal{S}_{test} = \emptyset, \mathcal{S}_{val} \cap \mathcal{S}_{test} = \emptyset \quad (17)$$

equation (17) ensures that no scenario overlap exists between the training, validation, and testing subsets. This condition was explicitly verified to prevent data leakage during model evaluation.

Once the data has been split, pre-processing is performed on the training data, and the resulting pre-processing parameters are applied to the validation and test data. The first step in pre-processing is median imputation. If a value (x_{ij}) is missing, it is replaced using the median value of the (j)-th feature calculated from the training data, as defined in equation (18):

$$x_{ij}^{imp} = \begin{cases} x_{ij}, & \text{If } x_{ij} \text{ available} \\ \text{median}(x_{1j}, x_{2j}, \ldots, x_{n_{\text{train}}j}), & \text{If } x_{ij} \text{ missing} \end{cases} \quad (18)$$

The median value in equation (18) was selected instead of the mean because it is more robust to outliers.

The second step involved standardisation through centring and scaling. For each feature (x_j), the mean of the training data was calculated using equation (19):

$$\mu_j = \frac{1}{n_{train}}\sum_{i=1}^{n_{train}} x_{ij} \quad (19)$$

$$\sigma_j = \sqrt{\frac{1}{n_{train}-1}\sum_{i=1}^{n_{train}} (x_{ij} - \mu_j)^2} \quad (20)$$

The standard deviation was computed using equation (20). Each observation was then transformed into a standardised score according to equation (21):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (21)$$

The transformation defined in equation (21) ensures that all features are placed on a comparable scale, which is essential for machine learning algorithms such as ANN and SVM. Following standardisation, feature engineering is carried out to enrich the representation of the power system data. The first engineered features are the average values across the three phases for current, voltage, and energy, which are calculated using (22), (23), and (24), respectively.

$$I_{mean} = \frac{I_a + I_b + I_c}{3} \quad (22)$$

$$V_{mean} = \frac{V_a + V_b + V_c}{3} \quad (23)$$

$$E_{mean} = \frac{E_a + E_b + E_c}{3} \quad (24)$$

Equations (22)–(24) are used to represent the aggregate behaviour of the three-phase system and to capture the overall operational trends of current, voltage, and energy. The next feature is a measure of the imbalance between phases relative to the norm (L_1). For current imbalance, the unbalance index is calculated using equation (25), while the voltage imbalance index is defined in Equation (26), and the energy imbalance index is computed using Equation (27):

$$I_{unbalance}^{L1} = |I_a - I_b| + |I_b - I_c| + |I_c - I_a| \quad (25)$$

$$V_{unbalance}^{L1} = |V_a - V_b| + |V_b - V_c| + |V_c - V_a| \quad (26)$$

$$E_{unbalance}^{L1} = |E_a - E_b| + |E_b - E_c| + |E_c - E_a| \quad (27)$$

The higher this value, the greater the degree of asymmetry between the phases, which in practice is often associated with the occurrence of faults. To capture the effective magnitude of the signal, three-phase RMS proxy features are calculated using (28), (29), and (30). equation (28) defines the RMS current feature, equation (29) defines the RMS voltage feature, and equation (30) defines the RMS energy feature:

$$I_{rms} = \sqrt{\frac{I_a^2 + I_b^2 + I_c^2}{3}} \quad (28)$$

$$V_{rms} = \sqrt{\frac{V_a^2 + V_b^2 + V_c^2}{3}} \quad (29)$$

$$E_{rms} = \sqrt{\frac{E_a^2 + E_b^2 + E_c^2}{3}} \quad (30)$$

This RMS feature is important because electrical disturbances often affect the effective amplitude of the signal. This study also develops phase-space radius features to geometrically represent the distance of a measurement point from the origin in three-dimensional space. The current phase-space radius is defined in (31), the voltage phase-space radius is defined in (32), and the energy phase-space radius is defined in (33):

$$R_I = \sqrt{I_a^2 + I_b^2 + I_c^2} \quad (31)$$

$$R_V = \sqrt{V_a^2 + V_b^2 + V_c^2} \quad (32)$$

$$R_E = \sqrt{E_a^2 + E_b^2 + E_c^2} \quad (33)$$

This feature provides a concise representation of the total energy of the system's state in the measurement space. In addition, the ratio of asymmetric energy to total energy is calculated using (34):

$$R_{asym} = \frac{Energy_{asym}}{|Energy_{total}| + \varepsilon} \quad (34)$$

with $(\varepsilon = 10^{-8})$ as a small constant to prevent division by zero. This ratio measures the magnitude of the asymmetric energy component relative to the total system energy, making it useful for capturing changes in conditions caused by faults.

Once the engineered features have been formed, median imputation is performed again to ensure that all new features are free of missing values, particularly if the formation of ratios or transformations results in undefined values for some observations.

The primary model used is an ANN for binary classification. Generally, an ANN maps an input vector $(\mathbf{x} \in \mathbb{R}^p)$ to a probability output $(\hat{y})$ via a combination of linear and non-linear activation functions. For a single neuron in the hidden layer, the activation value is calculated using equation (35):

$$a_k = \sum_{j=1}^{p} w_{jk}\, x_j + b_k \quad (35)$$

The weighted input (a_k) is then passed through the sigmoid activation function, as defined in equation (36):

$$h_k = \sigma(a_k) = \frac{1}{1 + e^{-a_k}} \quad (36)$$

If the hidden layer contains H neurons, the output layer value can be calculated using equation (37):

$$z = \sum_{k=1}^{H} v_k\, h_k + c \quad (37)$$

The probability of the fault class is calculated using the sigmoid activation function, as expressed in equation (38):

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (38)$$

Thus, $(\hat{y})$ lies within the interval ([0,1]) and can be interpreted as the probability that an observation belongs to the fault class. The ANN training process is carried out by minimising the cross-entropy loss function, as defined in equation (39):

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1-\hat{y}_i\right)\right] \quad (39)$$

Since the model employs weight decay regularisation, the optimisation objective is modified into the regularised loss function shown in equation (40):

$$\mathcal{J} = \mathcal{L} + \lambda \sum w^2 \quad (40)$$

where (λ) is the regularisation parameter, and $(\sum w^2)$ is the sum of the squares of the network weights. This regularisation aims to reduce the model's complexity, thereby mitigating the risk of overfitting.

In this study, model training was conducted using repeated cross-validation with a 5-fold scheme and 3 repeats on the training data. This means that the training data was divided into five folds, and the training and cross-validation process was repeated three times with different divisions. For each combination of hyperparameters, model performance was measured using the AUC. The AUC value was chosen as the optimisation metric because it is more stable in assessing the model's discriminative ability than accuracy alone, particularly when class proportions are not strictly identical. The hyperparameters tuned were the number of neurons in the hidden layer (size) and the regularisation parameter (decay), with a search space (size $\in$ 3,5,7,9), and (decay $\in$ 0.001,0.01,0.1). The best model was selected based on the highest ROC score in repeated cross-validation.

Once the best model has been obtained, it is evaluated on the validation and test data. For each observation, the model generates a predicted probability of the fault class, denoted as $(\hat{y}_i)$. Based on this set of probabilities, a Receiver Operating Characteristic (ROC) curve is constructed by calculating the True Positive Rate (TPR) and False Positive Rate (FPR) at various classification thresholds, where the True Positive Rate, also referred to as sensitivity, is defined in equation (41) as:

$$TPR = \frac{TP}{TP + FN} \quad (41)$$

and the False Positive Rate is defined in equation (42) as

$$FPR = \frac{FP}{FP + TN} \quad (42)$$

The overall classification performance is then evaluated using the Area Under the ROC Curve (AUC), which is expressed in equation (43) as

$$AUC = \int_0^1 TPR\{FPR^{-1}(u)\}\, du \quad (43)$$

The closer the value is to 1, the better the model's ability to distinguish between normal and fault classes. To determine the optimal classification threshold, this study employed the Youden Index, as expressed in equation (44):

$$\mathrm{J} = \mathrm{Sensitivity} + \mathrm{Specificity} - 1 \tag{44}$$

In equation (44), the specificity value is calculated using equation (45):

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{45}$$

Based on the Youden Index defined in equation (44), the optimal threshold $(\hat{\tau})$ is determined using equation (46):

$$(\hat{\tau}) = \arg\max_{\tau}\{\mathrm{Sensitivity}(\tau) + \mathrm{Specificity}(\tau) - 1\} \tag{46}$$

Once the optimal threshold $(\hat{\tau})$ has been obtained from the validation data, the class predictions for both the validation and test datasets were determined using the decision rule defined in equation (47):

$$\hat{c}_i = \begin{cases} \text{Fault}, & \text{If } \hat{y}_i \geq \tau^* \\ \text{Normal}, & \text{If } \hat{y}_i < \tau^* \end{cases} \tag{47}$$

From the results of this classification, a confusion matrix was constructed comprising True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Based on this confusion matrix, the evaluation metrics were calculated in detail. Accuracy is defined in equation (48) as

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{48}$$

Precision is calculated using equation (49):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{49}$$

Recall, also known as sensitivity, is defined in equation (50) as

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{50}$$

Specificity is calculated using equation (51):

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{51}$$

The F1-score, which represents the balance between precision and recall, is calculated using equation (52):

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{52}$$

Meanwhile, balanced accuracy is defined in equation (53) as

$$\mathrm{Balanced\ Accuracy} = \frac{\mathrm{Sensitivity} + \mathrm{Specificity}}{2} \tag{53}$$

In some parts of the analysis, the Kappa coefficient was also used to measure the model's agreement with the actual labels after adjusting for the probability of random agreement. The general formula for Cohen's Kappa is presented in equation (54):

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{54}$$

where (p_o) is the observed agreement proportion and (p_e) is the expected random agreement proportion.

To ensure fairness in the benchmark, this study also compares the ANN with several baseline models, namely Logistic Regression (GLM), Radial SVM, and Random Forest (RF). All models were trained on the same data subset, using the same features, and undergoing the same preprocessing, so that performance differences truly reflect the capabilities of each algorithm. In the baseline comparison, the classification threshold was standardised to $(\tau = 0.5)$ to maintain consistency in the evaluation.

In addition to classification evaluation, this study also performed probability calibration analysis. One of the evaluation metrics used in this study is the Brier Score, which is defined in equation (55):

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{55}$$

Equation (55) measures the mean squared difference between the predicted probabilities and the actual outcomes. A lower Brier Score indicates better agreement between the predicted probabilities and the true labels. The Expected Calibration Error (ECE), defined in equation (56), is used to measure the average calibration error between predicted probabilities and observed outcomes:

$$\mathrm{ECE} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \frac{|\mathrm{B_m}|}{\mathrm{N}} |\mathrm{acc}(\mathrm{B_m}) - \mathrm{conf}(\mathrm{B_m})| \tag{56}$$

where (B_m) is the (m)-th probability bin, $\{\mathrm{acc}(B_m)\}$ is the actual positive frequency in that bin, and $\{\mathrm{conf}(B_m)\}$ is the average predicted probability in that bin.

Robustness analysis was also conducted to assess the model's stability against external disturbances and variations in operating conditions. In the noise-addition scenario, the signal was modified according to equation (57):

$$x_i' = x_i + \epsilon_i \tag{57}$$

where the noise term (ϵ_i) follows a Gaussian distribution as defined in equation (58):

$$\epsilon_i \sim \mathcal{N}(0, \sigma_{noise}^2) \tag{58}$$

In the load variation scenario, the phase currents were scaled by a factor (α), as expressed in equation (59):

$$I_a' = \alpha I_a, I_b' = \alpha I_b, I_c' = \alpha I_c \tag{59}$$

Meanwhile, in the fault-level variation scenario, the total energy feature was scaled by a factor (β), as defined in equation (60):

$$\mathrm{Energy}'_{\mathrm{total}} = \beta \cdot \mathrm{Energy}_{\mathrm{total}} \tag{60}$$

For each perturbation condition described in equations (57)–(60), the AUC value was recalculated to evaluate the model's sensitivity to variations in the input data.

To test the reliability of performance, this study also conducted bootstrap-based experiments or repeated runs. In each repetition, the sample is resampled and the AUC is calculated. From the set of AUC values $(AUC_1, AUC_2, \ldots, AUC_R)$, the average model performance is calculated using equation (61):

$$\overline{AUC} = \frac{1}{R}\sum_{r=1}^{R} AU\,C_r \tag{61}$$

The variability of the AUC scores across repeated experiments is then measured using the standard deviation defined in equation (62):

$$SD_{AUC} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(AU\,C_r - \overline{AUC})^2} \tag{62}$$

Furthermore, the 95% confidence interval for the mean AUC is computed using equation (63):

$$CI_{95\%} = \overline{AUC} \pm t_{0.975,R-1} \cdot \frac{SD_{AUC}}{\sqrt{R}} \quad (63)$$

To assess the significance of the difference in performance between the ANN model and the baseline, statistical tests such as the paired t-test and the Wilcoxon signed-rank test were applied to the results of the experimental runs. The effect size was calculated using equation (64), which measures the standardized difference between the mean performances of the two models:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_d} \quad (64)$$

where (s_d) is the standard deviation of the difference in performance between the two models. This effect size is important for demonstrating that the difference in performance is not only statistically significant but also practically meaningful.

III. Results and Discussion

A. Experimental Setup and Evaluation Protocol

Based on the results of the data completeness analysis, the dataset used in this study is of very high quality, with a very low rate of data loss. Of the total 18 variables analysed, all key features, including three-phase currents (Ia, Ib, Ic), voltages (Va, Vb, Vc), energy (Ea, Eb, Ec), as well as aggregate features such as total energy and energy asymmetry, had absolutely no missing values (0%). This indicates that the data representing the dynamics of the power system has been recorded consistently and comprehensively, which is a crucial prerequisite for machine learning-based modelling. According to data quality theory in machine learning, the presence of missing values in key features can degrade model performance and increase estimation bias, particularly in non-linear systems such as artificial neural networks [9], [10].

Conversely, missing values were found only in a few derived features, namely the Energy Moving Average (Energy MA), Energy Standard Deviation (Energy SD), thresholds (upper and lower), and Energy Fault Dynamic, with 49 observations each, or approximately 0.408% of the total data. This proportion is very small and statistically insignificant in relation to the overall distribution of the dataset. The consistent pattern of missing values in these features indicates that they are most likely caused by a rolling window computational mechanism, where the initial number of observations is insufficient to generate valid derived values. This phenomenon is common in time-series analysis and has been widely reported in studies of both statistical and system dynamics-based feature extraction [11], [12], [13].

Thus, there is no indication of either missing data of a random nature (MCAR) or systematic nature (MAR/MNAR) in the main features that could cause model bias. The limited data loss in these derived features can be effectively addressed using simple imputation techniques without compromising the integrity of the primary information. This is consistent with previous findings stating that missing values in small proportions (<1%) do not have a significant impact on model performance, particularly if they do not occur in the primary features [14], [15].

The results of the data split show that the dataset partitioning process has produced a balanced and representative distribution across the subsets. Of the total data, 8,401 samples were allocated as training data, whilst 1,800 samples each were used for validation and testing, reflecting a split ratio of approximately 70%–15%–15%. This strategy is a commonly used approach in machine learning to maintain a balance between the model training process and generalisation evaluation [16], [17], [18].

In terms of class distribution, across the training, validation and test subsets, it is evident that the proportions of the Normal and Fault classes are relatively balanced. In the training data, the distribution stands at approximately 54% for normal conditions and 46% for fault conditions, whilst a similar pattern is maintained in the validation and test data. This consistency indicates that the data splitting process successfully preserved the statistical characteristics of the original dataset, thereby minimising potential distribution bias across subsets. According to the literature, a balanced class distribution is crucial in binary classification to avoid model bias towards the majority class and to improve the reliability of evaluation metrics such as AUC and F1-score [19], [20].

To improve the reliability of model evaluation, this study employs a repeated cross-validation scheme with a 5-fold configuration and three iterations. This approach aims to reduce variability resulting from random data splitting and to provide a more stable and robust estimate of performance. Repeated cross-validation has been recommended in various studies as a more reliable evaluation method than single-split validation, particularly for highly complex datasets [21], [22].

The results of the ANN model training demonstrated very high performance across almost all tested hyperparameter combinations. The (AUC/ Receiver Operating Characteristic (ROC)) values ranged from 0.9975 to 0.9999, indicating the model's very strong discriminatory ability in distinguishing between normal and abnormal conditions. Theoretically, an AUC value approaching 1 indicates that the probability distributions of the predictions for the two classes show almost no overlap, meaning the model exhibits excellent separability in the feature space [23].

Furthermore, the sensitivity (recall for the disturbance class) consistently remained above 0.997, indicating that almost all instances of disturbance were successfully detected. Meanwhile, specificity also falls within a high range (0.991–0.998), indicating the model's ability to avoid misclassification under normal conditions. This balance between sensitivity and specificity is crucial in the context of power protection systems, where both failure to detect faults and false alarms can have a significant impact on system reliability [24].

The best model was selected based on the highest ROC score, resulting in an optimal configuration with a hidden layer size of 3 and a regularisation parameter (decay) of 0.01. Interestingly, this configuration, despite its relatively low complexity, was able to deliver very high performance. This indicates that the relationship between features and the target has a clear structure and can be effectively represented without requiring a complex network architecture. This finding is consistent with the principle of Occam's Razor in machine learning, which states that simple models are often better able to generalise effectively than overly complex models [25], [26].

The stability of performance across various hyperparameter combinations suggests that the model is relatively insensitive to minor changes in the network configuration. This indicates that the features employed particularly those based on phase-space and energy possess strong representational power regarding fault characteristics in three-phase systems. Previous studies have also shown that features based on system dynamics and energy are capable of enhancing class separability in fault detection problems compared to conventional time-domain-based features alone [27], [28].

TABLE I. ROBUSTNESS OF ANN PERFORMANCE ACREOSS HIDDEN SIZE AND WEIGHT DECAY CONFIGURATIONS

Hidden Size	Decay	ROC (AUC)	Sensitivity	Specificity
3	0.001	0.997551	0.997945	0.991710
3	0.01	**0.999986**	0.997724	**0.998359**
3	0.1	0.999447	0.997137	0.993696
5	0.001	0.999958	0.997724	0.997928
5	0.01	0.999985	0.997504	0.998273
5	0.1	0.999982	0.998091	0.996719
7	0.001	0.999985	0.997504	0.997755
7	0.01	0.999986	0.997504	0.998359
7	0.1	0.999982	0.998312	0.996546
9	0.001	0.999935	0.997724	0.998014
9	0.01	0.999958	0.997504	0.998273
9	0.1	0.999984	**0.998312**	0.996978

It can be seen from Table I that the ANN model exhibits very high and stable performance across all tested parameter combinations. Changes in the values of hidden layer size and decay do not result in significant differences in model performance, indicating that the patterns in the data can be effectively learned even by a relatively simple architecture. The optimal configuration was obtained for a model with low complexity, which nevertheless provides a good balance between fault detection capability and the identification of normal conditions.

Test results show that the proposed approach has very high discriminatory power in distinguishing between normal conditions and faults in a three-phase system. This is demonstrated by AUC values of 0.985772 on the validation data and 0.987438 on the test data. The consistency of the AUC values between the validation and test data indicates that the model does not suffer from significant overfitting and possesses good generalisation ability towards previously unseen data. This finding is consistent with previous research on machine learning-based power system fault detection, which indicates that performance stability across datasets is an important indicator of model robustness [29].

The determination of the decision threshold using the Youden method yielded an optimal threshold value of 0.1634, which provides the best balance between sensitivity and specificity. Theoretically, the Youden index is used to maximise the trade-off between the true positive rate and the false positive rate in both medical and engineering classification systems [30]. At this threshold, the model achieves a sensitivity of 0.9826 and a specificity of 0.9890, indicating that it can detect most fault events whilst maintaining a very high ability to recognise normal conditions. In the context of power systems, this balance is crucial because detection errors, particularly false negatives can have a significant impact on system safety [31], [32].

Analysis of the confusion matrix on the validation data shows that classification errors are relatively small, with 14 undetected cases of disorder (false negatives) and 11 normal cases incorrectly classified as disorders (false positives). This results in an accuracy of 0.9861 with a 95% confidence interval ranging from 0.9796 to 0.9910, indicating a high level of model reliability. A Kappa value of 0.9719 indicates a very strong level of agreement between the model's predictions and the actual conditions after accounting for the possibility of random agreement. This value is significantly higher than the general threshold for interpreting Kappa (>0.8), which is categorised as "almost perfect agreement".

A balanced accuracy value of 0.9858 indicates that the model performs equally well in detecting both classes, meaning it is not biased towards any class. This is particularly important in power system applications, where an imbalance in performance could lead to failure in detecting critical faults. Positive predictive values and negative predictive values approaching 0.986 also indicate that both fault predictions and normal condition predictions have a high level of confidence. These findings are consistent with previous research showing that a combination of evaluation metrics provides a more comprehensive picture than using accuracy alone [33].

Evaluation of the test data shows that the model maintains very high performance. According to the confusion matrix, 966 normal data points were correctly classified with only 2 errors, whilst 808 fault data points were correctly detected with 24 errors. This pattern indicates that the model has a very strong ability to minimise false positives, although there are still a small number of false negatives. In the context of power system protection, false negatives are often more critical than false positives because undetected faults can lead to more widespread system damage [34]. Therefore, although the model's performance is very high, these results highlight the need for further attention to be paid to improving sensitivity.

Overall, the accuracy on the test data reached 0.9856 with a 95% confidence interval of 0.9789 to 0.9905, indicating the stability of the model's performance. This value is significantly higher than the No Information Rate of 0.5378, and statistical tests indicate that this improvement in performance is significant (p-value < 2.2×10^{-16}). This suggests that the model does not merely follow the class distribution but is genuinely capable of learning relevant patterns within the data. A Kappa value of 0.9709 further reinforces that the model possesses very high reliability in classification.

However, the significant result of the McNemar test (p-value = 3.814×10^{-5}) indicates an imbalance in the types of errors produced by the model, with the number of false negatives being higher than that of false positives. A sensitivity of 0.9712 indicates that most abnormalities were successfully detected, but approximately 2.88% of abnormalities remained undetected. Conversely, a specificity of 0.9979 indicates that the model is almost perfect at recognising normal conditions. A positive predictive value of 0.9975 indicates that almost all fault predictions are correct, whilst a negative predictive value of 0.9758 indicates that most normal predictions are also accurate.

A balanced accuracy of 0.9845 confirms that the model maintains high and balanced performance. These results are consistent with previous studies showing that domain-based feature approaches (such as phase-space and energy) can enhance a model's discriminatory ability in power system analysis [35], [36].

In terms of data, the dataset used comprises 12,001 samples with 13 predictor variables that have been carefully selected to avoid data leakage. The class distribution is relatively balanced, with 6,505 normal samples and 5,496 fault samples, thus avoiding significant bias during model training. The label cardinality and label density values of 0.458 indicate that the frequency of fault label occurrence is at a moderate level, which aligns with the characteristics of real-world power systems that are not always in a faulted state.

The scenario-based data partitioning strategy into 121 separate scenarios ensures that no information leakage occurs between datasets. This approach is crucial for enhancing the model's external validity, as recommended in the machine learning literature for dynamic systems [37], [38]. Consistent class distribution across the training, validation and test datasets indicates that the data splitting process was carried out in a representative manner, thereby supporting a fair and unbiased model evaluation.

It can be seen in Figure 1 that the dataset used has a relatively balanced class distribution between normal and fault conditions, thus not introducing significant bias into the model training process. The missing value profile shows that only certain features have missing values, and in general the level of missing data remains manageable and can be addressed through pre-processing. Furthermore, scenario-based data splitting produces a consistent class distribution across the training.

It can be seen in Figure 2 that the distributions of several key features show quite distinct differences in pattern between the normal and fault classes. For several variables, particularly those relating to current, voltage and energy, the density curves for the two classes do not fully overlap but instead exhibit a shift and differences in distribution shape. This indicates validation, and test datasets, indicating that the splitting process was carried out representatively without distorting the distribution. Analysis of the scenario levels also reveals variations in the proportion of faults across scenarios, yet these remain naturally distributed without any indication of abnormality, thereby supporting the validity of the model evaluation and minimizing the risk of data leakage.

It can be seen in Figure 2 that the distributions of several key features show quite distinct differences in pattern between the normal and fault classes. For several variables, particularly those relating to current, voltage and energy, the density curves for the two classes do not fully overlap but instead exhibit a shift and differences in distribution shape. This indicates validation, and test datasets, indicating that the splitting process was carried out representatively without distorting the distribution. Analysis of the scenario levels also reveals variations in the proportion of faults across scenarios, yet these remain naturally distributed without any indication of abnormality, thereby supporting the validity of the model evaluation and minimizing the risk of data leakage.

It can be seen in Figure 3 that the phase-space representation shows a clear separation between normal and fault conditions, although there are still areas of overlap indicating the complexity of the signal characteristics. On the energy feature map, the relationship between total energy and asymmetry energy reveals distinct distribution patterns across classes, with fault conditions tending to exhibit higher variation and dispersion. Meanwhile, the pre-processing diagnostic results indicate that the standardization process successfully stabilized the data distribution without obscuring the primary patterns, thereby enhancing the features' readiness for the model training process.

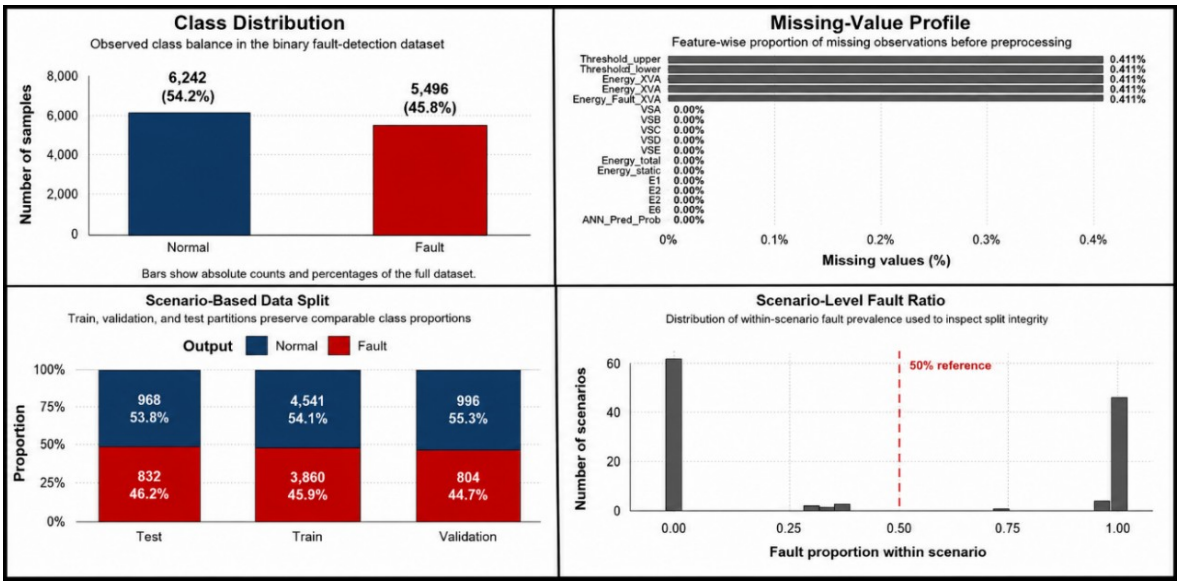

Figure 1. Class Distribution, Missing Values, and Consistency of Scenario-Based Data Partitioning

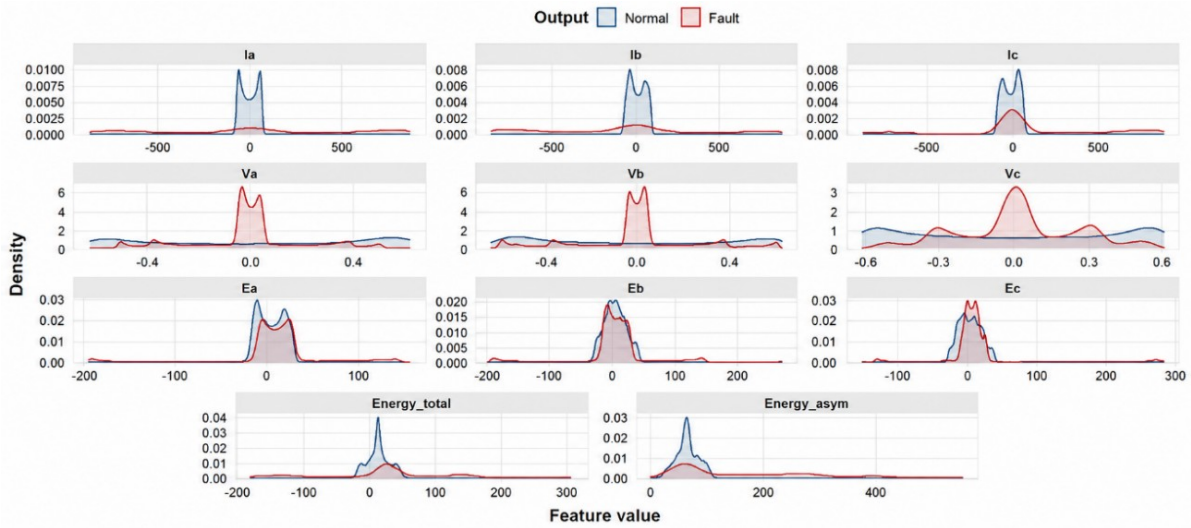

Figure 2. Distribution of Key Features Between Normal and Fault Conditions and Their Implications for the Model's Discriminatory Ability

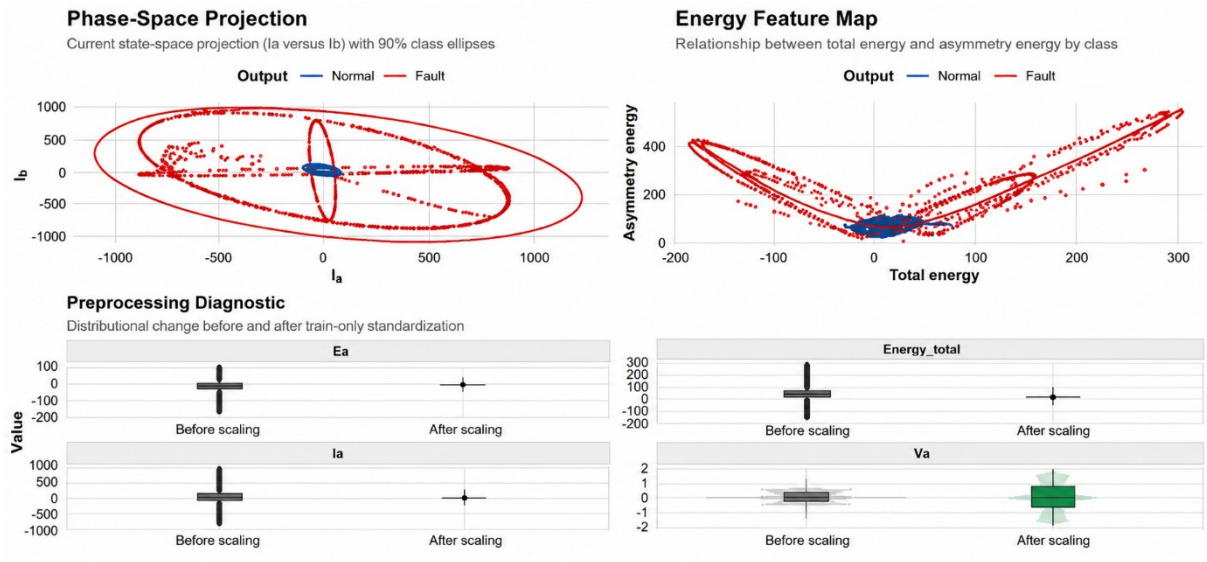

Figure 3. Phase-Space Separation, Energy Feature Mapping, and Preprocessing Effects in Fault Classification

TABLE II. COMPERATIVE PERFORMANCE OF MACHINE LEARNING MODELS FOR DISTURBANCE DETECTION IN THREE-PHASE POWER SYSTEM

Model	Accuracy	Kappa	Sensitivity	Specificity	Precision	Recall	F1-Score	Balanced Accuracy	AUC	Training Time (s)
GLM	0.814	0.616	0.643	0.96	0.931	0.643	0.76	0.802	0.834	0.18
SVM	0.994	0.987	0.992	0.995	0.994	0.992	0.993	0.994	≈ 0.9998	2.94
RF	**0.997**	**0.994**	**0.998**	**0.997**	**0.996**	**0.998**	**0.997**	**0.998**	**≈ 0.99997**	1.21
ANN	0.996	0.992	0.996	0.995	0.994	0.996	0.996	0.996	≈ 0.999	0.42

B. Baseline Models and Fair Comparison Setup

The results of the analysis show that the dataset used in this study comprises 11,952 samples with 16 features representing the characteristics of a three-phase power system, including current, voltage, and energy- and imbalance-based features. This representation of features is in line with modern power system analysis approaches, where a combination of basic electrical parameters and derived features such as energy and imbalance indices has proven capable of capturing disturbance dynamics more comprehensively than using raw signals alone [39].

The relatively balanced class distribution, with 6,456 samples in the normal class (≈54.02%) and 5,496 samples in the disorder class (≈45.98%) indicates that the dataset does not suffer from significant class imbalance. This is important because, according to [40] and [41], extreme class imbalance can lead to model bias towards the majority class and reduce the ability to detect minority events, particularly in fault detection. With a relatively balanced distribution, the evaluation of model performance in this study can be considered more representative without the need for additional balancing techniques such as oversampling or cost-sensitive learning.

The relatively balanced class distribution, with 6,456 samples in the normal class (≈54.02%) and 5,496 samples in the disorder class (≈45.98%) indicates that the dataset does not suffer from significant class imbalance. This is important because, according to [40] and [41], extreme class imbalance can lead to model bias towards the majority class and reduce the ability to detect minority events, particularly in fault detection. With a relatively balanced distribution, the evaluation of model performance in this study can be considered more representative without the need for additional balancing techniques such as oversampling or cost-sensitive learning.

The relatively balanced class distribution, with 6,456 samples in the normal class (≈54.02%) and 5,496 samples in the disorder class (≈45.98%) indicates that the dataset does not suffer from significant class imbalance. This is important because, according to [40] and [41], extreme class imbalance can lead to model bias towards the majority class and reduce the ability to detect minority events, particularly in fault detection. With a relatively balanced distribution, the evaluation of model performance in this study can be considered more representative without the need for additional balancing techniques such as oversampling or cost-sensitive learning.

Preprocessing is carried out using an approach that prevents data leakage, whereby the normalization parameters are learned solely from the training data and then applied to the test data. This is a crucial aspect in maintaining the integrity of the evaluation, as the use of information from the test data during the training phase can lead to an overly optimistic estimate of performance [46]. Furthermore, all models used the same decision threshold, namely $\tau = 0.5$, ensuring that classification decisions were made consistently across all methods and did not give an unfair advantage to any particular model.

The evaluation results revealed significant differences in performance between models when tested using a consistent protocol. The RF model emerged as the method with the best overall performance, with an accuracy of 0.997 and a Kappa coefficient of 0.994, indicating a near-perfect level of predictive agreement. The balanced sensitivity (0.998) and specificity (0.997) values indicate that this model can detect both disturbed and non-disturbed conditions with high accuracy without bias towards either class. An F1-score of 0.997 also indicates an optimal balance between precision and recall. These advantages are consistent with the literature, which states that tree-based ensemble methods such as Random Forest have a high ability to capture non-linear relationships and interactions between features, as well as being robust to noise [47], [48]. An AUC value approaching 1 further reinforces that the model possesses excellent discriminatory ability in separating the two classes in the probability space.

The SVM model demonstrated highly competitive performance, with an accuracy of 0.994 and an F1-score of 0.993. The high sensitivity (0.992) and specificity (0.995) values indicate that the SVM possesses strong generalization capabilities. This can be explained by the maximum margin principle employed by the SVM, which theoretically maximizes the distance between classes and enhances robustness against data that cannot be linearly separated through the use of a kernel [49]. In the context of power systems, this approach is effective in capturing the complex patterns generated by disturbance dynamics, as also reported in previous SVM-based fault detection studies.

Conversely, the Logistic Regression (GLM) model demonstrated relatively lower performance, with an accuracy of 0.814 and an F1-score of 0.760. Although it possesses high specificity (0.960), its sensitivity is significantly lower (0.643), indicating a tendency for the model to fail to detect a number of fault events (false negatives). This reflects the limitations of linear models in capturing complex non-linear relationships within power system data, particularly when features involve phase-space and energy transformations. These findings are consistent with the theory that Logistic Regression performs optimally on data with linear separability, whilst its performance declines on data with complex feature interactions [51].

In terms of computational efficiency, Logistic Regression is the fastest model to train, making it suitable for applications with limited resources or requiring real-time responses. However, the increase in computational time for models such as Random Forest and SVM has proven to be commensurate with a significant improvement in performance. This demonstrates the classic trade-off between model complexity and accuracy, whereby more complex models tend to deliver higher performance at the cost of greater computational overhead [52].

As can be seen in Table II, non-linear models consistently outperform linear models. Random Forest demonstrated the best overall performance with the most stable results across various evaluation metrics, followed by SVM and ANN, which also exhibited very high discriminatory power. Meanwhile, Logistic Regression shows the poorest performance, indicating its limitations in capturing the complexity of data patterns. From a computational perspective, there is a clear trade-off between speed and accuracy, whereby simpler models are faster but less accurate than more complex models.

The ROC curve in Figure 4 shows that Random Forest has the best discriminatory ability, with the curve lying closest to the top-left corner, followed very closely by SVM and then ANN, which also demonstrates high performance but is slightly less stable. In contrast, Logistic Regression has a curve that lies further from the optimal region, indicating a lower ability to distinguish between classes.

Figure 5 shows that the high-performance model produces a very clear probabilistic separation between the disturbance and non-disturbance classes, with distributions concentrated well away from the decision boundary. In contrast, the lower-performance model exhibits an overlap of distributions around the threshold, indicating uncertainty in the classification.

Figure 6 is a confusion matrix showing that high-performance models such as Random Forest, SVM and ANN are dominated by values on the main diagonal, indicating that the majority of predictions are correct and errors are minimal. In contrast, Logistic Regression still exhibits more noticeable errors off the diagonal, particularly in detecting the ‘disorder’ class, thereby reflecting the limitations of linear models compared to non-linear approaches.

It can be seen in Figure 7 that the model with the best performance has a calibration curve that most closely approximates the diagonal line, indicating that the predicted probabilities align with actual events. Meanwhile, deviations from this line indicate a mismatch between predictions and reality, with narrow confidence intervals signalling greater model stability.

As can be seen in Figure 9, the cumulative contribution of features increases sharply in the early stages, indicating that a small subset of features plays a dominant role in determining the model’s decisions. After a certain point, the curve begins to flatten, indicating that the addition of further features contributes only marginally to improving the model’s performance. This pattern confirms that the Random Forest model relies primarily on a small number of key features, whilst the remaining features serve as complementary elements with a more limited influence.

C. Robustness and Reliability Analysis

The results of the analysis show that the developed model performs exceptionally well and remains consistently stable across various test scenarios. An evaluation of variability conducted through 30 experimental repetitions shows that the average AUC value reaches 0.9954 with a very small standard deviation (±0.00043) and a narrow 95% confidence interval (0.995266–0.995587). This stability indicates that the model possesses very strong and consistent discriminatory ability across experiments. In the context of machine learning, low variability in repeated experiments is an important indicator of a model’s reliability and reproducibility, as emphasised in a study by [53] and [54], which suggests that models with high performance, but low variance are more reliable than unstable models.

Consistency in performance is also reflected in other evaluation metrics, where the accuracy (0.9948), precision (0.9957), recall (0.9947), and F1-score (0.9952) show very little variation across trials. This indicates that the model not only excels in a specific metric but also exhibits a good overall balance of performance. Theoretically, this cross-metric stability suggests that the model is not biased towards any particular class and possesses good internal generalisation ability regarding the training data distribution [55].

In terms of robustness, the test results show that the model’s performance is relatively unaffected by variations in the operational conditions tested. For variations in load and fault resistance, the AUC values remained within an almost identical range, with standard deviations and sensitivities approaching zero. These findings indicate that the features used primarily phase-space and energy-based features exhibit invariance to changes in system parameters. This is consistent with previous research showing that phase-space representations are capable of capturing system dynamics in a more fundamental manner

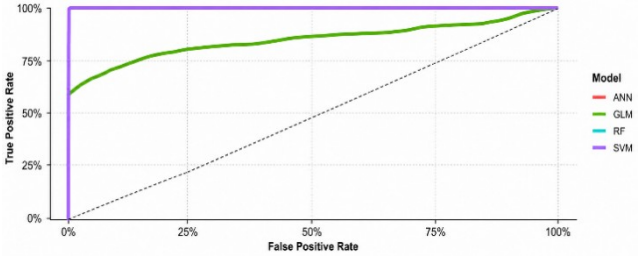

Figure 4. ROC Curves Comparing the Performance of Classification Models (Random Forest, SVM, ANN, and Logic Regression)

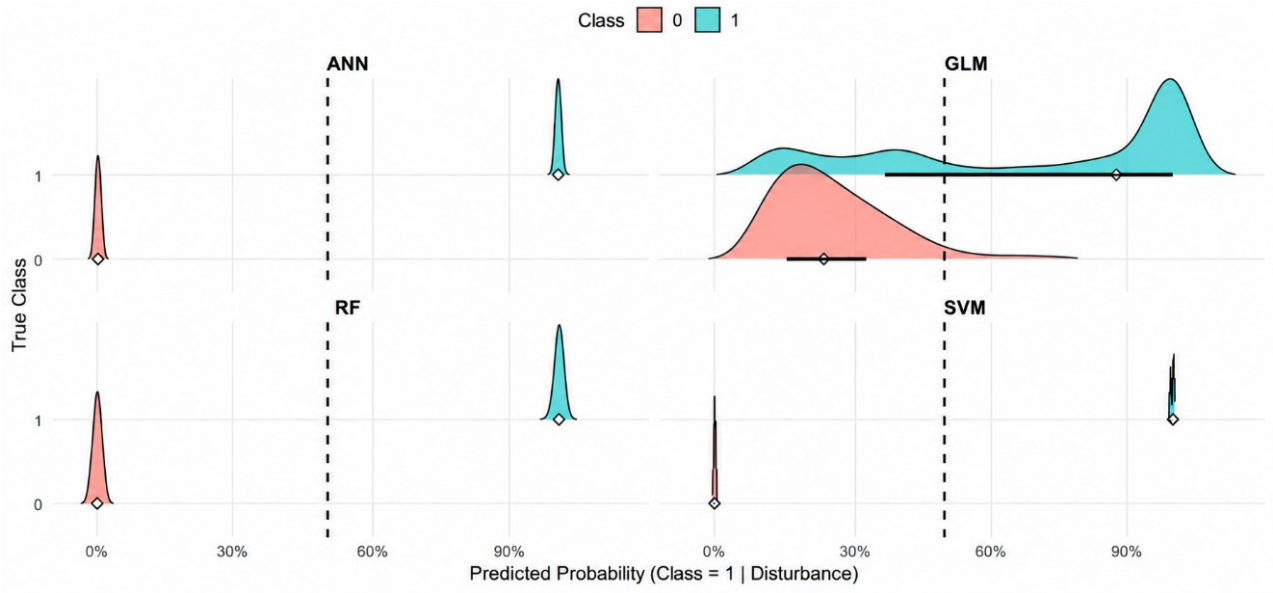

Figure 5. Comparison of Classification Probability Distributions between High- and Low-Performance Models

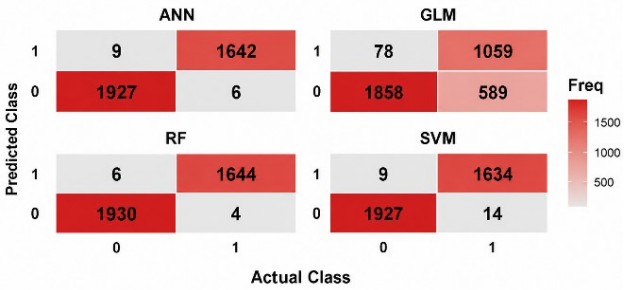

Figure 6. Confusion Matrix Comparison of Machine Learning Models for Disturbance Classification

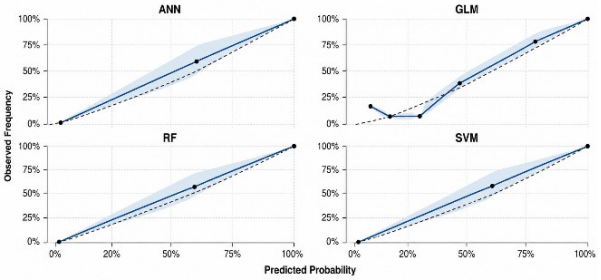

Figure 7. Calibration Curves of Classification Models with Confidence Intervals

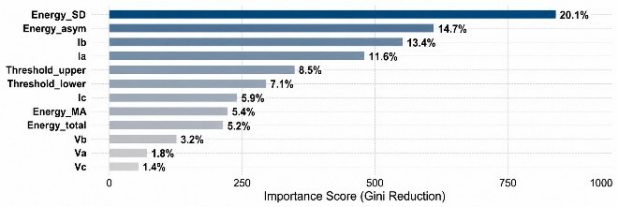

Figure 8. Feature Importance Distribution in Disturbance Classification Model

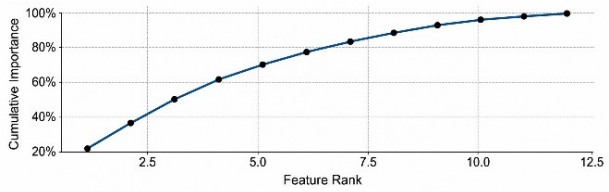

Figure 9. Cumulative Feature Contribution Curve in Random Forest Model

and are less sensitive to changes in external parameters [56], [57].

In the noise injection scenario, although there was some variation in performance, the changes observed remained very small (standard deviation ±0.000186 and sensitivity 0.000422). This indicates that the model possesses good resistance to external disturbances. In the literature, resistance to noise is one of the key indicators of a model's robustness, particularly in power system applications which inherently involve signal fluctuations and disturbances [58]. The model's ability to maintain performance under noisy conditions indicates that the features used have a good signal-to-noise ratio and are capable of effectively extracting important information.

Generalisation testing via out-of-distribution (OOD) scenarios revealed a decline in performance from an AUC of 0.9989 on the training data to 0.9897 on the test data with a different distribution, with a generalisation gap of approximately 0.0092. Although there is a decline, this value remains relatively small, indicating that the model is still capable of maintaining high performance under conditions that are not entirely identical to the training data. Theoretically, this phenomenon is consistent with the principle of generalisation in machine learning, whereby a good model should be able to maintain performance on slightly different data distributions, even if it is not entirely free from a decline in performance [59], [60]. These results also indicate that the model has not yet experienced significant overfitting, although there is still a tendency for better performance on data similar to the training data.

From a statistical significance perspective, the test results show that the difference in performance between the proposed model and the baseline is highly significant, with a very small p-value ($p < 0.001$). Furthermore, the very large effect size (approximately 60.7) indicates that the performance improvement achieved is not only statistically significant but also has a very strong practical impact. In the evaluation of machine learning models, the combination of a small p-value and a large effect size is an indicator that the performance difference is substantive and not merely an artefact of sample size [61]. Therefore, these results provide strong evidence that the proposed approach consistently outperforms the baseline method.

It can be seen in Figure 10 that the model demonstrates very high and consistent performance, as indicated by the extremely narrow distribution of AUC values around the mean, which is close to 1. This stability is reinforced by a distribution of values that shows almost no variation or outliers, thus indicating strong reliability across repetitions. Furthermore, the robustness curve indicates that the model's performance is

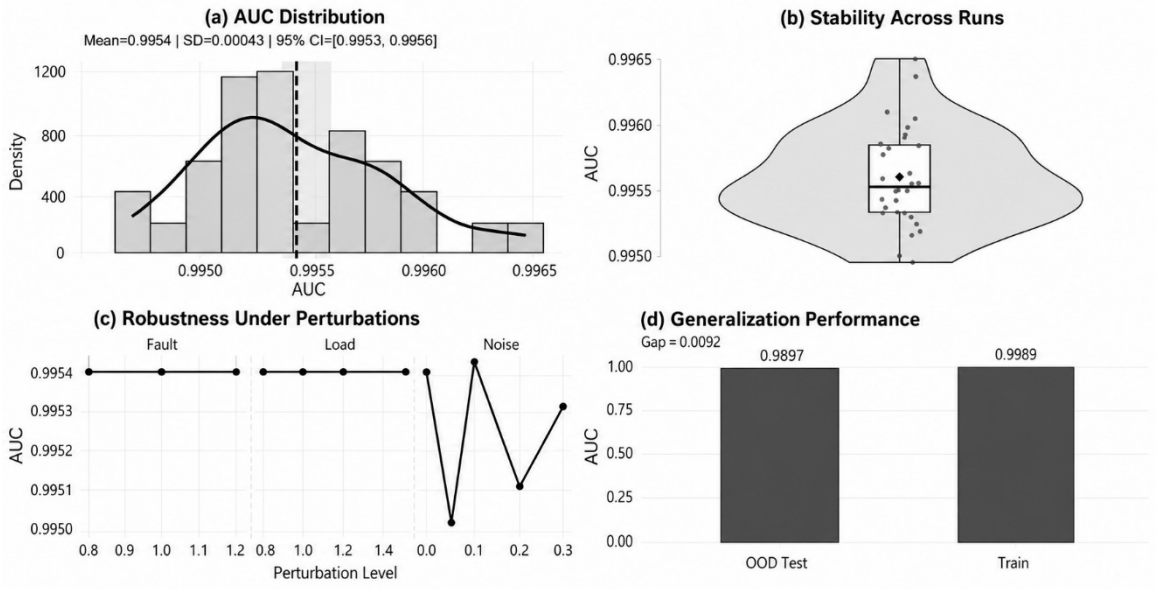

Figure 10. Model Performance Stability, Robustness, and Generalization Across Multiple Experimental Conditions

relatively unaffected by variations in conditions such as noise, load changes, and faults, with only very minor changes in the AUC. Regarding generalisation, although there is a slight decline in performance on out-of-distribution data, the resulting gap remains small, thus demonstrating good generalisation ability.

D. Ablation Analysis

The results of the ablation study provide a deeper understanding of the relative contribution of each feature group to the model's performance in detecting and classifying faults in three-phase power systems. In general, the full model, which combines all features, including raw signals, phase-space features, and engineered features demonstrates very high performance with an AUC of 0.9944 and an accuracy of 98.58%. This indicates that the integration of multi-representation data is capable of enhancing the model's discriminatory ability, as also reported in the literature that the combination of time-domain features and non-linear transformations can enrich the representative information within power systems [62], [63].

However, more interesting results emerged when the phase-space features were removed. In this configuration, the model's performance improved, with an AUC of 0.9992 and an accuracy of 98.77%. These findings suggest that the phase-space feature in the current implementation does not yet make a significant contribution to improving the model's performance. Theoretically, phase-space representations are designed to capture non-linear dynamics and the characteristics of dynamic systems [57], [64]. However, in this case, it is highly likely that this information is already implicitly represented in the raw signal or other features, resulting in information redundancy. This phenomenon is consistent with the findings [65] which states that redundant features not only fail to improve performance, but can also increase the model's complexity without providing any additional benefit.

Furthermore, when the engineered features were removed, the model continued to perform well, achieving an AUC of 0.9962 and an accuracy of 98.63%. This indicates that engineered features, such as signal energy and derived statistics, do contribute to performance, but are not the dominant factors. In the literature, feature engineering is often used to improve signal representation and highlight specific characteristics [66], However, in some cases, modern machine learning models, particularly data-driven ones are capable of directly extracting key patterns from raw data without the need for further transformation [67].

The most significant results were observed in the 'raw-only' scenario, where the model achieved its best performance with an AUC of 0.9995 and an accuracy of 98.99%. These findings indicate that the key information required for fault classification is already strongly embedded in the raw current and voltage signals. Theoretically, electrical signals contain fundamental information regarding the system's condition, including fault characteristics such as changes in amplitude, frequency, and phase [68]. Raw signal-based approaches often deliver competitive or even superior performance compared to feature-based approaches, particularly when the model has sufficient capacity to learn complex representations.

Conversely, in scenarios using only engineered and phase-space features (feature-based only), the model's performance declined significantly, with an AUC of 0.8740 and an accuracy of 86.08%. This decline indicates that derived features alone are unable to optimally represent the complexity of disturbance patterns without the support of information from raw signals. This can be explained by the theory of information loss in the feature extraction process, where data transformation often results in the loss of some important information contained in the original signal [69]. These findings are also consistent with previous research showing that the use of overly compressed or restricted features can reduce a model's ability to capture complex variations in power systems [70].

IV. CONCLUSION

Research shows that machine learning approaches based on Artificial Neural Networks (ANNs) are highly effective at distinguishing between normal conditions and faults in three-phase systems, with very high accuracy and AUC values approaching 1. Electrical signal-based features such as current, voltage, and energy have proven capable of effectively representing system conditions, while derived features can enrich the representation although they are not always more dominant than raw signals. Non-linear models such as ANN, Random Forest, and SVM have also proven superior to linear models in capturing the complexity of power systems. Furthermore, the use of scenario-based splitting has proven to be more valid for dynamic systems because it prevents

temporal data leakage and provides a more realistic performance evaluation. In practice, the developed model is capable of detecting faults with very high sensitivity and specificity while remaining robust against noise, load variations, and changes in operating conditions. However, false negatives still occur and need to be minimized, as they have the potential to pose significant risks to the power system. Therefore, future research is recommended to focus on improving the model's sensitivity, developing more complex deep learning methods, implementing explainable AI, and conducting tests using real-time data to further enhance the model's performance and reliability in real-world implementations.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Majid Jamil, Sanjeev Kumar Sharma, and Rajveer Singh for their work published in SpringerPlus (2015), which provided the primary dataset on electrical fault detection and classification used in this study. The dataset, containing line current and voltage measurements for various fault conditions in three-phase transmission systems, was critical for the development, training, and validation of the proposed machine learning models. Their contribution to the field of power system fault analysis has enabled further research into accurate and reliable disturbance detection and classification.

REFERENCES

[1] T. Gönen, C.-W. Ten, and A. Mehrizi-Sani, *Electric power distribution engineering*. CRC press, 2024.

[2] S. Sharma, R. Chaudhary, and P. Mani, *Fundamentals of electric power system*. Shineeks Publishers, 2024.

[3] M. A. Sheikh, S. T. Bakhsh, M. Irfan, N. bin M. Nor, and G. Nowakowski, "A review to diagnose faults related to three-phase industrial induction motors," *Journal of Failure Analysis and prevention*, vol. 22, no. 4, pp. 1546–1557, 2022.

[4] M. Vileiniskis, R. Remenyte-Prescott, D. Rama, and J. Andrews, "Fault detection and diagnostics of a three-phase separator," *Journal of Loss Prevention in the Process Industries*, vol. 41, pp. 215–230, 2016.

[5] O. Prakash, A. K. Samantaray, and R. Bhattacharyya, "Optimal adaptive threshold and mode fault detection for model-based fault diagnosis of hybrid dynamical systems," in *Fault Diagnosis of Hybrid Dynamic and Complex Systems*, Springer, 2018, pp. 45–78.

[6] H. S. Khalsa, "Generalised Power Components Definitions for Single and Three-phase Electrical Power System Under Non-sinusoidal and Nonlinear Conditions," 2007.

[7] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander, and M. S. H. Sunny, "Application of big data and machine learning in smart grid, and associated security concerns: A review," *Ieee Access*, vol. 7, pp. 13960–13988, 2019.

[8] M. Jamil, S. K. Sharma, and R. Singh, "Fault detection and classification in electrical power transmission system using artificial neural network," *SpringerPlus*, vol. 4, no. 1, p. 334, 2015.

[9] C. A. Leke and T. Marwala, *Deep learning and missing data in engineering systems*. Springer, 2019.

[10] A. F. Ahmad, K. Alshammari, I. Ahmed, and M. Sayed, "Machine learning for missing value imputation," *arXiv preprint arXiv:2410.08308*, 2024.

[11] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, p. 6085, 2018.

[12] F. K. Mirza, Ö. Pekcan, M. Hekimoğlu, and T. Baykaş, "Stock price forecasting through symbolic dynamics and state transition graphs with a convolutional recurrent neural network architecture," *Neural Computing and Applications*, vol. 37, no. 20, pp. 15855–15890, 2025.

[13] P. Chattopadhyay, "Data-Driven Modeling and Pattern Recognition of Dynamical Systems," 2018.

[14] B. Radišić, S. Seljan, and I. Dunđer, "Impact of missing values on the performance of machine learning algorithms," presented at the 5th International Conference on Recent Trends and Applications in Computer Science and Information Technology (RTA-CSIT), University of Tirana, Faculty of Natural Sciences, Department of Informatics, 2023, pp. 54–62.

[15] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *Journal of clinical epidemiology*, vol. 110, pp. 63–73, 2019.

[16] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, and R. Forghani, "Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls," *Radiology: Artificial Intelligence*, vol. 5, no. 1, p. e220028, 2022.

[17] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," *PeerJ Computer Science*, vol. 10, p. e2245, 2024.

[18] J. Tan, J. Yang, S. Wu, G. Chen, and J. Zhao, "A critical look at the current train/test split in machine learning," arXiv preprint *arXiv:2106.04525*, 2021.

[19] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digital Health*, vol. 2, no. 11, p. e0000290, 2023.

[20] R. Diallo, C. Edalo, and O. O. Awe, "Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score," in *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*, Springer, 2024, pp. 283–312.

[21] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, "Cross validation for model selection: a review with examples from ecology," *Ecological monographs*, vol. 93, no. 1, p. e1557, 2023.

[22] V. W. Lumumba, D. Kiprotich, M. Lemasulani Mpaine, N. Grace Makena, and M. Daniel Kavita, "Comparative analysis of cross-validation techniques: LOOCV, K-folds cross-validation, and repeated K-folds cross-validation in machine learning models," K-*folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models (June 01, 2024)*, 2024.

[23] Z. Yang, Q. Xu, S. Bao, X. Cao, and Q. Huang, "Learning with multiclass AUC: Theory and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7747–7763, 2021.

[24] H. Belgacem and I. Chihi, "Toward Reliable and Intelligent Sensor Systems: A Comprehensive Study of Fault Diagnosis and Mitigation," *IEEE Sensors Reviews*, 2025.

[25] F. J. Bargagli Stoffi, G. Cevolani, and G. Gnecco, "Simple models in complex worlds: occam's razor and statistical learning theory," *Minds and Machines*, vol. 32, no. 1, pp. 13–42, 2022.

[26] T. F. Sterkenburg, "Statistical Learning Theory and Occam's Razor: The Core Argument: TF Sterkenburg," *Minds and Machines*, vol. 35, no. 1, p. 3, 2024.

[27] L. Zhang *et al.*, "A data-driven approach to anomaly detection and vulnerability dynamic analysis for large-scale integrated energy systems," *Energy conversion and management*, vol. 234, p. 113926, 2021.

[28] J. Chen, S. Zhou, Y. Qiu, and B. Xu, "An anomaly detection method of time series data for cyber-physical integrated energy system based on time-frequency feature prediction," *Energies*, vol. 15, no. 15, p. 5565, 2022.

[29] O. A. Alimi, K. Ouahada, and A. M. Abu-Mahfouz, "A review of machine learning approaches to power system security and stability," *IEEE access*, vol. 8, pp. 113512–113531, 2020.

[30] T. Taner and J. Antony, "The assessment of quality in medical diagnostic tests: a comparison of ROC/Youden and Taguchi methods," *International journal of health care quality assurance*, vol. 13, no. 7, pp. 300–307, 2000.

[31] S. C. Savulescu, *Real-time stability in power systems: techniques for early detection of the risk of blackout*. Springer, 2014.

[32] S. A. Aleem, N. Shahid, and I. H. Naqvi, "Methodologies in power systems fault detection and diagnosis," *Energy Systems*, vol. 6, no. 1, pp. 85–108, 2015.

[33] A. Reinke *et al.*, "Common limitations of image processing metrics: A picture story," *arXiv preprint arXiv:2104.05642*, 2021.

[34] J. Hare, X. Shi, S. Gupta, and A. Bazzi, "Fault diagnostics in smart micro-grids: A survey," *Renewable and sustainable energy reviews*, vol. 60, pp. 1114–1124, 2016.

[35] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Speech recognition using reconstructed phase space features," presented at the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)., IEEE, 2003, p. I–60.

[36] A. C. Lindgren, "Speech recognition using features extracted from phase space reconstructions," 2003.

[37] L. Chen, "Design and applications of scenario-based machine learning models," 2025.

[38] A. Hossain, "A Taxonomy and Comprehensive Survey of Scenario Generation for Autonomous Driving: Methods, Challenges, and Emerging Trends in Safety-Critical Testing," 2025.

[39] Y. Liu, D. Yuan, H. Fan, T. Jin, and M. A. Mohamed, "A multidimensional feature-driven ensemble model for accurate classification of complex power quality disturbance," *IEEE transactions on instrumentation and measurement*, vol. 72, pp. 1–13, 2023.
[40] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, p. 42, 2018.
[41] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of big data*, vol. 6, no. 1, p. 27, 2019.
[42] P. Szymański and T. Kajdanowicz, "A network perspective on stratification of multi-label data," presented at the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2017, pp. 22–35.
[43] M. Merrillees and L. Du, "Stratified sampling for extreme multi-label data," presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2021, pp. 334–345.
[44] A. Beutel *et al.*, "Putting fairness principles into practice: Challenges, metrics, and improvements," presented at the Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 453–459.
[45] M.-Y. Chan and S.-M. Wong, "A comparative analysis to evaluate bias and fairness across large language models with benchmarks," *OSF,, doi: https://doi. org/10.31219/osf. io/mc762*, 2024.
[46] G. Vandewiele *et al.*, "Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling," *Artificial Intelligence in Medicine*, vol. 111, p. 101987, 2021.
[47] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression," *Energy*, vol. 164, pp. 465–474, 2018.
[48] K. Taha, "Tree-based ensemble learning models for protein-protein interactions detection: a review and experimental evaluation," *BioData Mining*, 2025.
[49] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. Swamy, "Exploring kernel machines and support vector machines: Principles, techniques, and future directions," *Mathematics*, vol. 12, no. 24, p. 3935, 2024.
[50] K. R. Ahmed, M. E. Ansari, M. N. Ahsan, A. Rohan, M. B. Uddin, and M. A. H. Rivin, "Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP: KR Ahmed et al.," *Scientific Reports*, vol. 15, no. 1, p. 26355, 2025.
[51] J. J. Levy and A. J. O'Malley, "Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning," *BMC medical research methodology*, vol. 20, no. 1, p. 171, 2020.
[52] D. A. Gzar, A. M. Mahmood, and M. K. Abbas, "A Comparative Study of Regression Machine Learning Algorithms: Tradeoff Between Accuracy and Computational Complexity.," *Mathematical Modelling of Engineering Problems*, vol. 9, no. 5, 2022.
[53] S. S. Alahmari, D. B. Goldgof, P. R. Mouton, and L. O. Hall, "Challenges for the repeatability of deep learning models," *IEEE Access*, vol. 8, pp. 211860–211868, 2020.
[54] A. Bustillo, R. Reis, A. R. Machado, and D. Y. Pimenov, "Improving the accuracy of machine-learning models with data from machine test repetitions," *Journal of Intelligent Manufacturing*, vol. 33, no. 1, pp. 203–221, 2022.
[55] A. Palikhe *et al.*, "Towards transparent ai: A survey on explainable language models," *arXiv preprint arXiv:2509.21631*, 2025.
[56] F. Zhao, "Extracting and representing qualitative behaviors of complex systems in phase space," *Artificial Intelligence*, vol. 69, no. 1–2, pp. 51–92, 1994.
[57] M. Zhang, J. Zhang, A. Hou, A. Xia, and W. Tuo, "Dynamic system modeling of a hybrid neural network with phase space reconstruction and a stability identification strategy," *Machines*, vol. 10, no. 2, p. 122, 2022.
[58] L. V. Gambuzza, A. Buscarino, L. Fortuna, M. Porfiri, and M. Frasca, "Analysis of dynamical robustness to noise in power grids," *IEEE Journal on Emerging and Selected topics in Circuits and Systems*, vol. 7, no. 3, pp. 413–421, 2017.
[59] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[60] T. Freiesleben and T. Grote, "Beyond generalization: a theory of robustness in machine learning," *Synthese*, vol. 202, no. 4, p. 109, 2023.
[61] D. Rajput, W.-J. Wang, and C.-C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC bioinformatics*, vol. 24, no. 1, p. 48, 2023.
[62] W. Cui, W. Yang, and B. Zhang, "A frequency domain approach to predict power system transients," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 465–477, 2023.
[63] M. D'Apuzzo and M. D'Arco, "A time-domain approach for the analysis of nonstationary signals in power systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 9, pp. 1969–1977, 2008.
[64] N. Bartolovic, M. Gross, and T. Günther, "Phase space projection of dynamical systems," presented at the Computer Graphics Forum, Wiley Online Library, 2020, pp. 253–264.
[65] J.-A. Chen, W. Niu, B. Ren, Y. Wang, and X. Shen, "Survey: Exploiting data redundancy for optimization of deep learning," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–38, 2023.
[66] A. M. Alqudah and Z. Moussavi, "Bridging Signal Intelligence and Clinical Insight: A Comprehensive Review of Feature Engineering, Model Interpretability, and Machine Learning in Biomedical Signal Analysis," *Applied Sciences*, vol. 15, no. 22, p. 12036, 2025.
[67] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A survey on active deep learning: From model driven to data driven," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–34, 2022.
[68] M. Mishra, "Power quality disturbance detection and classification using signal processing and soft computing techniques: A comprehensive review," *International transactions on electrical energy systems*, vol. 29, no. 8, p. e12008, 2019.
[69] B. C. Geiger and G. Kubin, *Information loss in deterministic signal processing systems*. Springer, 2018.
[70] S. Wang and H. Chen, "A novel deep learning method for the classification of power quality disturbances using deep convolutional neural network," *Applied energy*, vol. 235, pp. 1126–1140, 2019.