Scientific Reports

https://doi.org/10.1038/s41598-026-35449-x

Article in Press

Detection of disturbances and cyber-attacks in smart grids using explainable machine learning

Received: 29 September 2025

Accepted: 6 January 2026

Published online: 19 February 2026

Cite this article as: Farsi M., Alwateer M., Alsaedi S.A. *et al.* Detection of disturbances and cyber-attacks in smart grids using explainable machine learning. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-35449-x

Mohamed Farsi, Majed Alwateer, Shatha Abed Alsaedi, Abdulrhman I. AlSahafi, Hossam Magdy Balaha, Moustafa M. Aboelnaga, Mahmoud Badawy & Mostafa A. Elhosseini

Detection of Disturbances and Cyber-Attacks in Smart Grids Using Explainable Machine Learning

Mohamed Farsi[1], Majed Alwateer[2], Shatha Abed Alsaedi[2], Abdulrhman I AlSahafi[3], Hossam Magdy Balaha[4,5], Moustafa M. Aboelnaga[5,6], Mahmoud Badawy[3,5,*], and Mostafa A. Elhosseini[1,5]

[1]Department of Information Systems, College of Computer Science and Engineering, Taibah University, Yanbu 46421, Saudi Arabia.
[2]Department of Computer Science, College of Computer Science and Engineering, Taibah University, Yanbu 46421, Saudi Arabia.
[3]Department of Computer Science and Information, Applied College, Taibah University, Madinah, 41461 Saudi Arabia.
[4]Bioengineering Department, J.B. Speed School of Engineering, University of Louisville, Louisville, KY 40292, USA.
[5]Department of Computers and Control Systems Engineering , Faculty of Engineering, Mansoura University, 35516, Mansoura, Egypt.
[6]Department of Software Engineering. SolarWinds Company Holandská 873, 639 00 Brno-střed,Czech repbulic.
[*]Corresponding author: Mahmoud Badawy: engbadawy@mans.edu.eg

ABSTRACT

Modern power systems are subjected to natural disruptions and cyberattacks, both of which have the potential to have catastrophic consequences on the grid's stability and security. Besides, due to the sophistication of cyber-physical threats, including techniques like false data injection and command tampering, comprehensive detection strategies to counter the vulnerabilities have become an absolute necessity. Traditional detection methods are inherently constrained in their capabilities since they treat physical failures and cyber intrusions as independent problems and use unclear models that hardly suffice for the enormous trustworthiness required in making high-stakes decisions. This study presents a heterogeneous data-driven framework that seeks to unify disturbance and intrusion detection using time-synchronized measurements. This framework utilizes advanced pre-processing techniques, multi-strategy feature selection approaches, and ensemble machine learning model implementations, all of which were optimized using Optuna. The framework employed permutation SHAP to enhance explainability and transparency by delivering interpretable insights regarding feature contributions. The experiments performed across 37 different event scenarios in binary, three-class, and multi-class settings prove the superior performance of the proposed framework. The best models showed precision, recall, F1-score, accuracy, and specificity exceeding 96%. Besides, the average performance across the aggregated datasets surpassed 93%. These results prove the effectiveness and the practicality of the framework toward the awareness and resilience of the smart grid, serving as an interpretable and scalable approach to countering ever-evolving cyber-physical threats.

Introduction

Modern power systems are prone to various forms of physical and cyber disturbances that affect power systems' stability, reliability, and security [1, 2]. The amalgamation of advanced sensing, communication, and control technologies with the need for real-time situational awareness and automation has also expanded the attack surface for malicious actors seeking to compromise critical infrastructure [3, 4, 5]. Adversaries manipulate sensor data, inject false commands, and exploit system vulnerabilities to the maximum extent. Their actions mimic natural human-generated faults or operational anomalies [6, 7]. Convergence of both these cyber and physical threats requires the development of appropriate technologies that can identify an easily visible disturbance from the tricks within the shortest time [8, 9, 10].

Grid cybersecurity practices today are concerned with the ability to detect heterogeneous, time-synchronized data streams as events with network characteristics [11, 12]. Most traditional fault detection and classification perform relatively well under stable assumptions with static feature extraction; there are only rare exceptions in enforcing nonstationary or adversarial behaviors as initiated by measurement data [13, 14]. Particularly, cyberattacks, some falsified data injection, customized command injection [15] [16], and replay attacks [17], are becoming very sophisticated in the existing detection strategies. As a result, they become a part of the natural operating environment [18]. In this vien, there is a crucial need for some expert tools, which are good in modeling temporal dependencies, good in modeling discriminative feature extractions, so that they

can in fact model very high-level complex event types at various granularities ranging from binary (attack vs. non-attack) to multi-class (specific attack types, natural faults, and normal operations) [19, 20, 21].

Intelligent power grids have continuously integrated information technology (IT) and operational technology (OT) solutions to facilitate data-centric computation, real-time event monitoring, and efficient data processing, thereby enhancing the reliability and resilience of the energy system [22, 23]. IT and OT systems offer great opportunities for improving smart grid communication networks; however, they lead to issues concerning security and privacy [24, 25, 26]. The convergence of these technologies exposes them to threats such as the compromise of data security of sensitive personal information (for example, user location, social security numbers, and home address) as well as critical power usage data theft [27, 28]. Figure 1 shows the interaction across different entities involved in the smart grid, including control centers, power plants, homes, smart buildings, and industries, where power is efficiently distributed among them. Artificial intelligence (AI) techniques are increasingly employed to ensure flexibility and responsiveness to meet power demands dynamically and diversely all over the system.

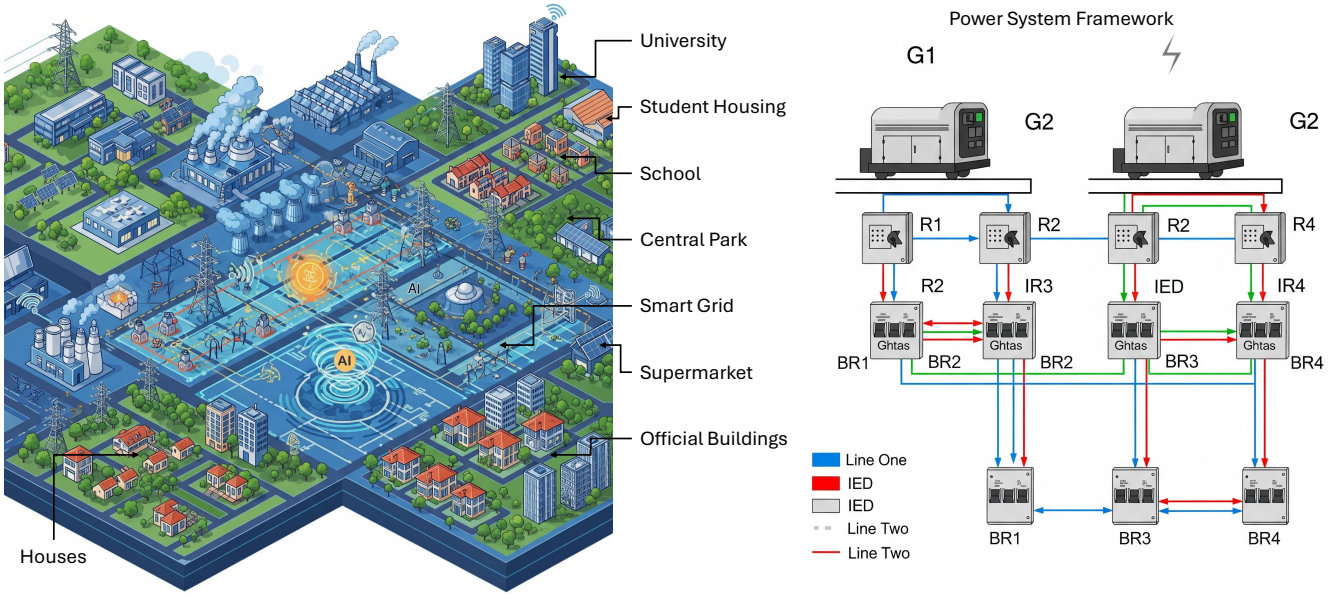

Figure 1. Architecture of a smart power grid illustrating the interaction among distribution centers, power plants, homes, smart buildings, and industries. The system offers dynamic power distribution and integrates IT and OT to support real-time monitoring and control.

This study addresses these challenges through introducing a novel methodology that employs machine learning (ML) to classify disturbances and cyber penetrations of a power system using heterogeneous time-synchronized data. The proposed approach utilizes rich temporal information from synchrophasor measurements (e.g., voltage phase angles, current magnitudes, frequency deviations), device logs, intrusion detection alerts, and relay records in building unique temporal signatures associated with diverse event types. This signature, represented as common sequential patterns, is a high-endpoint feature that is last trained by a robust classifier capable of differentiating between contingency, which forms both natural faults and stealthy cyber intrusions.

The proposed approach presents several methodological innovations. First, it applies fairly advanced preprocessing methods to enable quality data from the different sources. Second, feature selection methods are incorporated, such as Random Forest Importance (RFI), Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and Mutual Information (MI), for dimensionality reduction while maintaining informative characteristics. Third, it applies hyperparameter optimization using Optuna in tune with ML models such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting, and Neural Networks to improve the accuracy of classification and performance for generalization. The proposed approach is validated across three classification experiments: binary classification (attack versus non-attack), three-class classification (attack, natural, and no-event), and multi-class classification of 37 specific cases, including short-circuit faults, line maintenance, normal operation, and various cyber-attack vectors. The dataset on which this study relies was created by researchers from Oak Ridge National Laboratory (ORNL) and contains 128 features per instance, with instances derived from PMUs, control panel logs, Snort alerts, and protective relay logs [29, 30]. Each instance has been tagged with scenario identifiers, fault locations, and load conditions, facilitating an extensive evaluation of classification performance across various operating contexts. The significance of the proposed approach lies in its potential to improve situational awareness, automate

response directives, and enhance the smart grid's resilience against the evolution of attacks on the cyber-physical infrastructure. The proposed approach provides a unified framework through which disturbances and cyber attacks would be classified with high fidelity (even if the attacks are engineered to look natural) and assists in the evolution of intelligent and adaptive defense methodologies for next-generation power systems. Accordingly, this study contributes greatly to the evolution of cybersecurity in the smart grid in the following dimensions:

- Unified detection framework that detects natural disturbances and cyber-attacks of various kinds in parallel, tackling the problem of binary, three-class, and multi-class classification in one architecture.
- Feature selection in combination with RFI, PCA, RFE, and Mutual Information, considering dimensionality reduction while keeping domain-relevant and interpretable features across the board.
- Hyperparameter tuning via Optuna offered the best set of machine-learning models for generalization and robustness across Logistic Regression, SVM, Random Forest, Gradient Boosting, and Neural Nets.
- Explainable AI via Permutation SHAP gives interpretable attributions that bolster operator trust, accountability, and situational awareness in the understudied yet critical power system environment.
- Validation on 37 event scenarios, consistently outperforming individual datasets on the metrics of precision, recall, F1-score, accuracy, and specificity with performance levels $> 96\%$, and sustaining across datasets at $> 93\%$.

The remaining parts of this study are structured as follows: Section 2 presents a review of works in the relevant area, highlighting the development of intrusion detection approaches to power systems. Section 3 introduces a description of the suggested framework, focusing on data-pre-processing, feature selection, machine learning classifications, and explainability methods. Section 4 contains the experimental setup and results obtained therefrom, demonstrating the workings of this methodology for binary, three-class, and multi-class classifications and some commentary on the implications of the findings and contributions of this work. Section 5 addresses the study limitations. Finally, Section 6 summarizes the key conclusions and outlines future research directions.

Related Studies

Pan et al. [31] conducted a case study to evaluate an intrusion detection system (IDS) methodology based on sequential pattern mining and Bayesian networks (BN). They focused on a modified 2-bus, 2-generator power transmission system derived from the IEEE 9-bus, 3-generator system. An experimental study was developed based on nine scenarios, including four power system disturbances and five cyber-attacks. Time-synchronized synchrophasor data, relay logs, and control signals were collected from a hardware-in-the-loop. Sequential patterns were extracted to represent events with unique temporal signatures. A BN-based model was then constructed using these patterns to classify events. All nine scenarios were successfully classified post hoc by the IDS. This proves the effectiveness of the proposed method in distinguishing between natural faults and cyber intrusions.

In [11], the authors implemented a common path-mining approach to detect and classify disturbances and cyber-attacks in a simulated three-bus, two-line transmission system. They stored the heterogeneous time-synchronized data from Phasor Measurement Units (PMUs), relays, and system logs in one database. The sensor data was continuously quantized into discrete states. The FP-growth algorithm was used to mine frequent sequential patterns. These mined sequences constitute unique event signatures. They trained a classifier on these patterns to distinguish among symmetric and asymmetric faults (e.g., single-line-to-ground, line-to-line, double-line-to-ground) and cyber-attacks that acted like fault behaviors. Their methodology showed high accuracy in identifying both physical disturbances and adversarial activities, validating its potential for real-time intrusion detection in smart grids.

To support classification of disturbances and cyber-attacks, Pan et al. [32] developed a framework for automatic discovery of common paths from labeled power system data logs. Data were gathered from a simulated three-bus, two-line transmission system under various fault conditions and cyber-attack scenarios. The FP-growth algorithm extracted frequent sequential patterns from quantized sensor measurements and system logs. These patterns served as input features for a classifier capable of distinguishing between different types of faults (e.g., 1LG, LL, 2LG, 3LG) and cyber-attacks such as command injection, Aurora attacks, and fault replay attacks. The case study showed that the proposed method could effectively identify unique temporal signatures for each event type and achieve accurate classification across multiple attack and disturbance categories.

In addition, Hink et al. [30] applied multiple machine learning algorithms to distinguish between power disturbances and cyber attacks from heterogeneous data obtained from a gas pipeline SCADA system and an electric transmission system. They further studied several classification schemes, including binary (normal against attack), three-class (attack, natural, no-event), and multi-class (specific fault and attack types). The learners included RFs, JRip, Adaboost+JRip, SVM, and rule-based

classifiers. Experiments were done using the 10-fold cross-validation strategy over 15 datasets. Evaluation of performance used feature importance analysis and precision metrics. Random Forests, JRip, and Adaboost+JRip showed the best precision for cyber-attack detection. The results highlighted that machine learning is fairly applicable for intrusion detection applications in industrial control systems. They underlined the necessity for collecting balanced and labeled training data to ensure robustness when dealing with unseen data.

Following them, a study by Zaman et al. [33] validated a machine learning-based IDS design framework using the ORNL dataset, which emulates a power transmission system testbed. Recursive Feature Elimination with Random Forest (RFE-RF) was employed to reduce feature dimensionality. The top N features (varying in number from 10 to 110) were chosen for model-training purposes. The final cross-validation was through augmented training data; binary classification (normal versus intrusion) was considered. The experiment has analyzed fifteen CSVs consisting of 128 features (116 PMU measurements and 12 synthetic features). Results determined that for informative features retained, fewer features improved the generalization of the model. The data augmentation and feature selection methods were also shown to be necessary for enhancing the robustness and scalability of the IDS model in power system applications.

Similarly, a hybrid optimization technique named BGWO-EC is proposed in [34] to boost performance in intrusion detection of smart grid environments. The name BGWO originates from "binary grey wolf optimization and ensemble classifier". The BGWO algorithm was applied to extract optimal subsets of features from the ORNL dataset, which includes 128 features per instance. An ensemble classifier was trained based on this reduced feature set for the binary and multi-class classification experiments. Experimental analysis confirmed that the BGWO-EC scheme outperformed all existing benchmark methods concerning accuracy, precision, recall, and F1-score on all 15 datasets. The technique achieved comparative classification accuracy as high as 98.63% for some datasets, thus demonstrating the merits of metaheuristic-based feature selection combined with ensemble learning for a robust IDS design.

A recent work of Naeem et al. [35], thus far, explored deep-stacked ensemble learning strategies on power systems for intrusion detection using the ORNL dataset. They investigated the effects of feature selection techniques applied, such as genetic principal component analysis and grey wolf optimization (GWO), on the performance of the models. Deep stacking classifiers, combining base learners such as RF and GB, with a meta-classifier, improved detection accuracy and reduced false alarms. Compared with the traditional artificial neural networks employing hand-crafting of features, it widened the effectiveness scope. In this regard, stacked ensembles with GWO-selected features attained better classification accuracy and generalization for unseen data during training, thus substantiating the emphasis on advanced ensemble architectures in smart grid cybersecurity.

Increased research considering the issues of adversarial false data injection attacks (AFDIA), as well as state estimation and multi-label correlational vulnerabilities in modern deep learning models, has been the hallmark trend. Tian et al. [36] proposed EVADE, which is a targeted adversarial attack framework that perturbs a small number of state variables through the use of saliency maps to be able to bypass conventional Bad Data Detectors (BDD) and Neural Attack Detectors (NAD) to achieve very high stealth and success rates. In their next work [37], the general multi-label adversarial assault framework is recommended for deep learning-based FDIA markers, showing how multiple attack labels are manipulated simultaneously under physical constraints, bringing serious threats for any resource that operates under fine-granularity classification. Meanwhile, improvements are realized by establishing ADMM-based generalizations of adversarial attacks for multi-label detectors, emphasizing cost-effectiveness and realism when designing perturbations.

Moreover, their earlier work [38] examined joint adversarial example and FDI attacks (AFDIA), demonstrating that perturbations to state variables, as opposed to raw measurements, leave attackers therein undetected by either BDDs or deep learning models. Collectively, these studies reveal an important gap: deep learning provides powerful detection, but such power might also be implicated in introducing other attack surfaces that take advantage of model fragility. In contrast, our framework does not rely on state estimation or deep neural architectures; rather, we use ensemble tree-based models and those trained on heterogeneous time-synchronized sensor and log data; hence, we are intrinsically less amenable to gradient-based adversarial manipulation. Furthermore, the use of Permutation-based SHAP explainability offers transparency in feature attribution, a defense mechanism that many adversarial attack papers lack, thus enabling validation of decisions by operators and detection of potential spoofing via anomalous SHAP patterns. Therefore, while prior studies have focused on modeling methods employed by adversaries to evade detection, this study proposes the construction of interpretable classifiers that resist a wide spectrum of sophisticated and stealthy attacks, mimicking natural fault scenarios during the activities of adversaries.

Collectively, these studies trace the evolution of intrusion detection methodologies tailored for modern power systems, from early sequential pattern mining and probabilistic modeling to more sophisticated machine learning and ensemble-based frameworks. Initially, researchers relied on extracting temporal state transitions via algorithms like FP-growth to build event-specific signatures. Later works introduced feature engineering, recursive feature elimination, and synthetic data augmentation to improve model generalizability. Recent studies have shifted toward hybrid approaches that combine metaheuristic optimization (e.g., GWO) with deep ensemble learning to enhance classification accuracy and robustness. Across binary, three-class,

and multi-class classification experiments, these studies consistently demonstrate the value of integrating domain-specific knowledge with data-driven techniques to address the dual threats of physical disturbances and cyber-attacks in smart grids.

Research gaps

Although significant advancements have been made in developing data-driven and machine learning approaches to intrusion detection on power systems, a large number of serious open issues remain, including:

- The major contribution of existing studies is either physical fault detection or cyber-attack identification. Still, in the narrow sense, there is very little scope that may enable the detection of coordinated or blended cyber-physical attacks, among the most dangerous threats to modern power grids.

- Although the accuracy of machine learning and ensemble models appears very high, this is undermined by their lack of transparency and interpretability, which diminishes operator confidence. The unexplainable "black box" outputs cannot support situational awareness or automatic decision-making in power grids.

- Accuracy and interpretability limitations, as recent frameworks do not incorporate evidence coming from the power system domain into practices of feature engineering or model design.

- High computational complexities due to using advanced methods like meta-heuristics optimization and explainable AI, which make them unsuitable for either real-time or large-scale applications until improved.

- Most methods generically classify events (e.g., an attack vs. no attack), which is insufficient for steering operation.

In this vein, smart grids may remain vulnerable to ever-evolving cyber-physical threats that can go undetected and break down trust in automation, compromising the backbone of national infrastructure. A unified framework for detecting real-time natural disturbances and cyber intrusions with high reliability and computational efficiency should address these gaps. Table 1 shows a gap-contribution mapping.

Table 1. The gap-contribution mapping.

No.	Identified Research Gap	Proposed Contribution
1	Fragmented focus: Existing studies treat physical faults and cyber-attacks as separate problems, failing to detect blended or coordinated threats.	A unified framework that simultaneously classifies natural disturbances and multiple cyber-attack types (binary, three-class, and multi-class) ensures comprehensive cyber-physical scenario coverage.
2	Black-box limitations: High-performing ML models lack interpretability, undermining operator trust in critical decision-making.	Integration of explainable AI via Permutation SHAP enables transparent feature attribution and improves trust, accountability, and situational awareness.
3	Underuse of domain-specific knowledge: Limited incorporation of power system expertise in model design reduces accuracy and interpretability.	A hybrid feature selection strategy (RFI, PCA, RFE, MI) that blends statistical/data-driven insights with domain relevance ensures both informative and interpretable feature sets.
4	Computational inefficiency: Advanced optimization and explainability methods introduce overhead, restricting real-time scalability.	Efficient hyperparameter optimization with Optuna, improving performance while reducing trial-and-error overhead, making the framework scalable and deployment-friendly.
5	Limited fine-grained classification: Many frameworks oversimplify detection into binary outputs, which are inadequate for real-world response protocols.	Fine-grained event classification across 37 scenarios, achieving $> 96\%$ performance on individual datasets and $> 93\%$ on combined datasets, thus supporting precise, actionable defense mechanisms.

Methodology

This study proposes a methodology constituted systematically for the classification of power system disturbances and cyber-attacks with time-synchronized heterogeneous data (see Figure 2). It involves feature engineering, machine learning classification, and meta-heuristic optimizations to ensure a greater classification accuracy in binary, three-class, and multi-class experiments.

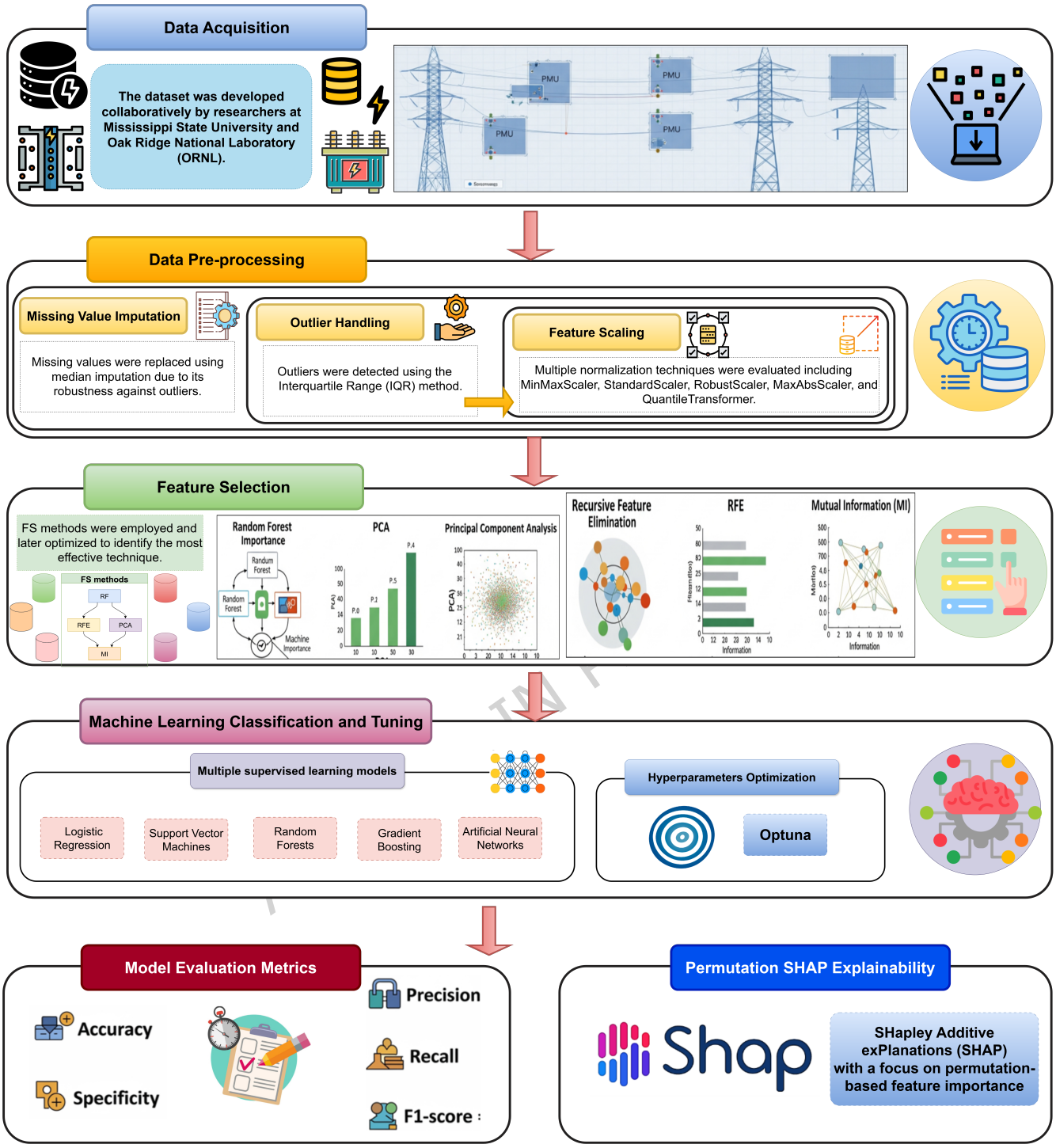

Figure 2. The proposed framework to classify disturbances and cyber-attacks is illustrated in the figure. It involves data pre-processing, feature selection, and machine learning classification, complemented by metaheuristic optimization and explainability through permutation SHAP. This systematized end-to-end method guarantees accurate and robust classification in binary, three-class, and multi-class cases using time-synchronized heterogeneous data from PMUs, control logs, and intrusion detection systems.

Materials and Partitioning

The used dataset was developed jointly by Mississippi State University and Oak Ridge National Laboratory (ORNL) [11, 31, 32]. It is derived from a simulated high-fidelity environment for the power transmission system designed to simulate real-world

operational dynamics while allowing the controlled injection of both natural disturbances and cyber-attacks. This simulation environment provides precise reproducibility and labeling of events, which are tremendously important for training supervised machine learning models in cybersecurity, where ground truth is mandatory.

This dataset comprises 37 unique event scenarios, classified under three main categories: (i) *natural events*: single-line-to-ground faults, double-line-to-ground faults, line maintenance, and load shedding; (ii) *no-event conditions*: representing normal steady-state operations to differentiate load levels; and (iii) *cyber attacks*: remote tripping command injection, false data injection on PMU measurements, replay attacks on relay logs, and Aurora-style oscillatory attacks. These engineered scenarios emulate real threats to modern smart grids; many of these attack vectors have been designed to resemble signatures of natural faults, thus enabling the model's ability to discriminate between subtle adversarial manipulations and benign anomalies. Each instance in the dataset contains 128 features generated from different time-synchronized sources: synchrophasor measurements (voltage magnitude, phase angle, frequency deviation, current magnitude), gathered from four appropriately spaced PMUs; discrete control panel logs recording breaker operation and manual intervention; Snort intrusion detection alerts indicating network-level anomalies; and protective relay logs specifying trip signals and fault-clearing actions. Timestamped records ensure temporal alignment across all data sources, a prerequisite for modeling dynamic system behavior during transient event conditions. To facilitate the multi-class classification experiments of binary (attack vs. non-attack), three-channel (attack, natural, member of no-event) scenarios, and fine-grained sub-multi-class (37 scenarios), the original dataset is partitioned into 15 subsets (Data 1 to Data 15). Each subset corresponds to a distinct combination of scenario type, fault location, load condition, or attack vector. Some examples are: (i) *Data 1*: pertain to short-circuit faults at Bus 1 under high load; (ii) *Data 5*: concerns a remote tripping command injection targeting Relay R2 under medium load; (iii) *Data 10*: concerns a false data injection attack on PMU voltage readings near Generator G3; (iv) *Data 13*: pertains to line maintenance activities performed without any active faults or attacks.

The fifteen datasets were created by random sampling about 1% of the entire ORNL dataset for computational convenience while upholding a high degree of statistical significance. The "Combined" datasets (used for evaluation of generalization over heterogeneous conditions) were created by pooling together samples of multiple individual datasets. More precisely, "Combined (80%:20%)" and "Combined (90%:10%)" refer to stratified train-test splits in which 80% or 90% of the combined dataset was assigned for training, respectively. At the same time, the rest was reserved for testing. Through stratification, class representatives of all 37 scenarios were preserved in the training set and the test set to prevent any bias due to class imbalance.

For each dataset, in the last column, the markers contain metadata specifically stating: (a) identifier of the scenario (e.g., Scenario 7 = "False Data Injection at PMU R4"), (b) location of the fault or attack (e.g., "Bus 2", "Relay R3"), and (c) load condition (e.g., "High Load", "Normal Load"). This equally rich contextual labeling allows proper classification and further fosters post-hoc analysis linking model decision-making upon specific physical or cyber events; a fundamental requirement for explainable AI in safety-critical systems.

There exists a public availability of all datasets in the official repository maintained by the ORNL team (https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets), thus assuring reproducibility and comparison to be made with any future research. Due to its comprehensive feature set, labeled diversity, and realistic simulation fidelity, this dataset represents an ideal benchmark for developing, validating, and comparing machine learning approaches to detect blended cyber-physical threats to modern power systems.

Data pre-processing

Before model training, several pre-processing steps were applied. The first step is "Missing Value Imputation," where missing values were replaced using median imputation due to its robustness against outliers. The second step is "Outlier Handling" where outliers were detected using the Interquartile Range (IQR) method IQR $= Q_3 - Q_1$. Any value below $Q_1 - 1.5 \times$ IQR or above $Q_3 + 1.5 \times$ IQR was considered an outlier and replaced with the nearest non-outlier value. The third step is "Feature Scaling": Multiple normalization techniques were evaluated, including MinMaxScaler, StandardScaler, RobustScaler, MaxAbsScaler, and QuantileTransformer. For example, MinMax scaling transforms a feature x as $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$.

Feature Selection

Four feature selection methods were employed to reduce dimensionality while preserving informative features; four feature selection methods were later optimized to identify the most effective technique for classifying power system disturbances and cyber-attacks [39]. The number of selected features was considered a hyperparameter, varying from 25% to 100% of the original feature set (totaling 128 features). Every technique presents distinct advantages based on the qualities of the data and the requirements of the model. The utilized techniques are described below:

- **RFI**: This filter-based feature selection approach ranks features according to their importance scores derived from a trained Random Forest classifier [40]. Each tree in the ensemble computes feature importance based on the decrease in impurity (e.g., Gini index or entropy) when a feature is used for splitting. The overall importance score I_j for feature j

is calculated using 1 where T is the number of trees in the forest. Features with higher importance scores are retained, ensuring that the most discriminative attributes contribute to classification performance.

$$I_j = \frac{1}{T} \times \sum_{t=1}^{T} \Delta \text{ impurity}_j^{(t)} \tag{1}$$

- **Principal Component Analysis (PCA)**: PCA is an unsupervised linear dimensional reduction technique that changes the input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ into a lower-dimensional space after a few operations to identify orthogonal components that capture the most variance [41]. The forward transformation is given by $\mathbf{T} = \mathbf{X} \cdot \mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{d \times k}$ has the top k eigenvectors of the covariance matrix $\mathbf{C} = \frac{1}{n-1}\mathbf{X}^\top \cdot \mathbf{X}$; the transformed dataset $\mathbf{T} \in \mathbb{R}^{n \times k}$ is now reduced to k dimension. PCA handles multicollinearity, allowing good generalization by learning the most informative latent variables.
- **Recursive Feature Elimination (RFE)**: RFE is an iterative wrapper method that removes the least essential features using a base estimator to score the features on each removal [42]. It trains the base estimator (SVM, RF, etc.) at each iteration and calculates the importance of the features; the weakest is booted till the number of features is achieved k. With the loss function $L(\hat{y}, y)$ minimized over each successive feature subset, the procedure goes on to bring subsets with lower loss as it removes more features with each iteration of RFE as in Equation 2, where F is the gene set of all features, S is the gene set of selected features in the current iteration, and $f_S(x)$ is a prediction function using only the features S. RFE picks a final set of features with maximum predictive power and lowest redundancy.

$$\min_{S \subseteq F, |S|=k} L(f_S(x), y) \tag{2}$$

- **Mutual Information (MI)**: MI quantifies the amount of information obtained about one random variable (e.g., class label Y) through another (e.g., feature X) [43]. It is beneficial for capturing non-linear dependencies between features and labels. MI is defined in Equation 3 where $p(x,y)$ is the joint probability distribution and $p(x), p(y)$ are marginal distributions. Features with higher mutual information are prioritized, as they provide greater insight into the underlying class structure.

$$\text{MI}(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \times \log\left(\frac{p(x,y)}{p(x) \times p(y)}\right) \tag{3}$$

These feature selection techniques were evaluated across multiple datasets and classification experiments (binary, three-class, multi-class), with the optimal method and feature count determined via hyperparameter tuning using Optuna.

Machine Learning Classification and Tuning

Among various supervised learning models tried, namely, Logistic Regression, SVM, Random Forests, Gradient Boosting, and Artificial Neural Networks [44], hyperparameter tuning was done using Optuna [45], an automated library for hyperparameter optimization with efficient sampling and pruning strategies inside to fill the search space. The objective function maximizes a weighted average of evaluation metrics as defined in Equation 4, where w_1 to w_5 are user-defined weights; for this study, it has been set to 0.2.

$$\text{Objective} = w_1 \times \text{Accuracy} + w_2 \times \text{Precision} + w_3 \times \text{Recall} + w_4 \times \text{F1} + w_5 \times \text{Specificity} \tag{4}$$

Optuna employs the Tree-structured Parzen Estimator (TPE) algorithm as its default sampler for constructing probabilistic models of the objective function [46]. TPE maintains two distributions over hyperparameters: $l(x)$ for configurations that resulted in good objective values, and $g(x)$ for those that did not. The next candidate configuration x is selected by maximizing the expected improvement (EI) using Equation 5 where y^* is the current best objective value, γ is a mixing coefficient, and expectations are taken over the observed outcomes y [47]. This strategy allows TPE to balance exploration and exploitation during the search process [48].

$$\text{EI}(x) = \frac{l(x)}{g(x)} \times [\gamma \times \mathbb{E}_l[\max(0, y^* - y)] + (1-\gamma) \times \mathbb{E}_g[\max(0, y^* - y)]] \tag{5}$$

To accelerate training and reduce computational overhead, the Median Pruner was applied to early-stop unpromising trials [49]. For each trial at a given step t, the pruner compares the intermediate result (e.g., validation score) against the median of results from previous trials at the same step. Suppose the current trial's performance falls below the median. It is terminated early using Equation 6 where k is the number of completed trials up to that point. $\text{value}_t^{(i)}$ is the performance metric at step t for trial i. This dynamic mechanism ensures that only the most promising configurations are fully trained, significantly improving optimization efficiency without affecting the model quality [50].

$$\text{If value}_t < \text{median}(\{\text{value}_t^{(i)}\}_{i=1}^{k}) \Rightarrow \text{Prune Trial} \tag{6}$$

Model Evaluation Metrics

The evaluation of the classification performance was carried out using standard metrics in Equation 7 [51]. In Equation 7 TP, TN, FP, and FN indicate true positives, true negatives, false positives, and false negatives predictions, respectively. Cross-validation and holdout validation were employed to guarantee the model's generalization capability [52]. The performance analysis was detailed by constructing confusion matrices, ROC curves, and Precision-Recall curves [53].

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{F1-score} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \\
\text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \\
\text{Specificity} &= \frac{\text{TN}}{\text{TN}+\text{FP}}
\end{aligned} \tag{7}$$

Explainability using Permutation SHAP

The interpretability of our classification models was carried out using SHapley Additive exPlanations (SHAP) with permutation-based feature importance. SHAP is a game-theoretic approach that assigns an importance value for a given prediction to each feature with consistency and local accuracy [54]. For the event classification scenario within the power system, knowing how such parameters as synchrophasor measurements, relay logs, or intrusion detection alerts contribute to the vote of the model is crucial for validating its decisions and earning the system's trust to be deployed.

The Permutation SHAP methodology is based on estimating the feature's marginal contribution towards the model's output by permuting the respective feature values and observing the performance metrics. Specifically, for any instance, x, the SHAP value, $\phi_j(x)$, for feature j is computed as in Equation 8, where F is the set of all features, S is the subset of features excluding feature j, and $f(S)$ refers to the model's prediction using solely the feature set S. The SHAP value quantifies the average effect of including feature j over all possible subsets of features. It thus represents a fair and consistent measurement of feature importance.

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \big(f(S \cup \{j\}) - f(S)\big), \tag{8}$$

In practice, SHAP value computation, to be precise, is an expensive operation, especially with high-dimensional datasets such as ours. As a remedy, we used a permutation-based approximate method that estimates SHAP values by stochastic shuffling of feature values and measuring the change in model performance. The permutation importance of a feature j is defined as in Equation 9 where D is the original dataset, D_j is the dataset with feature j permuted, $L(\cdot,\cdot)$ is the loss function (for example, cross-entropy or mean squared error), and y being the true label. The greater the PI_j value, the more the feature impacts modeling.

$$\text{PI}_j = \mathbb{E}_{x \sim D}\big[L(f(x), y)\big] - \mathbb{E}_{x' \sim D_j}\big[L(f(x'), y)\big], \tag{9}$$

Analyzing permutation SHAP revealed the relative importance of different sources of data. For instance, synchrophasor measurement parameters such as voltage phase angle and frequency deviation were assessed as some of the most important

features for distinguishing between cyber attacks and natural disturbances. This is corroborated by domain knowledge of the events, as they reflect changes in the power system from these measurements. In this respect, Snort log intrusion detection alerts have also been weighted significantly for identifying specific types of cyber attack, demonstrating the complementarity of different data sources.

Permutation Shapley added both interpretability and justification from domain knowledge relative to the results achieved by the high performance of the machine-learning models. This is the most important aspect for successfully deploying machine-learning technology in applied settings for critical infrastructure, such as power grid security.

The Framework Pseudocode

Algorithm 1 defines the detailed pseudocode for the proposed intrusion detection framework for the classification of power system disturbances and their associated cyber-attacks. This pseudocode depicts the methodology that proceeds from the data preprocessing stage to feature selection to the final machine learning classification stage, assisted by metaheuristics optimization techniques like TPE. The representation design captures in a quite straightforward manner the modularity of how heterogeneous, time-synchronized data from PMUs, relay logs, and intrusion detection alerts undergo processing, transformation, and classification in binary, three- and multi-class contexts. This algorithmic blueprint aims to guide the development, evaluation, and improvement across various smart grid contexts.

Experiments and Discussion

In this section we describe the experiments carried out on the proposed methodology for classifying power grid disturbances and cyber-attacks using heterogeneous time-synchronized data, along with analysis outcomes. Three classification tests were conducted: binary classification (attack vs. non-attack), three-class classification (attack, natural event, no-event), and multi-class classification with 37 different scenarios, including short-circuit faults, line maintenance, normal operations, and various types of cyber assaults. Each experiment had its set of feature scaling procedures, feature selection techniques, and machine learning classifiers, with performance evaluated using precision, recall, F1-score, accuracy, and specificity.

The software environment was developed in Python, running on Windows 11, with Anaconda as a package and environment manager. The hardware setup consisted of an Intel Core i7 CPU, 128 GB of RAM, and an NVIDIA GPU with 6 GB of dedicated memory. Such a hardware setup provided enough computational power to train and properly evaluate the models used throughout this study.

Binary classification results. Table 2 summarizes results from the binary classification experiment. For this experiment, instances were classified with the Extra Trees classifier (for all subsets) either as attacks (cyber-attacks) or as normal (non-attacks). The feature subsets were prepared from various feature selection methods, such as RFE, Random Forest (RF)-based importance, and MI, with feature ratios from 30% up to 100%. Among the moderately successful classifiers, the one applying D1, Min-Max scaling and RFE on the features at 45% received the highest average score of 98.07%. Other configurations that also performed well were MI-feature selection on Quantile-scaled data (D10, 97.62%) and RF-feature selection on MaxAbs-scaled data (D12, 97.12%). The above shows that high classification accuracy can still be maintained with very small feature sets, thereby increasing the confidence in the usefulness of feature selection strategies.

Three-Class Classification Results: The three-class classification results are presented in Table 3, where events were categorized into one of three classes: attack, natural event, or no event. The table shows metrics such as precision, recall, F1-score, accuracy, specificity, and averages, giving insights into different pre-processing techniques, feature selection methods, and scaling strategies.

The performance is consistently high for the individual datasets, with most metrics over 96%. The best average performance of 98.05% driven by remarkable precision (98.19%), recall (98.19%), F1 score (98.19%), accuracy (98.46%), and specificity (97.19%) was achieved for Datapoint 5 using Standard scaler with no feature selection. Data 13 and Data 11 are notable datasets, which achieved average scores above 97%. The choice of scaler and feature selection method impacts performance. For instance, datasets defaulting to MinMax scaling (e.g., Data 1, Data 4) and those implementing MI for feature selection (e.g., Data 4, Data 11) seem to have success. This proves that one or both of these pre-processing manipulations may have been particularly useful in this classification experiment.

The performance metrics for combined (i.e., all scenarios) datasets, incorporating data from several sources, are slightly lower than those of the best individual datasets, yet still very good. In the “Combined (80%:20%)” dataset, Robust scaling and recursive feature elimination (RF) with a 40% feature selection ratio yielded an average of 93.18%. The “Combined (90%:10%)” dataset, with no scaling, but RF at a 50% selection ratio, scored slightly higher with an average of 93.52%. The consistency across different training and testing data splits showcases the importance of combining datasets. The combined datasets do not outperform the best individual datasets, but they step in as a practical solution when it becomes important or beneficial to aggregate data from diverse sources (for a restricted amount of data or heterogeneous data).

Algorithm 1: Pseudocode for Classification of Power System Disturbances and Cyber-Attacks.

Input : Raw power system dataset $\mathscr{D}$, Labels $\mathscr{Y}$, Hyperparameter bounds Θ, Number of trials T
Output : Trained classifier $\hat{\mathscr{C}}$, Predicted event class $\hat{\mathscr{Y}}$, Evaluation metrics $\mathscr{M}$

// Load and prepare heterogeneous time-synchronized data
Stage 1: Data pre-processing for *each instance* $D_i \in \mathscr{D}$ **do**
  // Handle missing values using median imputation
  Replace missing values in D_i with median value
  // Detect and cap outliers using the IQR method
  Compute interquartile range: $IQR = Q_3 - Q_1$
  Cap values below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$
  // Normalize features using selected scaler
  Apply feature scaling: $x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$
// Reduce dimensionality while preserving informative features
Stage 2: Feature Selection // Evaluate multiple feature selection techniques
Select technique $\mathscr{F} \in \{\text{RFE}, \text{RF}, \text{PCA}, \text{MI}\}$
Set feature ratio $r \in [0.25, 1.0]$
Apply selected method to reduce feature set: $\mathscr{X}_{\text{reduced}} = \mathscr{F}(\mathscr{X}, r)$
// Extract temporal patterns from time-synchronized data
Stage 3: Model Training and Hyperparameter Optimization Initialize Optuna study with TPE sampler and Median Pruner
Define objective function:

$$\text{Objective} = w_1 \times \text{Accuracy} + w_2 \times \text{Precision} + w_3 \times \text{Recall} + w_4 \times \text{F1-score} + w_5 \times \text{Specificity}$$

for $t \in 1 \dots T$ **do**
  // Sample hyperparameters from search space
  Sample $\theta_t \sim \Theta$
  // Train classifier using selected features and patterns
  Train $\mathscr{C}_{\theta_t}$ on $\mathscr{X}_{\text{reduced}}, P, \mathscr{Y}$
  // Prune unpromising trials early
  If trial underperforms median, prune: $\text{value}_t < \text{median}(\text{value}_{1:t})$
  // Compute evaluation metrics for current trial
  Evaluate $\mathscr{M}_t = \{\text{Accuracy}, \text{F1}, \text{AUC}, \text{etc.}\}$
// Select best-performing model based on cross-validation
$\hat{\mathscr{C}} = \arg\max_{\mathscr{C}_\theta} \mathscr{M}_{\text{avg}}$
// Classify disturbances and attacks across multiple experiments
Stage 4: Multi-Level Classification for *input signal* x_j **do**
  // Predict class label using trained model
  Predict $\hat{y}_j = \hat{\mathscr{C}}(x_j)$
  if *experiment* == *binary* **then**
    Classify as $\hat{y}_j \in \{\text{Attack}, \text{Non-attack}\}$
  else if *experiment* == *three-class* **then**
    Classify as $\hat{y}_j \in \{\text{Attack}, \text{Natural}, \text{No-event}\}$
  else if *experiment* == *multi-class* **then**
    Classify as $\hat{y}_j \in \{c_1, c_2, \dots, c_{37}\}$
// Evaluate performance using standard metrics
Stage 5: Performance Evaluation Compute:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Visualize results using confusion matrices, ROC curves, and precision-recall curves.
Return: Best classifier $\hat{\mathscr{C}}$, predicted labels $\hat{\mathscr{Y}}$, evaluation metrics $\mathscr{M}$

Multi-Class Classification Results: The multi-class classification results here can be seen in Table 4. These results reflect the performance of different configurations in distinguishing one of the 37 possible distinct power system event scenarios from the others, including natural faults, no-event conditions, and many forms of cyber-attack. The fine granularity of classes and the possible overlap of similar fault and attack patterns introduce a significant challenge. However, despite all these obstacles, the results lead to success for ensemble classifiers such as Extra Trees and Random Forests, which captured consistent, strong

Table 2. Performance metrics of binary classification across 15 datasets using various feature selection methods and scalers. Each dataset was evaluated using precision, recall, F1-score, accuracy, specificity, and average.

File	Scaler	Feature Selection (FS)	FS Ratio	Precision	Recall	F1	Accuracy	Specificity	Average
Data 1	MinMax	RFE	45	99.37%	95.15%	97.21%	98.79%	99.83%	98.07%
Data 2	Quantile	N/A	100	94.22%	91.38%	92.78%	95.66%	97.54%	94.32%
Data 3	MinMax	RFE	60	98.69%	93.78%	96.17%	97.79%	99.48%	97.18%
Data 4	N/A	N/A	100	98.44%	93.70%	96.02%	97.31%	99.22%	96.94%
Data 5	MinMax	RF	60	98.10%	93.24%	95.61%	97.55%	99.28%	96.76%
Data 6	MaxAbs	RFE	60	97.13%	91.44%	94.20%	96.65%	98.85%	95.65%
Data 7	MinMax	RFE	30	96.41%	94.47%	95.43%	97.71%	98.81%	96.57%
Data 8	MinMax	RF	80	96.90%	94.40%	95.63%	97.49%	98.76%	96.64%
Data 9	N/A	RF	70	96.89%	93.61%	95.22%	96.88%	98.50%	96.22%
Data 10	Quantile	MI	70	99.57%	93.93%	96.67%	98.09%	99.83%	97.62%
Data 11	Robust	RFE	60	98.86%	90.10%	94.28%	97.34%	99.66%	96.05%
Data 12	Quantile	RF	45	97.68%	95.11%	96.38%	97.58%	98.84%	97.12%
Data 13	N/A	N/A	100	98.16%	92.49%	95.24%	97.98%	99.51%	96.67%
Data 14	MaxAbs	RFE	45	96.46%	94.09%	95.26%	97.53%	98.76%	96.42%
Data 15	N/A	MI	75	97.42%	94.62%	96.00%	97.22%	98.64%	96.78%

Table 3. Classification results for the three-class experiment (attack, natural, no-event) across 15 datasets. The table shows performance metrics including precision, recall, F1-score, accuracy, specificity, and their average.

File	Scaler	Feature Selection (FS)	FS Ratio	Precision	Recall	F1	Accuracy	Specificity	Average
Data 1	MinMax	N/A	100	98.12%	98.12%	98.12%	98.21%	94.72%	97.46%
Data 2	N/A	RFE	75	96.51%	96.45%	96.48%	96.77%	92.97%	95.84%
Data 3	Standard	RF	50	96.66%	96.68%	96.67%	96.97%	94.67%	96.33%
Data 4	MinMax	MI	95	97.56%	97.57%	97.57%	97.80%	97.03%	97.51%
Data 5	Standard	N/A	100	98.19%	98.19%	98.19%	98.46%	97.19%	98.05%
Data 6	Quantile	N/A	100	97.22%	97.18%	97.20%	97.32%	94.16%	96.62%
Data 7	N/A	RF	70	97.47%	97.46%	97.46%	97.62%	95.31%	97.06%
Data 8	MinMax	MI	45	96.60%	96.62%	96.61%	96.90%	94.23%	96.19%
Data 9	N/A	RF	35	96.74%	96.75%	96.75%	97.12%	95.37%	96.55%
Data 10	MinMax	RF	55	97.18%	97.13%	97.16%	97.37%	93.93%	96.55%
Data 11	Standard	MI	70	97.96%	97.97%	97.96%	98.03%	95.52%	97.49%
Data 12	MaxAbs	RFE	35	97.45%	97.45%	97.45%	97.69%	96.25%	97.26%
Data 13	MaxAbs	RFE	35	98.10%	98.10%	98.10%	98.19%	94.92%	97.48%
Data 14	MaxAbs	RF	50	97.66%	97.66%	97.66%	97.69%	94.75%	97.08%
Data 15	N/A	MI	80	96.72%	96.72%	96.72%	97.20%	96.86%	96.84%
Combined (80%:20%)	Robust	RF	40	94.11%	94.13%	94.12%	94.57%	88.98%	93.18%
Combined (90%:10%)	N/A	RF	50	94.43%	94.44%	94.44%	94.86%	89.41%	93.52%

performances with such different preprocessing methods and feature-selection strategies.

Among these individual datasets, Data 7 and Data 11 produced the highest average performance of 96.66%, using the complete feature set without pre-processing or feature selection. By implication, in some cases, maintaining the complete feature set may give rise to just higher performance than any relevant feature reduction technique would conjure up. The latter may, however, save some computational effort. Other noteworthy configurations include Data 4, using Random Forest with Robust scaling and RF-based feature selection at a 65% ratio, achieving an average score of 96.65%, and Data 10, using Extra Trees with MinMax scaling and RF-based feature selection at an 85% ratio, scoring 96.56%. Across all 15 datasets, it is thus confirmed that several configurations clearly exceed 96% average performance for a firmer endorsement of fine-grained classification for heterogeneous power system events by the proposed method.

Regarding the combined datasets' performance for more diverse analyses, a slightly lower performance level than the best among individual datasets was observed; however, it was still resilient. The Composite Rating (90%:10%) dataset, being processed by MinMax scaling and without any feature selection, gathers an average score of 92.91, with precision, recall, F1-score, accuracy, and specificity being 88.55%, 88.49%, 88.52%, 99.33%, and 99.64%, respectively. Although the performance rate of these results does not match that of the best-performing individual datasets, it underscores the importance of joining datasets, as it establishes consistent values across many test and training splits. Such an approach may serve one well when aggregating data from so many sources is viable or necessary, perhaps when dealing with limited or heterogeneous sets.

Table 4. Evaluation outcomes of multi-class classification involving 37 distinct power system event scenarios. Metrics include precision, recall, F1-score, accuracy, specificity, and average performance across 15 datasets.

File	Classifier	Scaler	Feature Selection (FS)	FS Ratio	Precision	Recall	F1	Accuracy	Specificity	Average
Data 1	Extra Trees	Quantile	MI	45	92.03%	91.81%	91.92%	99.44%	99.70%	94.98%
Data 2	Extra Trees	Quantile	MI	70	93.15%	92.90%	93.03%	99.55%	99.75%	95.68%
Data 3	Extra Trees	MaxAbs	RF	85	94.24%	93.85%	94.04%	99.59%	99.79%	96.30%
Data 4	Random Forest	Robust	RF	65	94.76%	94.49%	94.63%	99.60%	99.77%	96.65%
Data 5	Extra Trees	MinMax	RFE	45	93.32%	93.03%	93.18%	99.56%	99.78%	95.77%
Data 6	Extra Trees	N/A	RFE	35	94.32%	93.97%	94.14%	99.61%	99.82%	96.37%
Data 7	Extra Trees	N/A	N/A	100	94.70%	94.53%	94.62%	99.66%	99.81%	96.66%
Data 8	Extra Trees	MinMax	N/A	100	93.28%	92.86%	93.07%	99.49%	99.70%	95.68%
Data 9	Extra Trees	MinMax	RFE	50	93.70%	93.26%	93.48%	99.55%	99.76%	95.95%
Data 10	Extra Trees	MinMax	RF	85	94.58%	94.38%	94.48%	99.58%	99.78%	96.56%
Data 11	Extra Trees	N/A	N/A	100	94.71%	94.54%	94.62%	99.63%	99.79%	96.66%
Data 12	Extra Trees	Quantile	N/A	100	94.67%	94.39%	94.53%	99.61%	99.78%	96.59%
Data 13	Random Forest	N/A	MI	60	92.89%	92.54%	92.72%	99.46%	99.72%	95.47%
Data 14	Extra Trees	MaxAbs	RFE	40	92.59%	91.93%	92.26%	99.45%	99.73%	95.19%
Data 15	Extra Trees	MinMax	MI	50	96.45%	96.45%	96.45%	96.69%	95.25%	96.26%
Combined (90%:10%)	Extra Trees	MinMax	N/A	100	88.55%	88.49%	88.52%	99.33%	99.64%	92.91%

The following were the pattern trends attained over the classification experiments:

- **Feature Selection**: RFE, MI, and RF-based importance together really increased the efficacy of the models, with perhaps no loss or even a gain to the accuracy.
- **Scaling Methods**: Min-Max and Quantile scalers have surpassed the rest regarding normalization, especially when working with effective feature selection.
- **Classifier Performance**: Ensemble methods showed consistency in performance over Extra Trees and Random Forest under binary-three class-multi-class settings.
- **Model Generalization**: The fact that we had good generalization owing to high training and testing splits is reflected in immense performance, especially when data were augmented synthetically and via cross-validation.

The experimental results indicate the robustness and versatility of our method in differentiating between physical disturbances and cyber-attacks in modern power systems under varied and challenging threats.

Permutation SHAP Explainability for Three-Class Datasets

A bar plot shown in Figure 4 contains information regarding the top-10 most important features that a combined testing subset could gain from the three-class classification. It can be observed in this bar plot that the importance of each feature is presented through its average effect on model predictions. It is worth mentioning that, in most occasions, synchrophasor measurements such as the voltage phase angles (VH) and zero-sequence currents (ZH) have been rated as very important features, which further highlight their important role in discriminating an ongoing attack from natural events and no-events. Such a finding would not be an artifact in statistics; it obeys the fundamental approaches of power systems in generating dynamic stress, where measurement deviations are often precursors of physical disturbances or malevolent tampering.

Permutation SHAP identified the following top-10 features: R3-PA2:VH, R1-PA3:VH, R4-PA7:VH, R2-PA:ZH, R2-PA5:IH, R4-PA3:VH, R1-PA7:VH, R4-PM2:V, R3-PA6:IH, and R2-PA6:IH. These features predominantly represent voltage-related measurements (denoted by "VH" and "V") and current-related measurements (denoted by "IH") from various PMUs across different regions, such as R1, R2, R3, and R4. The significance of these features aligns with domain expertise, as voltage phase angles and current magnitudes are vital indicators of power system integrity and can identify anomalies resulting from cyber-attacks or natural disturbances.

For example, the prominence of R2-PA:ZH, the zero-sequence current at Region 2, points towards sensitivity to ground faults or asymmetric attacks targeting unbalanced loads commonly exercised in Aurora-style attacks, which are devised to cause thermal stress without alerting immediate protective relays. Likewise, R4-PA7:VH and R3-PA2:VH illustrate the truncation of PMUs for early detection of generation units or major substations; their sudden changes in phase angles are usually indicative of false data injection attacks meant to mislead state estimators or incite misoperation of distance relays. In contrast, gradual deviations of R1-PA3:VH might match with events of natural load-shedding, or line maintenance, where system inertia dampens transient response.

Moreover, Figure 3 provides a beeswarm plot of SHAP values for the same set of features using the combined testing subset, providing a complementary view to the bar plot. Each point in this figure corresponds to one instance: the horizontal axis is the

SHAP value (impact on model output), and the vertical axis lists the features. The color scale here indicates normalized feature value so that the specific range of a feature can be observed on predictions. Larger values of R2-PA:ZH were strongly correlated to attack classes (contrary to the physics path of ground fault propagation), whereas lower values tend to cluster around classes of natural or no event, which indicates normal operation or symmetric faults. This detailed representation can bring actionable intelligence to operators; the operators could monitor R2-PA:ZH and R4-PA7:VH for real-time threat detection, prioritizing alerts triggered when these features behave with anomalous spikes or sustained deviation.

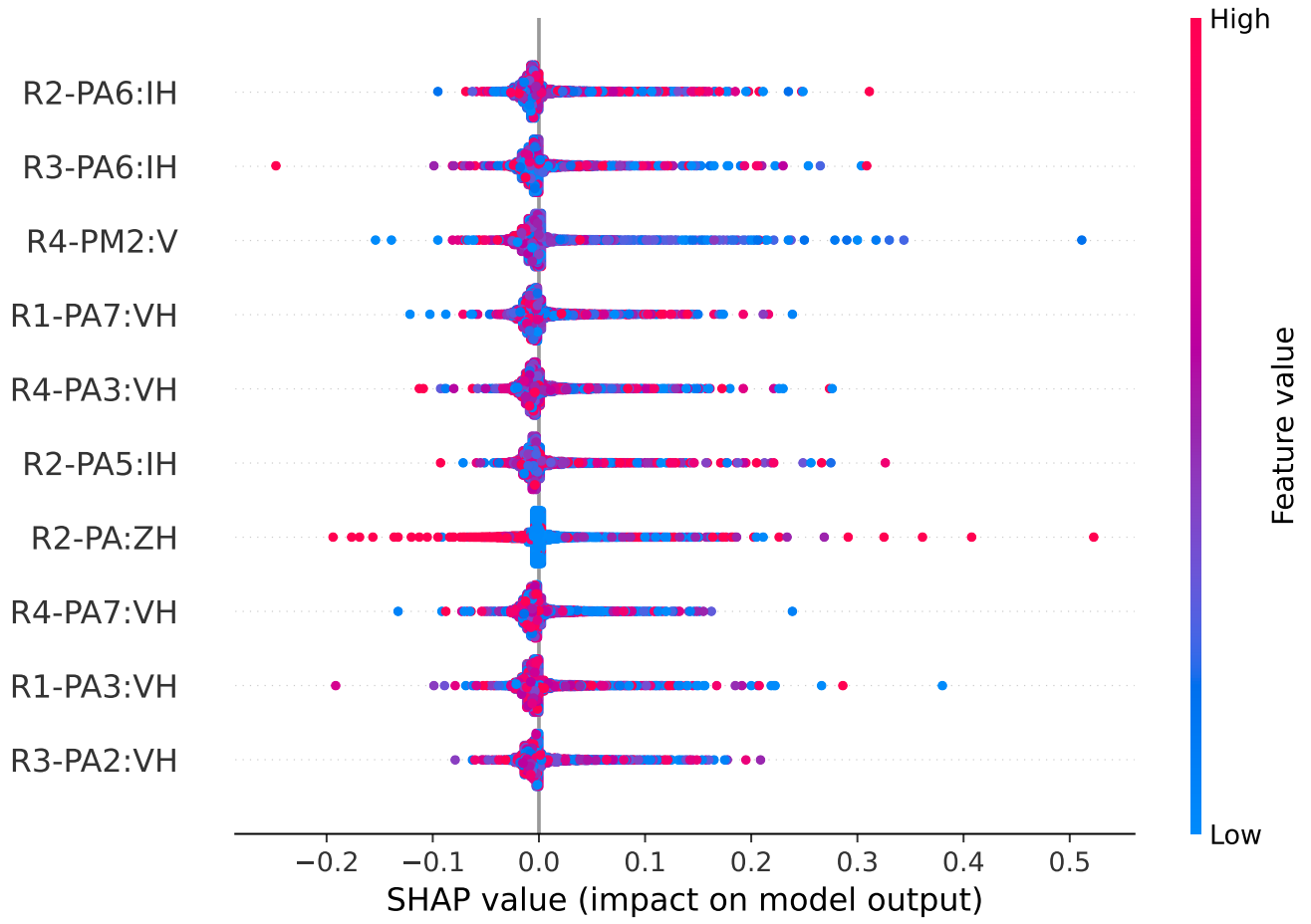

Figure 3. Summary plot of SHAP values for the top features in the three-class classification experiment.

Permutation SHAP Explainability for Multi-Class Datasets

Figure 6 provides a bar plot that captures the top-10 critical features in the performance of the multi-class classification task of distinguishing among 37 power system event scenarios. This presentation depicts the feature importance of each factor concerning its average impact on model prediction. The top features include R4-PM7:V, R2-PM1:V, R4-PA3:VH, R2-PM7:V, R3-PA:ZH, R2-PM3:V, R1-PM2:V, R2-PA:ZH, R2-PA11:IH, and R4-PM2:V; voltage measurements (indicated by “V” and “VH”) primarily constitute these features, while those for current measurements are denoted by “IH” from different PMUs operating from various regions, for instance, R1, R2, R3, R4. Such a prominence of these features is not coincidental. It shows that electrical observables are deeply coupled with the system state, where any minor manipulation would immediately instigate cascading failures if undetected.

An interesting feature is the overwhelming dominance of R4-PM7:V (the magnitude of voltage at PMU 7 in Region 4). PMU 7 is located near a vital generator bus in the simulated test. In power systems, sudden drops or spikes in voltage magnitude at such locations are strongly indicative of either a short-circuit fault or a false data injection attack targeting state estimation. The fact that this feature consistently drives the model toward the attack classifications may indicate that the model has learned to associate localized voltage instability with adversarial intent-a key element in identifying stealthy FDI attacks designed to blend in with natural faults. Similar arguments apply to R2-PM1:V and R2-PM7:V in Region 2, where multiple transmission

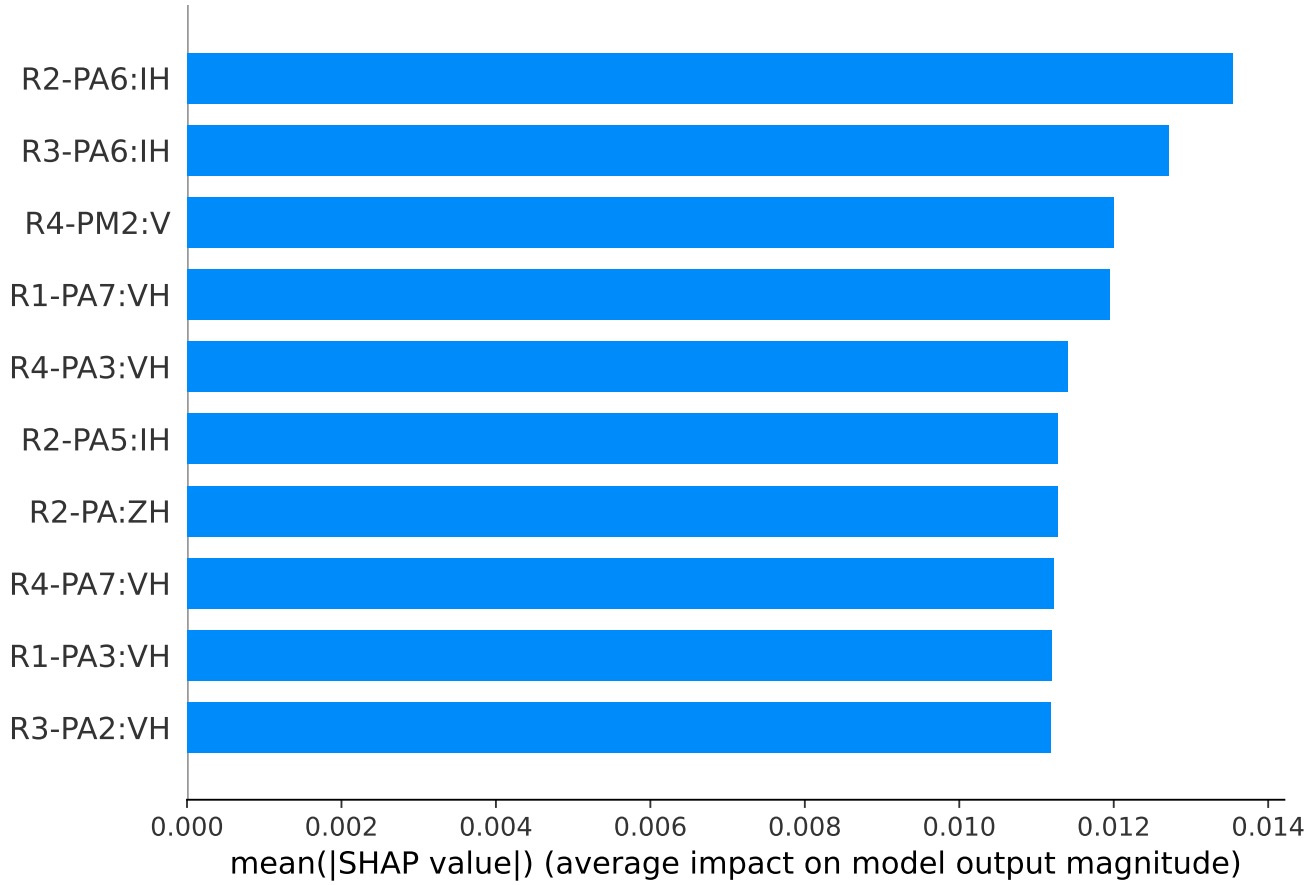

Figure 4. Bar plot summarizing the global importance of the top-10 features for the three-class classification experiment. The height of each bar represents the mean absolute SHAP value of the corresponding feature across all instances, providing a quantitative measure of its contribution to the model's predictions. Features are ranked by global importance, highlighting the most influential variables in distinguishing between attacks, natural events, and no-events.

lines and control relays exist, stronger together to suggest coordinated attacks on interconnected zones exploiting spatial dependencies in the grid.

Furthermore, the inclusion of R2-PA:ZH and R3-PA:ZH indicates zero-sequence currents in Regions 2 and 3, thereby supporting the model in detection efforts for ground-based anomalies, either originating from natural faults (e.g., tree contact) or cyber-induced imbalances (e.g., replay attacks on relay logs). The second-highest importance of R2-PA11:IH-current magnitude at PA11 in Region 2- suggests sensitivity to overcurrent conditions, typically triggered through command injection attacks forcing breakers to trip in an early manner or through load redistribution attacks aimed at overloading specific feeders.

Alongside the beeswarm of SHAP values of the same set of features being complemented with the bar plot, Figure 5 corresponds to a single instance. Each point denotes a specific instance, while the horizontal axis depicts the SHAP value (influence on model output). The vertical axis lists the features. The color scale depicts normalized values of the feature concerned. It enables the simultaneous observation of how distinct ranges of a feature influence predictions. For instance, positive SHAP values of R4-PM7:V, particularly coinciding with high normalized voltage values, strongly indicate the false data injection scenarios targeting generator buses, while negative values cluster around those scenarios with line-to-ground faults. These fine-grained insights would help forensic analysts reconstruct attack vectors post-event or activate automated mitigation protocols based on real-time SHAP attribution.

Those figures 5 and 6 prove the usefulness of Permutation SHAP as a statistical and physical-op rationale in explaining complex machine-learning models. By quantifying and visualizing feature importance in terms of power system dynamics and known attack signatures, we will validate the robustness of our classification framework and ensure its decisions comply with domain expertise. Such transparency is key in high-stakes applications like power grid security, where trust and accountability are a priority. The operators can now interpret the model's outputs not as black-box probabilities but as diagnostic signals

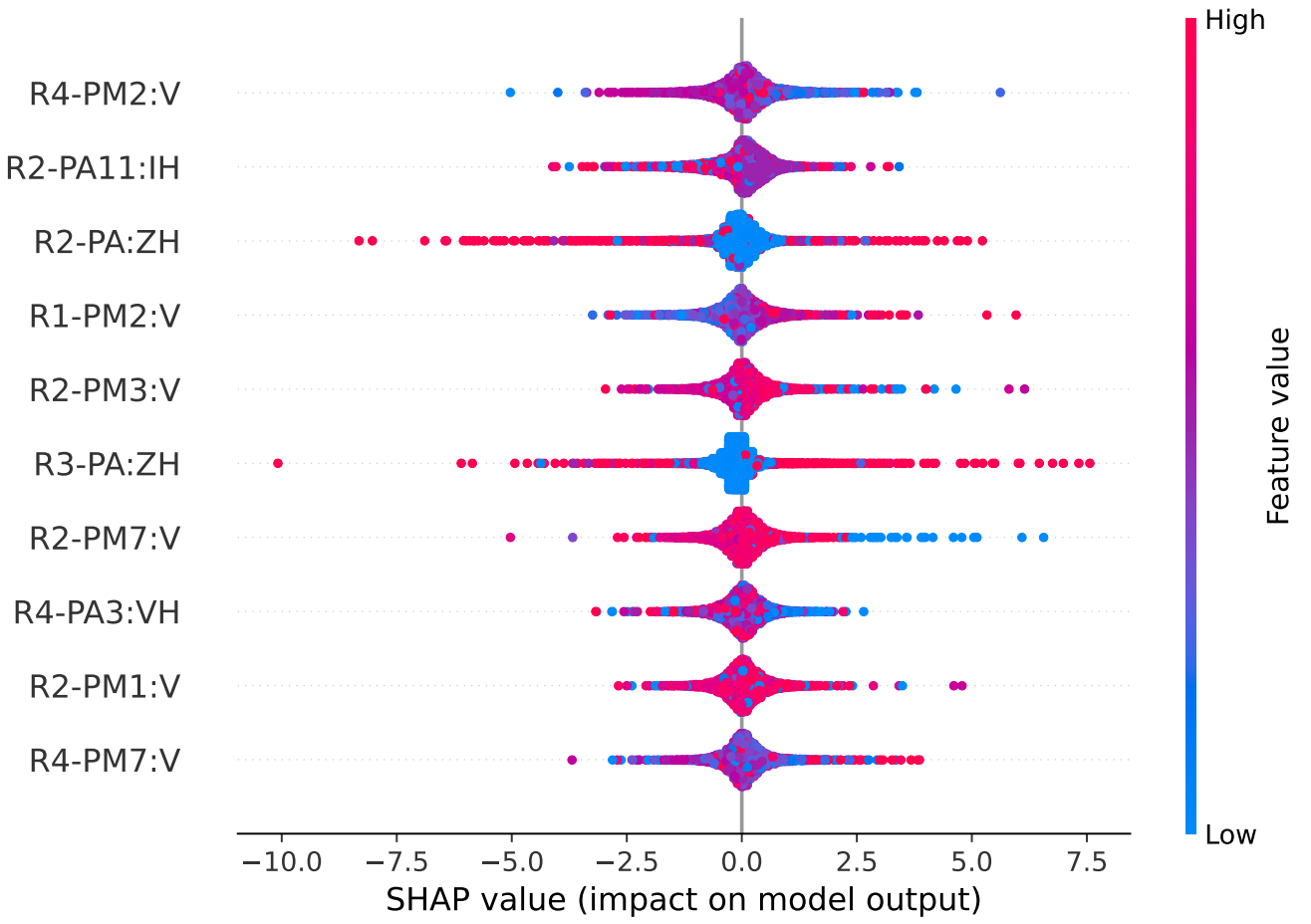

Figure 5. Summary plot of SHAP values for the top features in the multi-class classification task: SHAP values were calculated for the training dataset of only the top-10 features and much more directly correspond to the SHAP values in the annotations. Each point corresponds to a specific instance, with the horizontal axis representing the SHAP value (impact on model output) and the vertical axis listing the features. The color scale reflects the normalized feature value, allowing observation of how specific ranges of a feature influence predictions. Positive SHAP values indicate contributions toward classifying an event as a specific type of attack or disturbance, while negative values suggest contributions toward classifying it as no event or a different scenario.

grounded in measurable electrical phenomena; this enables faster, more confident decision-making during critical incidents.

Computational Efficiency and Real-Time Feasibility

The design of the proposed framework allows scalablility; however, we also need to investigate computational cost (especially hyperparameter tuning (Optuna) and SHAP-based explainability) to validate its suitability for operation. Therefore, we present some preliminary latency measurements and deployment strategies based on experiments performed on our platform running with hardware specifications of an Intel Core i7 CPU with 128 GB RAM, NVIDIA GPU with 6 GB VRAM.

Timing Metrics for Key Stages: We measured the time average per sample for validation of inference efficiency in three key stages:

- **Feature Extraction & Pre-processing**: Approx. **0.5 seconds for 1000 samples**; This includes treatments for missing value imputation, outlier capping using IQR, and feature scaling. Normally, the feature extraction and pre-processing task is carried out only once on data ingestion or initialization of the streaming pipeline and does not add any load to the real-time inference latency.

- **Model Inference**: Full inference with the best classifier (the optimized classifier for the speed-accuracy trade-off) takes **less than 0.1 seconds per sample**. This performance is on par with common SCADA system response requirements

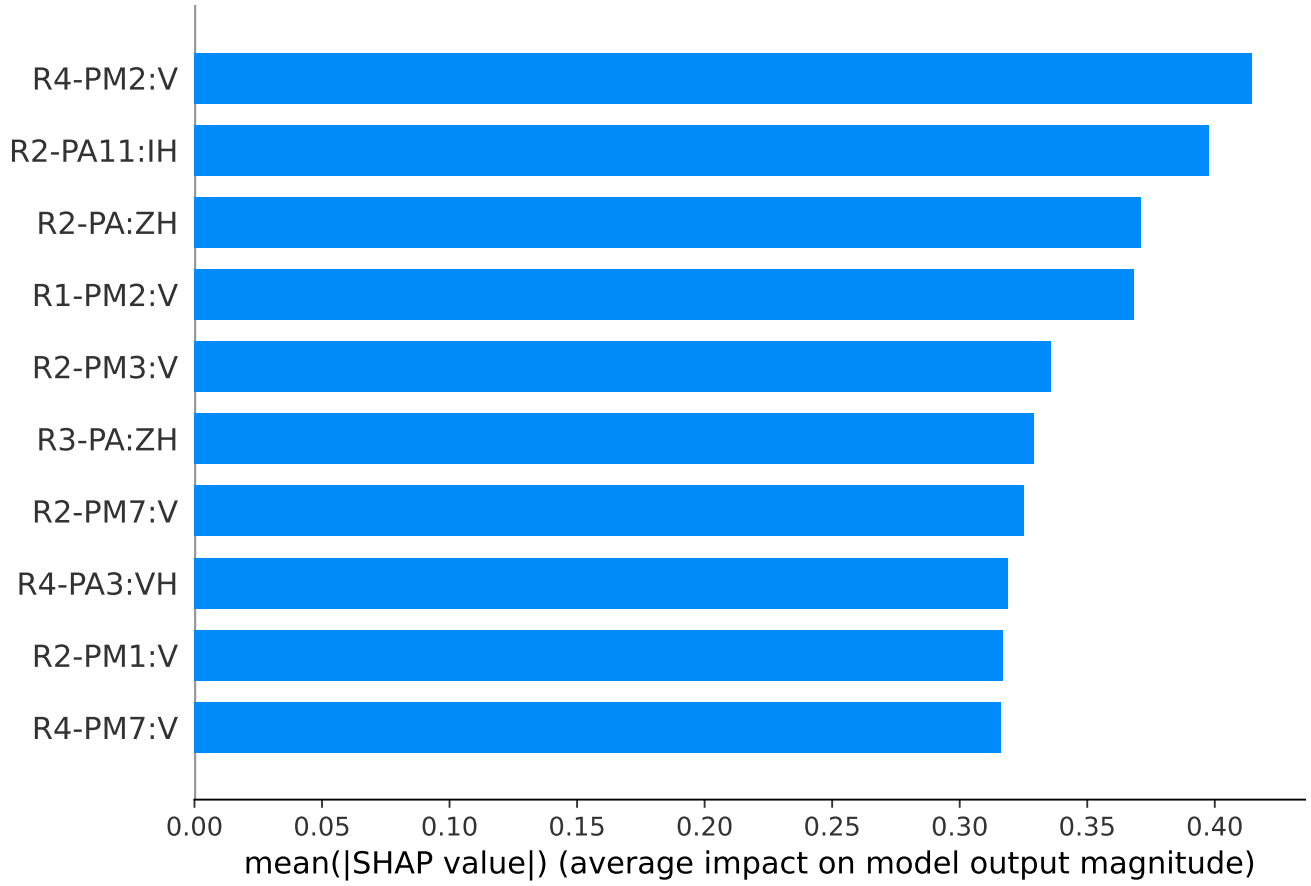

Figure 6. Bar plot summarizing the global importance of the top-10 features for the multi-class classification task: The height of each bar will give an idea about the mean absolute SHAP value of the corresponding feature across all instances, which could be representative of how much the particular feature contributes toward the model's overall predictions. Features are ranked by their global importance when weighed at the global level to separate the 37 different power system event scenarios.

(latency for event classification being less than a second) and avouches deployment at either control center or edge devices [55, 56].

- **Permutation SHAP Explanation**: As a rough estimate, local explanations for the top-10 features take around **2-5 seconds per sample**. Importantly, this step is *not required during real-time operations*; instead, it serves solely for forensic analysis post-event, operator demand dashboards, or model validation during development phases.

Offline vs. Real-Time Operation: We stress that the computationally heavy processes of hyperparameter optimizations (Optuna) and SHAP-based explainability are *offline activities* carried out during the model training, validation, or periodic retraining cycles. After deployment, a final model only conducts lightweight inference operations; this makes it feasible for real-time environments. For example, while the TPE sampler and Median Pruner from Optuna were capitalized on to find an optimal configuration over hundreds of trials, the trials will not be carried out in the live operation; only the ground-truth best-performing model will be deployed.

Deployment Strategies for Edge and Operational Environment: For the real-time feasibility, we suggest two realistic options for deployment architectures:

- **Edge Deployment with Lightweight Models**: For resource-constrained field devices, like RTUs or PMU gateways, we propose the deployment of pruned or quantized versions of Random Forest or Extra Trees models [57, 58].

- **Centralized Explainability Layer**: Reserve SHAP explanation generation for operator dashboards, centralized or post-incident investigation tools. Operators may explicitly request detailed attributions when events are flagged for investigation without interfering with the real-time detection pipeline [59, 60].

With these strategies, the proposed framework remains interpretable and operationally viable. The separation between offline and online model development enables us to ensure high integrity in detection while maintaining stringent latency constraints for power grid applications.

Comparison with Related Studies

The advancement of the suggested methodology over the latest works done so far in intrusion detection systems for power systems is all-time phenomenal. For example, Zaman et al [33] applied recursive feature elimination with random forest for dimensionality reduction upon the validated machine learning-based IDS framework using the ORNL dataset. Regarding classifiers in the study, the RF model trained on the augmented and balanced dataset on unseen test data produced an F1 score of 94.09%. The importance of feature selection and data augmentation has been sufficiently highlighted. However, the performance metrics of this proposed method surpass those of theirs in consistently high accuracy and robustness on binary, three-class, and multi-class classification scenarios.

Similar considerations were presented by Panthi and Das [34], who proposed an adaptive hybrid optimization method to the BGWO-EC model for smart grid intrusion detection. Reported classification accuracies on some datasets reached as high as 98.63%, corroborating the advantages of combining metaheuristic-based feature selection and ensemble learning. However, the new methodology proposed in this work advances the classification performance further through integration with high-level pre-processing techniques, diversified scaling methods, and explainable AI methods such as Permutation SHAP. These aspects assure greater accuracy and offer interpretable insights into the model's decisions, which is a direly needed solution for one major disadvantage of previous frameworks.

Besides that, Naeem et al. [35] carve an even greater niche in the body of knowledge by elaborating on deep-stacked ensembles for intrusion detection oriented toward the ORNL database. In their experimentation, they reported an accuracy of 96.6% using the original feature set and of 99% using the optimized selected reduced feature set. The limitation being, their work focuses more on binary classification, whereas the real challenge is in fine-grained multi-class situations, giving the edge to the concept presented here that classifies 37 different event scenarios, sustaining a performance of above 96% across all tasks. This capability affirms its relevance for real-field applications, which indicates the promise of striking a balance between accurate classification of natural faults and sophisticated cyber-attacks.

In this way, the proposed framework builds upon previous research and extends it through extensive data pre-processing, advanced strategies for feature selection, and explainability methods. The resulting enhancements will fortify classification performance and ensure greater scalability, interpretability, and robustness, thus making the framework well-suited for deployment in the power grid's complex and dynamic environment. Table 5 provides a comparative analysis of the proposed framework against the recent IDS approaches for power systems.

Table 5. Comparative Analysis of the Proposed Framework Against Recent IDS Approaches for Power Systems (2023:2025).

Study	Core Methodology	Key Strengths	Limitations	Real-Time Feasibility	Explainability
This Work	Ensemble Tree Models + Permutation SHAP + Optuna Tuning on Heterogeneous Time-Synchronized Data	High accuracy ($>$ 96% across 37 scenarios), interpretable decisions via SHAP, scalable inference ($<$ 0.1 sec/sample), lightweight deployment options	Requires labeled data; performance on zero-day attacks not evaluated	**High**: Optimized for SCADA latency; SHAP offline	**Strong**: Permutation SHAP provides feature-level transparency aligned with domain physics
Tian et al. (2024) [36]	EVADE: Targeted adversarial FDI attack using saliency maps to bypass BDD/NAD	Demonstrates vulnerability of deep learning models; high stealth rate	Designed to evade detection, not to defend; no mitigation strategy provided	**Low**: Focuses on attack generation, not real-time defense	**None**: Attack-focused, not explainable for operators
Tian et al. (2024) [37]	LESSON: Multi-label adversarial attack framework against DL locators	Highlights multi-label vulnerabilities; realistic perturbations under physical constraints	Exploits model fragility; does not offer defense mechanisms	**Low**: Adversarial generation is computationally heavy	**None**: Attack-centric, not operator-facing
Nandanwar & Katarya (2024) [10]	GAO-Xgboost + ECC-Integrated Blockchain for IoT Security	Combines optimization (GAO) with cryptography for secure detection	Complex architecture; ECC adds encryption/decryption latency	**Medium**: XGBoost fast, but blockchain layer slows inference	**Limited**: Explainability not prioritized; focus on data security
Zaman et al. (2023) [33]	RFE-RF + Synthetic Data Augmentation on ORNL Dataset	Good generalization; validated on binary classification	Limited to binary output; no multi-class or explainability support	**High**: RF inference fast	**None**: No SHAP or similar explanation method integrated
Panthi & Das (2022) [34]	BGWO-EC: Binary Grey Wolf Optimization + Ensemble Classifier	High accuracy (up to 98.63%) on specific datasets	Computationally expensive tuning; no real-time evaluation	**Medium**: Metaheuristic tuning slow; inference fast	**None**: Black-box ensemble; no feature attribution
Naeem et al. (2025) [35]	Deep Stacked Ensemble + GWO Feature Selection	High accuracy (up to 99%); good generalization on unseen data	Binary classification only; lacks fine-grained scenario coverage	**Medium**: Deep stacking may introduce latency	**Weak**: No built-in explainability; relies on post-hoc analysis

Relevance of the Study

In modern power grids, threats are increasing due to the integration of cyber-physical systems that keep them vulnerable to natural disturbances and subtle cyber-attacks. Good operation of power system-controlled very-larger interconnection networks

must categorize these events. This is where this study stands in the originality of walking: it presents a unified method to view a physical attack and cyber threat as one and help classify them with high accuracy, even when the latter is made indistinguishable to the (natural) faults.

The proposed work has high relevance, consumingly instant for the welfare of situational awareness towards providing defense mechanisms in a cradle of intelligence to protect the next-gen grid. Machine-learning officials especially suggested for binary three-class and multi-class classification scenarios based on a strongly rooted concept of heterogeneous time-synchronized machine data with PMUs, control panel logs, Snort alerts, and protective relay logs. In such a context, this aspect is very class-oriented and plays an indirect role in preventing a downtime system from entering a zone of cascading functional damage where there have been incidents half understood in the past.

In addition, the present study indeed pushes the frontier of intrusion detection methodologies for power systems. However, whereas earlier developments involved sequential pattern mining and probabilistic approaches, a gradual emergence of machine-learning-based and ensemble system models is noted in current practice. Therefore, this paper builds further on these models to exploit feature engineering, synthetic data augmentation, and metaheuristic optimization to increase model generalization and accuracy. The incorporation of Permutation SHAP in the paper creates dependence on high levels of explainable AI processes, where the former order fosters trust and accountability simply by endorsing the decisions made by the models and aligning opinion with domain knowledge.

The output of the study could be very useful in the near future for almost real-time monitoring and an automated response application to better handle the operators' safety against lurking threats. The groundwork done in the current research for classifying complex power system events via a potentially uncomplicated system serves as a prelude to resilient and secure power systems that can face ever-evolving cyber-physical threats.

Limitations

Despite demonstrating robust performance in classifying power system disturbances and cyber-attacks, the proposed methodology has limitations. One primary limitation concerns the dependence on high-quality, time-synchronized data from PMUs, control logs, and intrusion detection systems. The actual conditions concerning data availability and quality may vary considerably in the field regarding sensor coverage, communication delays, or adversarial tampering. Furthermore, it is assumed that labels in the training and evaluation datasets represent all possible attack vectors and fault conditions; this seems reasonable for benchmarking but may not hold true in cases of novel, zero-day, or highly adaptive threats with which operational grids must contend.

More specifically, the actual performance of our framework on attack types that were never previously seen (e.g., ones built to avoid detection by extant decision logic or exploit previously-unmodeled system configurations) has not been assessed. It is also an open question how it behaves in the presence of dataset shift (changes in load profiles, topology, sensor calibration, and so on across time). While our models generalize reasonably well across the 37 defined scenarios, they may encounter extreme difficulty in the case of events that fall outside this labeled space, a glaring weakness within a smart grid type of environment that is so dynamic and evolving.

To alleviate this, future work will include unsupervised anomaly detection modules (e.g., autoencoders, isolation forests, or one-class SVMs) in combination with our supervised classifier. Such a hybrid architecture would flag statistically anomalous patterns for human review or escalation, even if those patterns did not match any known label, thereby extending the framework's coverage into new and emerging threat landscapes. Another limitation is understanding the computational complexities concerning feature selection and hyper-parameter tuning in applying metaheuristic optimization methods. On the one hand, these methods enhance the model performance. On the other hand, with a longer training duration and resource consumption, they create high complexity that can be used in a large-scale operation. While the SHAP explainability analysis provided helpful insights, it was computationally intensive for high-dimensional datasets, so it became challenging to apply interpretability during implementation in real-time.

Finally, the framework's performance highly depends on the granularity of the event labels provided for training. Cases of possible misclassifications arise where some attack types and creep natural faults have patterns that are so closely related that they sometimes become indistinguishable, especially when the attack is aimed at making the intrusion look like something actually benign. Addressing those limitations will require advanced research in adaptive learning techniques, anomaly detection, and unsupervised methods that deal with unknown or evolving threat contexts.

Conclusions and Future Directions

This research presents a fully explainable framework for classifying all disturbances and cyber-attacks happening to a power grid using heterogeneous time-synchronized data. Such a framework relies on further incorporation of advanced preprocessing, multi-strategy feature selection, hyperparameter-optimized machine-learning models, and Permutation SHAP explainability for

all binary-, three-class-, and fine-grained multi-class classification scenarios at high accuracy and robustness. Individual case datasets report above 96% coverage, while combined heterogeneous data showcase above 93%. Now it can be inferred that strategically placed PMUs have voltage phase angles, current magnitudes, and zero-sequence currents as the discriminators between natural and malicious actions in agreement with domain knowledge for increasing operator trust by transparent decision-making. Among other objectives, to solve the integration of physical and cyber event detection into a common architecture, this type of framework will also fill the gap in research that has substantial practical significance for situational awareness, automated response, and resilience building for modern smart grids.

Future research will focus on furthering this capacity of the proposed framework as much as possible toward the evolution of threats facing us. Also, the development of a hybrid architecture using both supervised and unsupervised methods for zero-day threat detection is another research direction. It includes, for instance, autoencoders or isolation forests-anomaly detection modules- creating a flag for statistically deviant patterns- in the cases where, even in instances, no labeled classes exist to be analyzed by human review or elicited. Incremental learning techniques will then be used to get the attack data labeled during operations attached to periodic retraining of the model; hence, if such data are made available, the model will keep adapting to the evolving tactics employed by adversaries. Other investigations would also investigate the robustness of the model against dataset shifts (changes in loading profiles, topology, or sensor calibration) and some computational efficiency optimizations for real-time deployment onto edge devices. Lastly, forming laws of nature-like Kirchhoff's laws and relay coordination logic within the modeling pipeline would increase generalization while reducing false positives, ensuring effectiveness even against the high in-memory stealth factors in cyber-physical attacks.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Data availability

The dataset utilized in this study was developed collaboratively by researchers at Mississippi State University and ORNL. The dataset is avaialble at: https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets

Funding

Not applicable.

Author Contribution

Conceptualization, M.F. and M.A.; Data curation, S.A.A. and M.A.; Formal analysis, M.M.A. and A.I.S.; Investigation, S.A.A. and A.I.S.; Methodology, M.F., H.M.B., and M.A.E.; Software, M.F, M.M.A. , H.M.B., and A.I.S.; Validation, M.M.A. and S.A.A.; Visualization, S.A.A., M.A., and M.B.; Writing—review and editing, M.F., M.A.E., and M.B.; Supervision, M.A.E.

Ethics approval and consent to participate

Not applicable.

References

1. Abdelkader, S. *et al.* Securing modern power systems: Implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks. *Results engineering* 102647 (2024).

2. Latvakoski, J., Mäki, K., Ronkainen, J., Julku, J. & Koivusaari, J. Simulation-based approach for studying the balancing of local smart grids with electric vehicle batteries. *Systems* **3**, 81–108 (2015).

3. Nafees, M. N., Saxena, N., Cardenas, A., Grijalva, S. & Burnap, P. Smart grid cyber-physical situational awareness of complex operational technology attacks: A review. *ACM Comput. Surv.* **55**, 1–36 (2023).

4. Stellios, I., Kotzanikolaou, P., Psarakis, M., Alcaraz, C. & Lopez, J. A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services. *IEEE Commun. Surv. Tutorials* **20**, 3453–3495 (2018).

5. Nandanwar, H. & Katarya, R. Securing industry 5.0: An explainable deep learning model for intrusion detection in cyber-physical systems. *Comput. Electr. Eng.* **123**, 110161 (2025).

6. Illiano, V. P. & Lupu, E. C. Detecting malicious data injections in wireless sensor networks: A survey. *ACM Comput. Surv. (CSUR)* **48**, 1–33 (2015).

7. Nandanwar, H. & Katarya, R. Privacy-preserving data sharing in blockchain-enabled iot healthcare management system. *The Comput. J.* bxaf065 (2025).

8. Xing, W. & Shen, J. Security control of cyber–physical systems under cyber attacks: A survey. *Sensors* **24**, 3815 (2024).

9. Duo, W., Zhou, M. & Abusorrah, A. A survey of cyber attacks on cyber physical systems: Recent advances and challenges. *IEEE/CAA J. Autom. Sinica* **9**, 784–800 (2022).

10. Nandanwar, H. & Katarya, R. Optimized intrusion detection and secure data management in iot networks using gao-xgboost and ecc-integrated blockchain framework. *Knowl. Inf. Syst.* 1–56 (2025).

11. Pan, S., Morris, T. & Adhikari, U. Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data. *IEEE Transactions on Ind. Informatics* **11**, 650–662 (2015).

12. Nandanwar, H. & Katarya, R. A secure and privacy-preserving ids for iot networks using hybrid blockchain and federated learning. In *International Conference on Next-Generation Communication and Computing*, 207–219 (Springer, 2024).

13. Timusk, M., Lipsett, M. & Mechefske, C. K. Fault detection using transient machine signals. *Mech. Syst. Signal Process.* **22**, 1724–1749 (2008).

14. Xu, X. & Karney, B. An overview of transient fault detection techniques. *Model. monitoring pipelines networks: Adv. tools for automatic monitoring supervision pipelines* 13–37 (2017).

15. Deng, R., Xiao, G., Lu, R., Liang, H. & Vasilakos, A. V. False data injection on state estimation in power systems—attacks, impacts, and defense: A survey. *IEEE Transactions on Ind. Informatics* **13**, 411–423 (2016).

16. Chakrabarty, S. & Sikdar, B. Detection of malicious command injection attacks on phase shifter control in power systems. *IEEE Transactions on Power Syst.* **36**, 271–280 (2020).

17. Ramanan, P., Li, D. & Gebraeel, N. Blockchain-based decentralized replay attack detection for large-scale power systems. *IEEE Transactions on Syst. Man, Cybern. Syst.* **52**, 4727–4739 (2021).

18. Abdi, N. M. *Deep Reinforcement Learning Based Moving Target Defense for Mitigating False Data Injection Attacks in Power Grids*. Master's thesis, Hamad Bin Khalifa University (Qatar) (2024).

19. Alserhani, F. & Aljared, A. Evaluating ensemble learning mechanisms for predicting advanced cyber attacks. *Appl. Sci.* **13**, 13310 (2023).

20. Aljabri, M. *et al.* Intelligent techniques for detecting network attacks: review and research directions. *Sensors* **21**, 7070 (2021).

21. Nandanwar, H. & Katarya, R. A hybrid blockchain-based framework for securing intrusion detection systems in internet of things. *Clust. Comput.* **28**, 471 (2025).

22. Negi, M. Towards the integration of it/ot technologies in electricity based digitalized energy systems. *Univ. VAASA* (2024).

23. Mchirgui, N., Quadar, N., Kraiem, H. & Lakhssassi, A. The applications and challenges of digital twin technology in smart grids: a comprehensive review. *Appl. Sci.* **14**, 10933 (2024).

24. Ma, R., Chen, H.-H., Huang, Y.-R. & Meng, W. Smart grid communication: Its challenges and opportunities. *IEEE transactions on Smart Grid* **4**, 36–46 (2013).

25. Kumar, P. *et al.* Smart grid metering networks: A survey on security, privacy and open research issues. *IEEE Commun. Surv. Tutorials* **21**, 2886–2927 (2019).

26. Nandanwar, H. & Katarya, R. Tl-bilstm iot: transfer learning model for prediction of intrusion detection system in iot environment. *Int. J. Inf. Secur.* **23**, 1251–1277 (2024).

27. Bekara, C. Security issues and challenges for the iot-based smart grid. *Procedia Comput. Sci.* **34**, 532–537 (2014).

28. Dalipi, F. & Yayilgan, S. Y. Security and privacy considerations for iot application on smart grids: Survey and research challenges. In *2016 IEEE 4th international conference on future internet of things and cloud workshops (FiCloudW)*, 63–68 (IEEE, 2016).

29. Ankitdeshpandey & Karthi, R. Development of intrusion detection system using deep learning for classifying attacks in power systems. In *Soft Computing: Theories and Applications: Proceedings of SoCTA 2019*, 755–766 (Springer, 2020).

30. Hink, R. C. B. *et al.* Machine learning for power system disturbance and cyber-attack discrimination. In *2014 7th International symposium on resilient control systems (ISRCS)*, 1–8 (IEEE, 2014).

31. Pan, S., Morris, T. & Adhikari, U. Developing a hybrid intrusion detection system using data mining for power systems. *IEEE Transactions on Smart Grid* **6**, 3104–3113 (2015).

32. Pan, S., Morris, T. H. & Adhikari, U. A specification-based intrusion detection framework for cyber-physical environment in electric power system. *Int. J. Netw. Secur.* **17**, 174–188 (2015).

33. Zaman, M., Upadhyay, D. & Lung, C.-H. Validation of a machine learning-based ids design framework using ornl datasets for power system with scada. *IEEE Access* **11**, 118414–118426 (2023).

34. Panthi, M. & Das, T. K. Intelligent intrusion detection scheme for smart power-grid using optimized ensemble learning on selected features. *Int. J. Critical Infrastructure Prot.* **39**, 100567 (2022).

35. Naeem, H., Ullah, F. & Srivastava, G. Classification of intrusion cyber-attacks in smart power grids using deep ensemble learning with metaheuristic-based optimization. *Expert. Syst.* **42**, e13556 (2025).

36. Tian, J. *et al.* Evade: targeted adversarial false data injection attacks for state estimation in smart grid. *IEEE Transactions on Sustain. Comput.* (2024).

37. Tian, J. *et al.* Lesson: Multi-label adversarial false data injection attack for deep learning locational detection. *IEEE Transactions on Dependable Secur. Comput.* **21**, 4418–4432 (2024).

38. Tian, J. *et al.* Joint adversarial example and false data injection attacks for state estimation in power systems. *IEEE Transactions on Cybern.* **52**, 13699–13713 (2021).

39. Jia, W., Sun, M., Lian, J. & Hou, S. Feature dimensionality reduction: a review. *Complex Intell. Syst.* **8**, 2663–2693 (2022).

40. Hopf, K. & Reifenrath, S. Filter methods for feature selection in supervised machine learning applications–review and benchmark. *arXiv preprint arXiv:2111.12140* (2021).

41. Learning, U. M. & Reduction, D. Principal component analysis. *PCA–a primer, Employing PCA, Introd. k* (2023).

42. Jeon, H. & Oh, S. Hybrid-recursive feature elimination for efficient feature selection. *Appl. Sci.* **10**, 3211 (2020).

43. Leiva-Murillo, J. M. & Artes-Rodriguez, A. Maximization of mutual information for supervised linear feature extraction. *IEEE Transactions on Neural Networks* **18**, 1433–1441 (2007).

44. Omar, E. D. *et al.* Comparative analysis of logistic regression, gradient boosted trees, svm, and random forest algorithms for prediction of acute kidney injury requiring dialysis after cardiac surgery. *Int. J. Nephrol. Renovascular Dis.* 197–204 (2024).

45. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, 2623–2631 (2019).

46. Watanabe, S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127* (2023).

47. Zhan, D. & Xing, H. Expected improvement for expensive optimization: a review. *J. Glob. Optim.* **78**, 507–544 (2020).

48. Hassanali, M., Soltanaghaei, M., Javdani Gandomani, T. & Zamani Boroujeni, F. Software development effort estimation using boosting algorithms and automatic tuning of hyperparameters with optuna. *J. Software: Evol. Process.* **36**, e2665 (2024).

49. Huber, K. T., Moulton, V., Lockhart, P. & Dress, A. Pruned median networks: a technique for reducing the complexity of median networks. *Mol. phylogenetics evolution* **19**, 302–310 (2001).

50. He, Y. & Xiao, L. Structured pruning for deep convolutional neural networks: A survey. *IEEE transactions on pattern analysis machine intelligence* **46**, 2900–2919 (2023).

51. Vujović, Ž. *et al.* Classification model evaluation metrics. *Int. J. Adv. Comput. Sci. Appl.* **12**, 599–606 (2021).

52. Barratt, S. & Sharma, R. Optimizing for generalization in machine learning with cross-validation gradients. *arXiv preprint arXiv:1805.07072* (2018).

53. Sathyanarayanan, S. & Tantri, B. R. Confusion matrix-based performance evaluation metrics. *Afr. J. Biomed. Res.* 4023–4031 (2024).

54. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. neural information processing systems* **30** (2017).

55. Enemosah, A. & Ifeanyi, O. G. Scada in the era of iot: automation, cloud-driven security, and machine learning applications. *Int. J. Sci. Res. Arch.* **13**, 3417–3435 (2024).

56. Šenk, I., Tegeltija, S. & Tarjan, L. Machine learning in modern scada systems: Opportunities and challenges. In *2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–5 (IEEE, 2024).

57. Kumar, R. & Sharma, A. Edge ai: A review of machine learning models for resource-constrained devices. *Artif. Intell. Mach. Learn. Rev.* **5**, 1–11 (2024).

58. Ngo, D., Park, H.-C. & Kang, B. Edge intelligence: A review of deep neural network inference in resource-limited environments. *Electronics* **14**, 2495 (2025).

59. Smith, J. Explainable ai for threat intelligence and incident response. *Available at SSRN 5140447* (2020).

60. Asaye, L. *et al.* Predicting and understanding emergency shutdown durations level of pipeline incidents using machine learning models and explainable ai. *Processes* **13**, 445 (2025).