

Article

Smart-Grid Cyber-Attack and Disturbance Detection via Interpretable and Accurate Genetic-Fuzzy Data-Mining/Machine-Learning Approach

Marian B. Gorzalczyński[†]  and Filip Rudziński^{*,†} 

Department of Electrical and Computer Engineering, Kielce University of Technology, Al. 1000-Lecia P.P. 7, 25-314 Kielce, Poland; m.b.gorzalczyński@tu.kielce.pl

* Correspondence: f.rudzinski@tu.kielce.pl

[†] These authors contributed equally to this work.

Abstract

This paper makes a contribution to smart-grid cybersecurity by proposing an application of our approach, previously published in this journal, to design—from smart-grid data—transparent, interpretable, and accurate systems for detection and classification of cyber-attacks, natural-event-related disturbances, and regular operation of the smart grid. The collection of 15 open-source simulated smart-grid data sets provided by Mississippi State University, USA, in collaboration with Oak Ridge National Laboratories, USA, is used in our experiments. To address this problem, we use our knowledge-based data-mining/machine-learning approach which ultimately generates a set of fuzzy rule-based classifiers—each with a different trade-off between its accuracy and the interpretability of its knowledge base (both are the subjects of maximization in a multi-objective optimization process using evolutionary algorithms). The paper also contributes an extensive cross-validation-based experiment showing that while our approach is not as accurate as alternative and inherently accuracy-oriented “black boxes,” it surpasses them in terms of transparency and interpretability of the generated classification decisions, while ensuring acceptable accuracy of these decisions. Therefore, it can act as a second cybersecurity layer that uses transparent rules to help SCADA operators instantly verify and explain alerts triggered by accurate, fast “black-box” detectors. This paper also provides a contribution to a more general and important area of the automatic and effective selection of input variables (features) that are essential in a given decision-making problem (such as smart-grid contingency classification considered in this work).

Keywords: smart-grid cyber-attacks and disturbances; data mining; machine learning; interpretable and accurate classifiers; fuzzy rule-based classifiers; multi-objective evolutionary optimization



Academic Editors: Giorgio Graditi, Amedeo Buonanno, Maria Valenti and Salvatore Fabozzi

Received: 20 May 2026

Revised: 10 July 2026

Accepted: 14 July 2026

Published: 17 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The smart grid—often referred to as the next-generation power system [1,2]—employs communication and information technologies in the process of generation, delivery, and consumption of electric power. Whereas the components of a traditional power system are physically isolated, those of the smart grid are interconnected by means of cyber-information infrastructure that creates a widely distributed and automated cyber-physical system. Such a system is equipped with new functionalities including real-time control, grid resilience, operational efficiency, and much better integration of renewable energy

technologies due to a gradual shift away from fossil fuel-based energy generation. On the other hand, however, expanding the physical energy infrastructure with a digital network exposes the smart grid to cyber-attacks, i.e., threats similar to those faced by traditional computer systems and network faces. Since the electric power grid plays an essential role in our socio-economic system, various recommendations and guidelines aiming at improving the cybersecurity of the power system have been provided by different organizations (see, e.g., [3,4]). However, according to the United States Government Accountability Office, the existing recommendations and guidelines are still not sufficient for secure implementation of the smart grid, and further research and development in this area should be carried out [5]. Accordingly, in this paper, we apply the approach developed by us that is based on genetically optimized fuzzy rule-based classifiers (FRBCs) to the interpretable and accurate detection and classification of different types of cyber-attacks (intrusions) and disturbances in the smart grid. Our approach belongs to the research areas of data mining (DM) and machine learning (ML)—see further comments later in this section.

Cyber-attacks targeting the smart grid usually exploit vulnerabilities in communication links and control devices to affect the measurement and control signals as well as monitoring systems [6–8]. It is essential, however, to distinguish between cyber-attacks and disturbances related to natural events. Incorrect classification of a cyber-attack as a disturbance can result in a wrong response with some negative impact on the power system. On the other hand, wrong classification of a disturbance as a cyber-attack can lead to wrong response within the communication and information system [9]. Moreover, various automated event response systems (e.g., wide-area protection systems, system integrity protection schemes [10,11], etc.) are based on a single classification system that can identify all types of smart-grid contingencies. Our approach effectively addresses all these issues.

Synchrophasor measurement systems and audit logs (chronological, tamper-evident records tracking events like logins, configuration changes, and data access) from multiple system devices generate huge amounts of data that describe the behavior of the smart grid including its regular operation, natural-event-related disturbances, and cyber-attacks. For these reasons, DM and ML techniques are particularly well suited for designing systems that are able to detect and to distinguish different types of smart-grid operation. It is worth stressing that there is a considerable gap between DM and ML terms, techniques, and application areas [12]. The “older” one, i.e., ML, was defined initially by A. L. Samuel in 1959 as a “field of study that gives computers the ability to learn without being explicitly programmed” [13]. The “younger” one, i.e., DM—introduced in the 1990s—includes methods that concentrate—through learning from data (i.e., similarly to ML methods)—on the automatic discovery of helpful and understandable patterns, decision-making mechanisms, and relationships in the data under consideration. Although various DM/ML approaches have been deployed in smart-grid research—as we briefly reviewed in the next section of the paper—their primary limitation lies in their opaque, black-box nature. By prioritizing accuracy over interpretability, these models fail to explain their decisions or offer meaningful insights into the underlying mechanisms governing the smart-grid cyber-attacks and disturbances. In turn, the few interpretable (to some extent) DM/ML approaches are characterized by poor decision justification semantics, as they treat decision variables referring to the same physical domains as independent rather than related. This work addresses that problem by employing our earlier-mentioned modern, knowledge-based DM/ML approach (i.e., FRBCs with genetically optimized compromises between accuracy and interpretability—see, e.g., our work [14] for details) for the identification of different types of smart-grid contingencies.

The main goal and contribution of this paper is the application of our DM/ML methodology to design interpretable and accurate classification systems from the collection of 15 open-source simulated smart-grid data sets describing various types of cyber-attacks, natural-event-related disturbances, and regular operation of the smart grid. These data sets are provided by Mississippi State University, USA, in collaboration with Oak Ridge National Laboratories, USA [15,16].

In our DM/ML methodology, we introduce separate measures of FRBC's interpretability and accuracy using them also as performance metrics and optimization goals in the process of designing FRBCs from the smart-grid data. The complementary/contradictory nature of a system's interpretability and accuracy fully justifies the use of multi-objective evolutionary optimization algorithms (MOEOAs) in our approach. In such a way, the classification-system-design problem becomes the FRBC's structure- and parameter-optimization task equivalent to the FRBC's interpretability-accuracy trade-off optimization task (related work is reviewed in [17]). Thus, we employ the our generalization of the well-known MOEOA, i.e., the strength Pareto evolutionary algorithm (SPEA2) [18]; that generalization—referred to as SPEA3—is characterized by better performance than SPEA2 (see [19,20], which demonstrates the superiority of SPEA3 over SPEA2).

A major algorithmic advancement introduced in this work (compared with our previous works) is the concept of shared fuzzy sets across groups of input attributes with uniform physical domains (e.g., a group of attributes representing currents, voltages, and phase angles measured at different nodes of a smart grid). Typical techniques of fuzzy rule generation from data treat all attributes independently (by defining different membership functions and even linguistic labels of the corresponding fuzzy sets), leading to semantic inconsistency and a massive search space. To address this gap, we incorporate a small amount of domain-specific knowledge into our previously published decision-making mechanism by forcing all input attributes representing the same physical entity to share exactly the same fuzzy sets, regardless of the measurement node. Far from being a simple heuristic adjustment, this strict domain-driven structural constraint alters the classifier's decision logic—the generated rules strictly obey the physical semantics of power systems.

The remainder of this paper is organized as follows. First, we review related work on DM/ML applications for cyber-attack and disturbance classification in smart grids. Next, we outline our proposed DM/ML methodology and characterize the 15 smart-grid data sets used in our experiments. Finally, we present and discuss the main contribution of this work—the design of accurate and interpretable classification systems for the smart-grid multi-class data, alongside a comprehensive comparative analysis against 31 alternative methods.

2. Related Works

Various DM/ML methods have been applied to the detection and classification of smart-grid faults, disturbances, and cyber-attacks. Among the often used approaches are different types of artificial neural networks (ANNs). Several ANN-based solutions are applied to smart-grid fault classification. In [21], the current phase is decomposed using wavelet transforms and fed into a multi-layer perceptron ANN trained by means of particle swarm optimization for fault classification. The fault classification performance of a Chebyshev ANN is evaluated on current signals in [22]. Moreover, an ANN integrated with a wavelet transform technique is applied in [23] to extract power-system fault patterns using energy variation of fault signals.

As far as power-quality disturbance classification is concerned, decision tree (DT) classifiers are employed in [24,25], and a support vector machine (SVM) classifier is also used in [24]. Methods based on k -nearest neighbor (k -NN), probabilistic naive Bayes (NB),

k-means clustering (*k*-MC), and Hoeffding tree (HT) algorithms are used in [26,27] for power-system fault and disturbance classification. However, those methods are designed to process power-system measurement data only. They do not include any other system information (e.g., control system log information), which limits their classification precision.

Several DM/ML algorithms, including a perceptron-type ANN, *k*-NN, SVMs, and sparse logistic regression (LR) with ensemble learning and feature-level fusion are employed in [28] for classifying and predicting a common class of smart-grid cyber-attacks, namely, false data injection (FDI) attacks. The results of experiments carried out in [28] demonstrate that DM/ML methods outperform traditional state-estimation-based approaches to FDI-attack detection. State estimation employs the power-system measurements to estimate the system state and compare the difference between the observed measurements and the estimated ones with a threshold for cyber-attack detection [28]. The application of SVMs, *k*-NN, and extended nearest neighbor (ENN) methods for detecting both direct and stealth FDI attacks in smart-grid systems is presented in [29]. A DT-based anomaly detection methodology using power-system measurements to distinguish the normal tripping coming from power line faults and malicious tripping attacks on physical relays is proposed in [30]. An adaptive-boosting (AdaBoost) meta-algorithm using the random forest (RF) as the base classifier operating on power-system measurement data is considered in [31] for detecting power-system disturbances and cyber-attacks. In turn, RF supported by an improved grid search (IGS) algorithm, i.e., the RF+IGS approach is proposed in [32] for power-system intrusion detection. Gradient boosting DT (GBDT) and extreme gradient boosting (XGBoost) approaches are applied in [33] to build a smart-grid cyber-attack detection model. A number of DM/ML methods—including RF, SVM, rule induction techniques, such as OneR, NNge, and JRipper, as well as probabilistic naive Bayes (NB) method and earlier-mentioned AdaBoost meta-algorithm—are used in [16] to discriminate power-system disturbances and cyber-attacks.

In recent literature, eXplainable AI (XAI) tools are becoming prominent, such as the SHAP (SHapley Additive exPlanations, e.g., [34–36]) framework or the LIME (decision tree-based Local Interpretable Model agnostic Explanations, e.g., [37]) framework. Both frameworks attempt to explain the predictions of ML models. SHAP uses cooperative game theory to assign a numerical contribution score (“Shapley value”) to every feature, revealing precisely how much a given feature influenced the model’s output. LIME creates plausible explanations that capture nonlinear interactions between data features, resulting in more reliable interpretability, fidelity, and efficiency. Although SHAP and LIME provide excellent local explanations for complex “black-box” models, they inherently rely on the underlying model’s correctness. If a deep neural network suffers from adversarial manipulation or fails to detect a stealthy FDIA, the frameworks will merely explain a flawed decision. Our approach focuses on an intrinsically interpretable architecture where the decision logic is fully transparent from the beginning, minimizing operational risks. Second, traditional symbolic rule learners (e.g., JRip, OneR) and decision trees generate understandable rules, but they rely on “crisp” (rigid) thresholds (e.g., IF Current > 150 A). Smart-grid telemetry is continuous, highly dynamic, and prone to measurement noise. Rigid thresholds often lead to high false-positive rates or overly sensitive classifiers. Our proposed framework utilizes fuzzy logic, which allows for smooth linguistic transitions (e.g., “low”, “medium”, “high”) that match human engineering reasoning and accommodate grid volatility far better than standard crisp rule learners. Furthermore, decision trees and random forests typically generate many rules (even hundreds of rules) that, taken as a whole, are very difficult to interpret.

Compared with the relatively large number of supervised-learning-based DM/ML methods, a small number of semi-supervised anomaly detection approaches are used for

smart-grid cyber-attack detection. One-class SVM (OCSVM) in [38] as well as OCSVM combined with k -means recursive clustering (k -MRC) in [39] are employed to detect malicious attacks in the supervisory control and data acquisition (SCADA) systems. Several semi-supervised anomaly detection algorithms combined with deep representation learning are used in [40] for detecting cyber-attacks in smart grids. Broad reviews of cybersecurity problems in the smart grid can be found in [41–44].

To the best of our knowledge, the existing studies on the smart-grid cyber-attack and disturbance detection and classification present approaches oriented only toward accuracy. Even studies applying rule-based methods (mainly decision trees and random forests) do not provide any insight into, interpretation, and explanation of internal mechanisms underlying the cyber-attacks and disturbances in smart grids. Our work is an attempt to fill this knowledge gap.

3. The Proposed Methodology—An Outline

Our methodology is a knowledge-based DM/ML approach having its roots in a modern branch of artificial intelligence—referred to as computational intelligence—and covering “the theory, design, application and development of biologically and linguistically motivated computational paradigms” [45]. Our methodology employs MOEOAs to generate FRBCs from data. Such FRBCs enable both accurate and interpretable identification and classification of all types of smart-grid contingencies. Details, discussions, and selected applications of our methodology can be found in several works we authored, e.g., [14]. This section outlines basic components of the proposed approach, i.e., FRBC’s input–output format, learning data, FRBC’s knowledge base, and FRBC’s MOEOA-related optimization objectives.

FRBC’s input–output format: A system (FRBC) with n input variables $x_1, x_2, \dots, x_n$, $x_i \in X_i$, and $i = 1, 2, \dots, n$, and an output variable expressed as a fuzzy set over the set $Y = \{y_1, y_2, \dots, y_c\}$ of c class labels is considered. The input variables include both numerical and categorical data. A given input variable $x_i \in X_i$, $i \in \{1, 2, \dots, n\}$ representing numerical data is characterized by a_i fuzzy sets $A_{ik_i} \in F(X_i)$, $k_i = 1, 2, \dots, a_i$. $F(X_i)$ and $F(Y)$ denote families of all fuzzy sets defined in the universes of discourse X_i , $i = 1, 2, \dots, n$ and Y , respectively. An S -type fuzzy set (associated with the linguistic value “Small”) is represented by A_{i1} . An L -type fuzzy set (associated with the linguistic value “Large”) is represented by A_{ia_i} . M -type fuzzy sets (associated with the linguistic values “Medium 1”, “Medium 2”, ..., “Medium $a_i - 2$ ”) are represented by $A_{i2}, A_{i3}, \dots, A_{i,a_i-1}$, respectively. For simplicity, A_{ik_i} s also denote the appropriate linguistic values. Membership functions with trapezoidal shapes will be assigned to fuzzy sets A_{ik_i} in the experimental part of this work. In turn, a given input variable $x_i \in X_i = \{x_{i1}, x_{i2}, \dots, x_{ia_i}\}$, or $i \in \{1, 2, \dots, n\}$, representing categorical data, is characterized by a_i fuzzy singletons $A_{ik_i} = A_{ik_i}^{(singl.)} \in F(X_i)$, $k_i = 1, 2, \dots, a_i$ defined for particular linguistic values x_{ik_i} of x_i as follows: $\mu_{A_{ik_i}^{(singl.)}}(x_i) = \begin{cases} 1, & \text{for } x_i = x_{ik_i}, \\ 0, & \text{elsewhere} \end{cases}$.

Learning data: The design process of our FRBC—using the genetic learning and MOEOA-based optimization—is based on the learning data set L containing K input–output data records:

$$L = \{x_k^{(lrn)}, y_k^{(lrn)}\}_{k=1}^K, \quad (1)$$

where $x_k^{(lrn)}$ and $y_k^{(lrn)}$ are the FRBC’s input data vector and the corresponding class label (desired FRBC’s output response), respectively, for the k -th learning data sample, $k = 1, 2, \dots, K$. $x_k^{(lrn)} = (x_{1k}^{(lrn)}, x_{2k}^{(lrn)}, \dots, x_{nk}^{(lrn)}) \in \mathbf{X} = X_1 \times X_2 \times \dots \times X_n$, where $\times$

denotes the Cartesian product of ordinary sets. In turn, $y_k^{(ltn)} \in Y$, where Y is the previously defined set of c class labels.

FRBC's knowledge base consists of R linguistic fuzzy classification rules mined from the learning data set L (1). The number R of fuzzy rules and their forms (including the number of input variables selected and the shapes of fuzzy sets) evolve during the FRBC's genetic learning combined with MOEOA-based optimization. The evolving r -th rule, $r = 1, 2, \dots, R$ can be expressed in a general form:

$$\text{IF } f_{sw_1^{(r)}}[x_1] \text{ AND } f_{sw_2^{(r)}}[x_2] \text{ AND } \dots \text{ AND } f_{sw_n^{(r)}}[x_n] \text{ THEN } y \text{ is } B_{j(r)}^{(singl.)}, \quad (2)$$

where

$$f_{sw_i^{(r)}}[x_i] = \begin{cases} [x_i \text{ is } A_{i,sw_i^{(r)}}], & \text{for } sw_i^{(r)} > 0, \\ [x_i \text{ is not } A_{i,|sw_i^{(r)}|}], & \text{for } sw_i^{(r)} < 0, \\ x_i \text{ is removed from the } r\text{-th rule,} & \text{for } sw_i^{(r)} = 0, \end{cases} \quad (3)$$

$$sw_i^{(r)} \in \{0, \pm 1, \pm 2, \dots, \pm a_i\}, i = 1, 2, \dots, n.$$

By means of $sw_i^{(r)}$ -parameters in (2) and (3), MOEOA can include/remove the i -th input variable x_i into/from the r -th rule. $B_{j(r)}^{(singl.)} \in F(Y)$ in (2) is a fuzzy singleton representing the class label $y_{j(r)}$ in the r -th fuzzy rule (2). Its membership function is defined as follows: $\mu_{B_{j(r)}^{(singl.)}}(y) = \begin{cases} 1, & \text{for } y = y_{j(r)}, \\ 0, & \text{elsewhere} \end{cases}$.

FRBC's MOEOA-related optimization objectives: In our MOEOA-based approach, FRBC's accuracy and interpretability measures are separate optimization objectives and performance indices. The FRBC's accuracy measure is expressed in the following form (that measure is subject to maximization):

$$Q_{ACC}^{(ltn)} = \frac{1}{K} \sum_{k=1}^K \delta(y'_k, y_k^{(ltn)}), \quad (4)$$

and

$$\delta(y'_k, y_k^{(ltn)}) = \begin{cases} 1, & \text{for } y'_k = y_k^{(ltn)}, \\ 0, & \text{elsewhere,} \end{cases} \quad (5)$$

where y'_k is the class label generated—as a non-fuzzy, actual response of the system (2)—for the learning data sample $x_k^{(ltn)}$. In turn, $y_k^{(ltn)}$ in (4) and (5) is the desired response of the system (2) (class label) for that sample taken from L (1); $Q_{ACC}^{(ltn)} \in [0, 1]$. In an analogous way, the accuracy measure $Q_{ACC}^{(tst)}$ for the test data has also been defined.

We use the term “FRBC's interpretability” in a more general sense that includes both the FRBC's semantic aspects and FRBC's complexity. The FRBC's complexity-related interpretability measure is defined as follows (this measure—similarly to the accuracy measure—is also subject to maximization):

$$Q_{INT} = 1 - Q_{CPLX}, \quad (6)$$

where

$$Q_{CPLX} = \frac{Q_{CPLX_R} + Q_{CPLX_A} + Q_{CPLX_F}}{3}, \quad (7)$$

and

$$Q_{CPLX_R} = \frac{1}{R} \sum_{r=1}^R \frac{n_A^{(r)} - 1}{n - 1}, Q_{CPLX_A} = \frac{n_A - 1}{n - 1}, Q_{CPLX_F} = \frac{n_F - 1}{\sum_{i=1}^n a_i - 1}, n > 1. \quad (8)$$

The measure Q_{CPLX} (7) ranges from 0 to 1, representing minimal and maximal complexities, respectively. The proposed complexity measure is an average of three sub-indices. The first one ($Q_{CPLX_R} \in [0, 1]$ (8)) expresses the complexity in relation to the average length of rules (0 and 1 mean that each rule has only one or all input variables, respectively). The second and third ones express the complexity in relation to the number of active system's inputs ($Q_{CPLX_A} \in [0, 1]$ (8)) and active system's fuzzy sets ($Q_{CPLX_F} \in [0, 1]$ (8)) (0 and 1 mean that the system has only one or all active inputs/fuzzy sets, respectively). $n_A^{(r)}$ in (8) is the length of the r -th rule (the number of its active input variables). n_A and n_F in (8) are the numbers of active inputs and active fuzzy sets (linguistic terms), respectively, in the whole system.

As far as the FRBC's semantics-related interpretability is concerned, it can be shown [46] that a special type of fuzzy partitions—referred to as strong fuzzy partitions (SFPs)—imposed upon the domains of all numerical input variables satisfies the desired requirements regarding that type of interpretability. SFPs are fuzzy partitions such that the sum of the values of all membership functions for any domain value is equal to 1. We introduced—as optimization constraints—the SFPs for all numerical input variables in our approach.

FRBC's accuracy measure $Q_{ACC}^{(lrm)}$ (4) and interpretability measure Q_{INT} (6) are processed using SPEA3—our multi-objective evolutionary algorithm previously developed and validated in [19,20]. SPEA3 modifies SPEA2 by substituting its environmental selection mechanism with a novel approach. This mechanism enhances the performance of SPEA3 by generating non-dominated solution sets with higher accuracy, wider spread, and a more uniform distribution in the objective space. Since the primary objective of this paper is to address smart-grid cybersecurity using interpretable models rather than proposing a new optimization algorithm, a detailed fundamental breakdown of SPEA3 is omitted here to preserve the thematic focus of the text. Instead, this study highlights how this established optimizer is structurally constrained by design using domain-specific shared fuzzy sets.

4. Smart-Grid Data Sets Used in the Reported Experiments

The collection of smart-grid data sets is provided—as already mentioned in the Introduction of the paper—by Mississippi State University, USA, in collaboration with Oak Ridge National Laboratories, USA [15,16]. This collection includes 15 data sets, each of which contains approximately 5000 data records (see Table 1 for detailed numbers of data records in particular sets). In this paper, these data sets are referred to as *Data_1* through *Data_15*. The data distribution in particular data sets is relatively uniform for particular simulated operational scenarios of the smart-grid system. Altogether, 37 operational scenarios are considered and can be grouped into three categories:

- (1) Natural-event-related disturbances (8 scenarios), including the following:
 - (a) Short-circuit faults, i.e., single line-to-ground faults that can occur in various locations along the line (the location is indicated by the percentage range).
 - (b) Line maintenance, i.e., opening one or more breakers (using the remote relay trip command) to perform maintenance on certain parts of the smart grid.
- (2) Cyber-attacks (28 scenarios), including the following:
 - (a) Data injection, i.e., changing the values of measurements coming from the phasor measurement unit (PMU) so as to mimic a valid fault causing the breakers to be switched off.
 - (b) Remote tripping command injection, i.e., sending commands to relays that cause the breakers to open.

- (c) Relay setting change, i.e., changing the settings to disable the relay operation in such a way that relays will not trip for a valid fault or a valid command.
- (3) Regular operation, i.e., no event (1 scenario).

Table 2 presents a summary of all 37 operational scenarios addressed in the considered 15 smart-grid data sets as well as the numbering of particular scenarios used in those data sets. Details of particular scenarios are presented in [15].

Table 1. Numbers of records in particular multi-class data sets.

Data Set	Number of Records	Data Set	Number of Records
<i>Data_1</i>	4966	<i>Data_9</i>	5340
<i>Data_2</i>	5069	<i>Data_10</i>	5569
<i>Data_3</i>	5415	<i>Data_11</i>	5251
<i>Data_4</i>	5202	<i>Data_12</i>	5224
<i>Data_5</i>	5161	<i>Data_13</i>	5271
<i>Data_6</i>	4967	<i>Data_14</i>	5115
<i>Data_7</i>	5236	<i>Data_15</i>	5276
<i>Data_8</i>	5315	Total	78,377

Table 2. Summary of operational scenarios and related events.

Event Type	Event Description	Scenarios Nos.
Natural * ¹	Short-circuit faults	1–6
Natural * ¹	Line maintenance	13, 14
Cyber-attack	Data injection	7–12
Cyber-attack	Remote tripping command injection	15–20
Cyber-attack	Relay setting change	21–30, 35–40
Regular	No event	41

*¹ Natural stands for natural-event-related disturbances.

Each data record in particular data sets is described by 128 input variables (features) including 116 variables coming from the PMU-based measurement system and 12—from the log information of the control room. There are four PMUs in the system. Each of them is integrated with the corresponding relay R1, R2, R3, or R4. Each PMU measures 29 variables. Thus, in total, 116 variables are obtained from four PMUs/ R_i , $i = 1, 2, 3, 4$. The remaining 12 input variables come from the corresponding four control panel log alerts, four relay status data, and four Snort-system alerts. Altogether, 128 input variables are collected in Table 3. In the column “Attribute name: original” of Table 3, the original (i.e., coming from the sources [15,16]) variable names are included. In turn, in the column “Attribute name: ours” of that table, shorter and (in our opinion) clearer names of those variables are presented. They will be used from now on in this work.

The cyber-attack, disturbance, and regular operation classification experiment will be performed in this paper using two different classification schemes, namely

- (a) A binary scheme in which the 37 scenarios are grouped into two classes: cyber-attack events (28 events) and the remaining ones (9 events)—see Table 4.
- (b) A three-class scheme in which the 37 scenarios are grouped into three classes: cyber-attack events (28 events), natural-event-related disturbances (8 events), and regular operation (1 event)—see Table 4.

Table 3. Names, types, and descriptions of input variables (features) used in data sets *Data_1* through *Data_15*.

No.	Attribute Name		Attribute Type ^{*1}	Attribute Description	
	Original	Ours			
Input variables for the relay R_1 :					
1.	R1-PA1:VH	$R_1: V_{\phi_A}$	num.	voltage phase angle	} of phase A
2.	R1-PM1:V	$R_1: V_{m_A}$	num.	voltage magnitude	
3.	R1-PA2:VH	$R_1: V_{\phi_B}$	num.	voltage phase angle	} of phase B
4.	R1-PM2:V	$R_1: V_{m_B}$	num.	voltage magnitude	
5.	R1-PA3:VH	$R_1: V_{\phi_C}$	num.	voltage phase angle	} of phase C
6.	R1-PM3:V	$R_1: V_{m_C}$	num.	voltage magnitude	
7.	R1-PA4:IH	$R_1: I_{\phi_A}$	num.	current phase angle	} of phase A
8.	R1-PM4:I	$R_1: I_{m_A}$	num.	current magnitude	
9.	R1-PA5:IH	$R_1: I_{\phi_B}$	num.	current phase angle	} of phase B
10.	R1-PM5:I	$R_1: I_{m_B}$	num.	current magnitude	
11.	R1-PA6:IH	$R_1: I_{\phi_C}$	num.	current phase angle	} of phase C
12.	R1-PM6:I	$R_1: I_{m_C}$	num.	current magnitude	
13.	R1-PA7:VH	$R_1: V_{\phi_1}$	num.	voltage phase angle	} of the positive
14.	R1-PM7:V	$R_1: V_{m_1}$	num.	voltage magnitude	
15.	R1-PA8:VH	$R_1: V_{\phi_2}$	num.	voltage phase angle	} of the negative
16.	R1-PM8:V	$R_1: V_{m_2}$	num.	voltage magnitude	
17.	R1-PA9:VH	$R_1: V_{\phi_0}$	num.	voltage phase angle	} of the zero
18.	R1-PM9:V	$R_1: V_{m_0}$	num.	voltage magnitude	
19.	R1-PA10:IH	$R_1: I_{\phi_1}$	num.	current phase angle	} of the positive
20.	R1-PM10:I	$R_1: I_{m_1}$	num.	current magnitude	
21.	R1-PA11:IH	$R_1: I_{\phi_2}$	num.	current phase angle	} of the negative
22.	R1-PM11:I	$R_1: I_{m_2}$	num.	current magnitude	
23.	R1-PA12:IH	$R_1: I_{\phi_0}$	num.	current phase angle	} of the zero
24.	R1-PM12:I	$R_1: I_{m_0}$	num.	current magnitude	
25.	R1:F	$R_1: f$	num.	frequency seen by the relay	
26.	R1:DF	$R_1: \frac{df}{dt}$	num.	the derivative of $R_1: f$ (frequency delta)	
27.	R1-PA:Z	$R_1: Z_m$	num.	apparent impedance seen by the relay	
28.	R1-PA:ZH	$R_1: Z_{\phi}$	num.	apparent imped. angle seen by the relay	
29.	R1:S	$R_1: S$	cat.	status flag for the relay	
Input variables for the relay R_2 :					
30.	R2-PA1:VH	$R_2: V_{\phi_A}$	num.	voltage phase angle	} of phase A
31.	R2-PM1:V	$R_2: V_{m_A}$	num.	voltage magnitude	
33.	} the remaining variables for the relay R_2 are analogous to the				
...					
58.	} variables Nos. 3–29 for the relay R_1				
Input variables for the relay R_3 :					
59.	R3-PA1:VH	$R_3: V_{\phi_A}$	num.	voltage phase angle	} of phase A
60.	R3-PM1:V	$R_3: V_{m_A}$	num.	voltage magnitude	
61.	} the remaining variables for the relay R_3 are analogous to the				
...					
87.	} variables Nos. 3–29 for the relay R_1				
Input variables for the relay R_4 :					
88.	R4-PA1:VH	$R_4: V_{\phi_A}$	num.	voltage phase angle	} of phase A
89.	R4-PM1:V	$R_4: V_{m_A}$	num.	voltage magnitude	
90.	} the remaining variables for the relay R_4 are analogous to the				
...					
116.	} variables Nos. 3–29 for the relay R_1				
117.	control_panel_log1	cat.	} three different log types for the relays R_1 , R_2 , R_3 , and R_4		
118.	control_panel_log2	cat.			
119.	control_panel_log3	cat.			
120.	control_panel_log4	cat.			
121.	relay1_log	cat.			
122.	relay2_log	cat.			
123.	relay3_log	cat.			
124.	relay4_log	cat.			
125.	snort_log1	cat.			
126.	snort_log2	cat.			
127.	snort_log3	cat.			
128.	snort_log4	cat.			

*¹ num. and cat. stand for numerical and categorical input variables, respectively.

Table 4. Binary and three-class classification schemes.

Binary scheme:	Cyber-attack events	Natural and regular events	
Scenarios Nos.:	7–12, 15–20, 21–30, 35–40	1–6, 13, 14, 41	
Three-class scheme:	Cyber-attack events	Natural events	Regular event
Scenarios Nos.:	7–12, 15–20, 21–30, 35–40	1–6, 13, 14	41

5. Experiments, Comparative Analysis, and Discussion

In this section, we present the following:

- A summary of the results of 30 experiments demonstrating the application of our knowledge-based DM/ML methodology to designing accurate and interpretable smart-grid cyber-attack and disturbance detection and classification systems. The 30 experiments include binary and three-class classification schemes applied to each of the considered 15 data sets (*Data_1* through *Data_15*).
- A cross-validation-based evaluation of our solutions (also for the purpose of comparative analysis, the k -fold cross-validation with $k = 10$ will be considered). To ensure consistency across all cross-validation runs, we used the same algorithm settings for every single run: population size of 200 individuals, crossover and mutation probabilities of 0.8 and 0.1, respectively, archive size of 10 individuals, the maximum number of 10,000 generations (stop condition). The parameter ranges for the membership functions are bounded by the data sets limits as follows: voltage magnitude from 0 to 160 kV, current magnitude from 0 to 2 kA, voltage and current angles from -180° to 180° , frequency from 55 to 65 Hz, the derivative of frequency from -5.5 to 5.5 Hz, apparent impedance from 0 to $1\text{ k}\Omega$, and apparent impedance angle from $-\pi$ to π radians.
- A broad cross-validation-based comparative analysis with 31 available alternative methods for binary classification scheme and 7 available alternative methods for three-class classification scheme. We start with the binary scheme because it is most frequently considered in alternative works.

Before we present the cross-validation experiments, we would like to show some details of the functioning of our methodology, i.e., designing FRBCs from data using the *Data_1* data set. Therefore, the results of the genetic learning and optimization process, which employs the SPEA3 MOEOA to processing of a single learning-test *Data_1* split (with 9:1 ratio) will be presented and briefly discussed. Figure 1a presents a set of solutions that forms the best available approximation—generated by our SPEA3 method for the considered *Data_1* set—to the optimal Pareto-front solutions. Numerical details related to the accuracy and interpretability of particular solutions from Figure 1a are collected in Figure 1b, in which $ACC^{(ltn)}$ and $ACC^{(tst)}$ are the percentages of correct decisions in the learning and test sets, respectively, and n_{AR} is the average number of input variables per rule (the remaining parameters have been defined earlier in Section 3 of the paper).

The solutions from Figure 1a illustrate the spectrum of the optimized accuracy-interpretability compromise. This boundary ranges from Solution No. 1, which prioritizes interpretability to Solution No. 10, which maximizes classification accuracy. Users can thus choose an individual FRBC that best satisfies their specific operational requirements.

A prominent candidate among these alternatives is the solution offering the best balance: it delivers the highest generalization capability on test data while ensuring a compact, transparent rule base. As shown in Figure 1, Solution No. 8 exemplifies such a trade-off. The model comprises merely 5 fuzzy rules ($R = 5$) utilizing a subset of 21 ($n_A = 21$) highly relevant input variables, automatically filtered from the initial 128 variables. This yields

a low average of 6.4 variables per rule ($n_{AR} = 6.4$), thereby substantially improving the overall interpretability of the system.

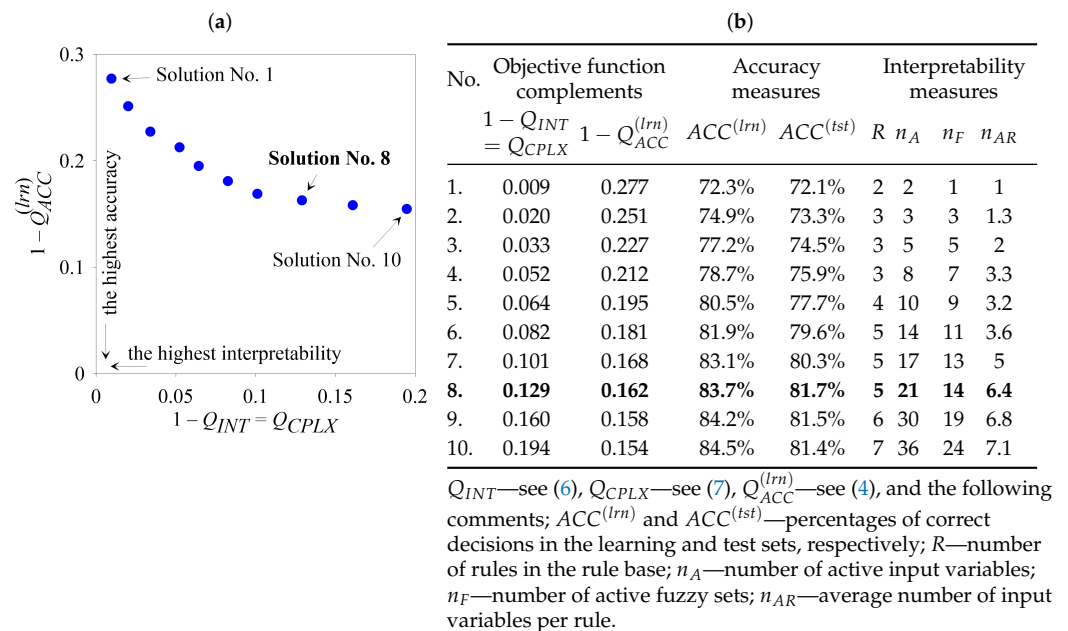


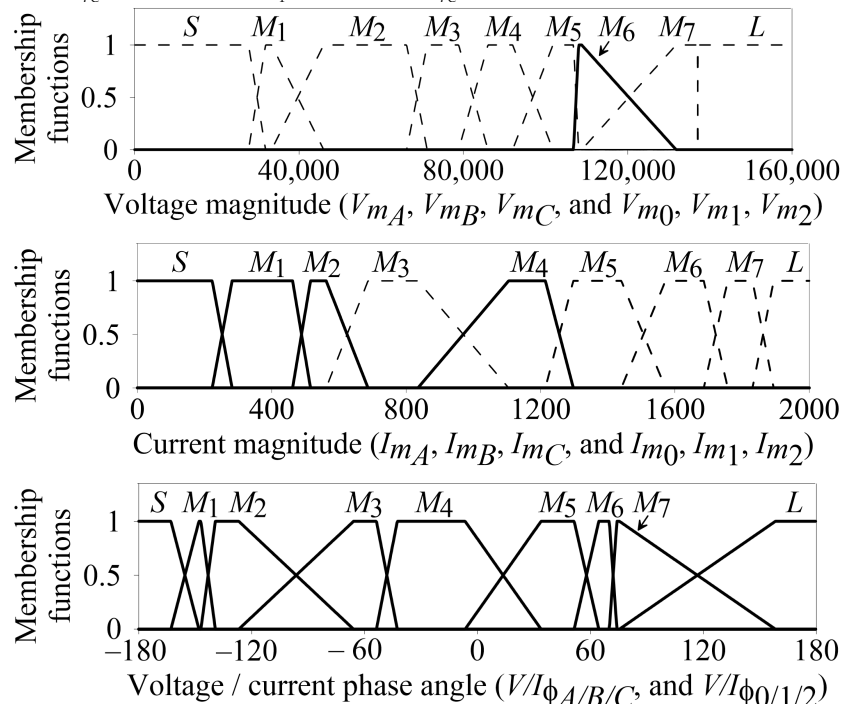
Figure 1. (a) The best available approximation of Pareto front generated by our SPEA3 method, (b) measures of accuracy and interpretability of solutions from (a) (binary classification scheme); the selected solution is marked in bold.

Table 5 presents fuzzy rule bases of selected solutions (FRBCs) from Figure 1; the solutions (FRBCs) Nos. 1, 2, 3, 4, and 8 are included. Fuzzy sets which represent numerical input variables occurring in fuzzy rules are also presented in Table 5. In our experiments, each numerical input variable is represented by 9 fuzzy sets, i.e., an S-type fuzzy set (corresponding to linguistic term “Small”), 7 M-type fuzzy sets $M_1, M_2, \dots, M_7$ (corresponding to linguistic terms “Medium1” through “Medium7”), and L-type fuzzy set (corresponding to linguistic term “Large”). Psychological research (see [46] and particularly [47]) indicates that humans can process at most 7 ± 2 conceptual entities simultaneously. Consequently, the number of conditions in a rule antecedent should not exceed this limit [47]. Extending this principle to fuzzy logic, we define the maximum upper bound of $7 + 2 = 9$ linguistic terms (fuzzy sets) for each numerical input variable. In the process of genetic learning, our approach adjusts the shapes of membership functions of particular fuzzy sets and selects the most appropriate fuzzy sets for particular fuzzy rules as shown in Table 5.

Our approach strictly enforces domain-specific shared fuzzy sets. All numerical variables representing, e.g., voltage (such as $R_3 : V_{m_B}$ or $R_4 : V_{m_A}$) are bound to exactly the same uniform base of 9 trapezoidal/triangular membership functions representing the physical semantics of voltage. Consequently, the linguistic label “M6 Voltage” has a single, invariant, and universally understood physical meaning across all voltage sensors in the entire smart grid topology. An expert or SCADA operator instantly knows what “M6 Voltage” represents without looking up a variable-specific table. Similarly, all remaining current-, phase-angle-, frequency-related attributes share independent, uniform physical-variable-specific fuzzy partitions. Moreover, that structural constraint intentionally prevents the evolutionary algorithm (SPEA3) from tailoring chaotic, sensor-specific fuzzy sets to overfit the learning data. Although it introduces a minor constraint on raw data-driven flexibility, it guarantees semantic consistency across uniform physical domains.

Table 5. Fuzzy rule bases of solutions (FRBCs) Nos. 1, 2, 3, 4, and 8 from Figure 1 (binary classification system); the knowledge base rule/attribute growth is marked in red.

No.	Fuzzy Classification Rules
Solution No. 1 ($ACC^{(tst)} = 72.1\%$):	
1.	IF $R_3: V_{\phi_B}$ is L THEN Attack
2.	IF $R_2: V_{\phi_B}$ is L THEN Natural
Solution No. 2 ($ACC^{(tst)} = 73.3\%$):	
1–2.	These rules are the same as rules 1 and 2 from Solution No. 1.
3.	IF $R_3: V_{\phi_B}$ is not S AND $R_4: V_{m_A}$ is M_6 THEN Natural
Solution No. 3 ($ACC^{(tst)} = 74.5\%$):	
1.	This rule is the same as rule 1 from Solution No. 1.
2.	IF $R_1: I_{m_1}$ is not M_2 AND $R_2: V_{\phi_B}$ is L THEN Natural
3.	IF $R_3: V_{\phi_B}$ is not S AND $R_4: V_{m_A}$ is M_6 AND $R_4: V_{\phi_B}$ is not M_4 THEN Natural
Solution No. 4 ($ACC^{(tst)} = 75.9\%$):	
1.	IF $R_3: V_{\phi_B}$ is L AND $R_4: V_{\phi_B}$ is not S THEN Attack
2.	IF $R_1: V_{\phi_C}$ is M_4 AND $R_2: V_{\phi_B}$ is not M_7 AND $R_2: I_{m_A}$ is M_1 AND $R_3: V_{\phi_0}$ is not M_1 THEN Natural
3.	IF $R_2: V_{\phi_C}$ is not L AND $R_3: V_{\phi_B}$ is not S AND $R_4: V_{m_A}$ is M_6 AND $R_4: V_{\phi_B}$ is not M_4 THEN Natural
...	
Solution No. 8 ($ACC^{(tst)} = 81.7\%$):	
1.	IF $R_1: I_{m_2}$ is not M_7 AND $R_3: V_{\phi_B}$ is L AND $R_4: V_{\phi_B}$ is not S THEN Attack
2.	IF $R_1: V_{\phi_C}$ is M_4 AND $R_1: I_{m_2}$ is not M_7 AND $R_2: V_{\phi_B}$ is not M_7 AND $R_2: I_{m_A}$ is M_1 AND $R_2: I_{\phi_2}$ is not M_3 AND $R_3: I_{\phi_A}$ is not M_4 AND $R_3: V_{\phi_0}$ is not M_1 THEN Natural
3.	IF $R_1: V_{\phi_B}$ is not M_5 AND $R_1: I_{m_2}$ is not M_5 AND $R_2: V_{\phi_B}$ is not M_2 AND $R_2: V_{\phi_C}$ is not L AND $R_2: I_{m_2}$ is not M_6 AND $R_2: I_{\phi_2}$ is not M_6 AND $R_3: V_{\phi_B}$ is not S AND $R_3: I_{\phi_A}$ is not M_5 AND $R_3: V_{\phi_0}$ is not M_3 AND $R_4: V_{m_A}$ is M_6 AND $R_4: V_{\phi_B}$ is not M_4 AND $R_4: I_{m_A}$ is not S AND $R_4: I_{m_B}$ is not M_2 THEN Natural
4.	IF $R_1: V_{\phi_A}$ is not L AND $R_2: V_{\phi_C}$ is not L AND $R_2: I_{\phi_B}$ is not M_3 AND $R_2: I_{m_1}$ is M_4 AND $R_2: I_{m_2}$ is not M_6 AND $R_3: I_{\phi_A}$ is not M_3 THEN Natural
5.	IF $R_1: I_{\phi_C}$ is M_1 AND $R_1: I_{m_1}$ is S AND $R_3: V_{\phi_C}$ is M_1 THEN Natural



An examination of Table 5 uncovers a noteworthy structural pattern. It is apparent that the rule base corresponding to Solution No. 2 inherits the two-rule framework of Solution No. 1, augmented by a new rule No. 3 (indicated in red font). Next, the rule base of Solution No. 3 retains the first rule from Solution No. 1, while extending the second and third rules of Solution No. 2 with additional conditions (marked in red font). In turn, Solution No. 4 builds upon these by extending the first rule of Solution No. 1, introducing a

new rule No. 2, and further expanding the third rule of Solution No. 3. Therefore, to achieve superior accuracy, the proposed algorithm systematically expands previously discovered fuzzy classification rules or integrates new rule structures. Through this incremental refinement, our methodology delivers a more granular and precise representation of the classification domain. It is worth emphasizing that such evolutionary behavior underscores the inherent stability and consistency of our framework.

A further, closer inspection of the Pareto-optimal solutions also reveals that as the framework approaches maximum predictive accuracy (e.g., Solution No. 8), the length of individual rules increases, sometimes exceeding ten antecedent conditions and introducing negated fuzzy sets. From a cognitive ergonomics perspective, such complex rules challenge immediate human comprehension. In practical power systems engineering, these lengthy rules should be categorized differently from compact ones. Short rules (3–5 antecedents) are suited for real-time operator alerts in SCADA interfaces, whereas hyper-complex rules are operationally useful for post-event forensics, automated anomaly logging, or formal firmware auditing, where precision is paramount and the logic is processed offline by engineers rather than during active real-time containment windows.

An essential part of this section is the evaluation of the performance of the proposed approach based on the k -fold cross-validation experiment with $k = 10$, i.e., with 9:1 learning-test data split. The selected value $k = 10$, is the same as in [16], which presents alternative solutions used in comparative analysis reported later in this section. A single, partial learning experiment for the *Data_1* data set begins with the generation of the best available Pareto-front approximation as shown above (see Figure 1 and Table 5). Next, a single solution characterized first by the highest accuracy in the test data and second by the highest interpretability is selected from the front approximation (it is Solution No. 8 in Figure 1 and Table 5). In turn, the results of all $k = 10$ partial experiments are averaged. Such an experiment is then repeated 10 times for different initializations of our method and the results are averaged. This ends the experiment for *Data_1* data set. Analogous experiments are carried out independently for data sets *Data_2* through *Data_15*. The final results—averaged over all 15 data sets—are shown in the first row of Table 6, which also presents the averaged results of as many as 31 alternative methods applied to the considered data sets in [16,32–34,36,40]. For comparison purposes, the same accuracy metrics as in those references are considered. They include

$$TPR = \frac{TP}{TP+FN}, PPV = \frac{TP}{TP+FP}, \text{ and } F_1 = 2 \frac{PPV \cdot TPR}{PPV+TPR} = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}, \quad (9)$$

where TPR , PPV , and F_1 denote true positive rate (or sensitivity), positive predictive value (or precision), and F_1 score (the harmonic mean of precision and sensitivity), respectively. In turn, TP , FP , and FN in (9) denote true positive (i.e., the cyber-attack instances correctly classified), false positive (the natural/regular instances incorrectly classified as cyber-attacks), and false negative (i.e., the cyber-attack instances incorrectly classified as natural/regular behavior), respectively.

The following alternative approaches are considered in the comparative analysis of Table 6:

- (i) Random Forest (RF), Support Vector Machine (SVM), One-Rule algorithm (OneR), Nearest Neighbor with generalization (NNge), Incremental Reduced Error Pruning algorithm (JRipper), Naive Bayes (NB), and Adaptive-Boosting meta-algorithm (Adaboost; AB) used to improve the performance of JRipper (AB+JRipper)—see [16].
- (ii) One-Class SVM (OCSVM), Histogram-Based Outlier Score (HBOS), Local Outlier Factor (LOF), Clustering-Based LOF (CBLOF), Angle-Based Outlier Detector (ABOD),

- k -Nearest Neighbor Outlier Detector (k -NNOD), Feature Bagging ensemble (FB), and Isolation Forest ensemble (iF)—see [40].
- (iii) SVM, NB, Decision Tree (DT), and RF supported by Improved Grid Search algorithm (RF+IGS)—see [32].
 - (iv) eXtreme Gradient Boosting approach (XGBoost), k -Nearest Neighbor (k -NN), DT, Artificial Neural Network (ANN), Logistic Regression (LR), Gradient Boosting DT (GBDT), and AdaBoost (AB)—see [33].
 - (v) Extra Trees classifier (ET) supported by feature selection techniques: RF, Recursive Feature Elimination (RFE), and Mutual Information (MI)—see [34].
 - (vi) A Deep-Stacked Ensemble (DSE), DT, RF and k -NN classifiers supported by a metaheuristic-based Grey Wolf Optimization (GWO) algorithm as a feature selection technique—see [36].

A significant majority of alternative approaches in Table 6 comprise non-transparent, non-interpretable, and black-box-type methods; all alternative approaches (except for the recent eight XAI/SHAP-assisted ones of [34,36]) concentrate mainly on the accuracy of the considered classification systems. Therefore, they are favored over our approach, which optimizes the compromise between the classifier's accuracy and interpretability, taking into account the strong constraints (oriented toward preserving the semantic consistency of the model) imposed by introducing common fuzzy sets for individual energy features (see Table 5) and strong fuzzy partitioning. This necessarily entailed a decrease in classification accuracy compared to unconstrained methods. Nevertheless, our approach generates interpretable classification decisions (i.e., using, on average, 6.6 fuzzy rules with 4.4 input variables per rule). In summary, the proposed approach intentionally sacrifices raw data-set-specific predictive performance to enforce such strict semantic constraints. While unconstrained black-box models or over-parameterized trees report accuracies between 96% and 99% by exploiting localized telemetry noise, they yield completely opaque structures or suffer from rule explosion. In a practical smart-grid deployment, our model's accuracy degradation may be acceptable, provided we assume the model will not replace black boxes as a standalone defense against attacks. It is rather intended as an additional layer of verification within a multi-layered defense architecture: the high-throughput black box handles basic filtering, while our transparent "IF-THEN" rules provide the real-time physical root cause analysis required by SCADA operators for incident containment.

Analogous experiments have also been carried out for the three-class classification scheme defined in Table 4. In Table 7, we present the final results of our approach and a comparison with alternative methods; the results of only 7 such methods are available—see [16]. As with the binary scheme, our approach generates both interpretable classification decisions (6.8 fuzzy rules with 3.9 input variables on average) and is also characterized by acceptable accuracy. Although the formulations in (9) represent standard binary classification metrics, the smart-grid classification problem evaluated in Table 7 involves three operational classes. To aggregate these metrics across the multi-class domain, a macro-averaging protocol was strictly implemented. Under this protocol, the TPR, PPV, and F1-score are computed independently for each individual class (treating each class as a one-vs-all binary sub-problem) and subsequently averaged. This ensures that the performance evaluation remains unskewed by class imbalances and treats the detection of rare cyber-physical anomalies with equal weight to the identification of normal operating states.

In conclusion, the proposed approach generates compact sets containing a few simple and easily comprehensible linguistic fuzzy rules that provide detailed insights into mechanisms allowing us to distinguish (i) between cyber-attacks and natural/regular events in smart grids (binary classification scheme) and (ii) between cyber-attacks, natural events, and regular events in smart grids (three-class classification scheme). It must also be ex-

plicitly noted that the comparative data compiled in Table 6 are sourced directly from the respective peer-reviewed publications of the 31 competing methods. Consequently, this overview combines results obtained under heterogeneous experimental protocols (ranging from 1:1 to 9:1 learn-to-test splits and varying cross-validation schemes). Although this direct cross-literature mapping introduces certain procedural variances, the approach was intentionally chosen to eliminate the so-called “experimenter’s bias”—wherein a proposed framework is inadvertently favored over poorly tuned, internally replicated baselines. Therefore, Table 6 should be interpreted as a broader comparative review of the reported smart-grid cybersecurity landscape rather than a perfectly controlled laboratory benchmark.

Table 6. Results of our approach and comparison with alternative methods (binary classification system).

Source	Method	Learn-to-Test Ratio	Number of Runs	Average Accuracy Measures for Learning and Test Data Sets					Average Interpretability Measures			
				$ACC^{(ltn)}$	$ACC^{(tst)}$	$TPR^{(tst)}$	$PPV^{(tst)}$	$F_1^{(tst)}$	$\bar{R}$	$\bar{n}_A$	$\bar{n}_F$	$\bar{n}_{AR}$
This paper	Our	9:1	10	80.9%	80.1%	76.3%	80.6%	78.4%	6.6	23.9	15.0	4.4
(Borges Hink et al. (2014)) [16]	RF	9:1	1	n/a	74.7%	18.5%	91.3%	29.2%	n/a	n/a	n/a	n/a
	SVM	9:1	1	n/a	71.4%	2.3%	92.6%	4.3%	-	-	-	-
	OneR	9:1	1	n/a	8.2%	95.7%	29.2%	43.6%	n/a	n/a	n/a	n/a
	NNge	9:1	1	n/a	63.9%	25.3%	52.5%	17.4%	-	-	-	-
	JRipper	9:1	1	n/a	87.6%	70.0%	84.8%	76.8%	n/a	n/a	n/a	n/a
	NB	9:1	1	n/a	20.1%	93.2%	30.3%	45.4%	-	-	-	-
	AB+JRipper	9:1	1	n/a	95.2%	87.9%	94.0%	91.4%	n/a	n/a	n/a	n/a
(Qi et al. (2021)) [40]	OCSVM	1:1	10	n/a	n/a	90.1%	94.5%	92.2%	-	-	-	-
	HBOS	1:1	10	n/a	n/a	70.2%	89.8%	78.7%	-	-	-	-
	LOF	1:1	10	n/a	n/a	60.5%	89.9%	71.2%	-	-	-	-
	CBLOF	1:1	10	n/a	n/a	89.5%	93.8%	91.6%	-	-	-	-
	ABOD	1:1	10	n/a	n/a	85.8%	93.4%	89.4%	-	-	-	-
	k-NNOD	1:1	10	n/a	n/a	90.0%	94.5%	92.1%	-	-	-	-
	FB	1:1	10	n/a	n/a	61.7%	86.9%	72.0%	-	-	-	-
	iF	1:1	10	n/a	n/a	58.7%	86.6%	69.8%	n/a	n/a	n/a	n/a
(Zhu et al. (2022)) [32]	SVM	5:1	1	n/a	78%	78%	60%	68%	-	-	-	-
	NB	5:1	1	n/a	63%	63%	69%	65%	-	-	-	-
	DT	5:1	1	n/a	95%	95%	95%	95%	n/a	n/a	n/a	n/a
	RF+IGS	5:1	1	n/a	98%	99%	98%	99%	n/a	n/a	n/a	n/a
(Lian et al. (2021)) [33]	XGBoost	n/a	n/a	n/a	96.2%	96.2%	96.2%	96.2%	-	-	-	-
	k-NN	n/a	n/a	n/a	79.4%	79.4%	80.2%	79.5%	-	-	-	-
	DT	n/a	n/a	n/a	88.2%	88.2%	88.1%	88.1%	n/a	n/a	n/a	n/a
	ANN	n/a	n/a	n/a	51.8%	51.8%	40.6%	45.4%	-	-	-	-
	LR	n/a	n/a	n/a	61.8%	61.8%	57.1%	57.7%	-	-	-	-
	GBDT	n/a	n/a	n/a	83.6%	83.6%	83.6%	83.0%	n/a	n/a	n/a	n/a
	AB	n/a	n/a	n/a	92.2%	92.2%	92.2%	92.2%	-	-	-	-
(Farsi et al. (2026)) [34]	ET	n/a	n/a	n/a	97.0%	92.5%	96.9%	94.7%	-	-	128	-
	ET+RFE	n/a	n/a	n/a	97.6%	93.3%	97.8%	95.4%	-	-	64	-
	ET+RF	n/a	n/a	n/a	97.4%	94.1%	97.4%	95.7%	n/a	n/a	81.6	n/a
	ET+MI	n/a	n/a	n/a	97.7%	94.3%	98.5%	96.3%	-	-	92.8	-
(Naeem et al. (2025)) [36]	DSE+GWO	7:3	n/a	n/a	96.8%	96.8%	96.8%	96.8%	-	-	47.6	-
	DT+GWO	7:3	n/a	n/a	98.2%	98.2%	98.2%	98.2%	n/a	n/a	47.6	n/a
	RF+GWO	7:3	n/a	n/a	97.7%	97.8%	97.8%	97.8%	n/a	n/a	47.6	n/a
	k-NN+GWO	7:3	n/a	n/a	97.9%	98.1%	98.1%	98.1%	-	-	47.6	-

n/a and “-” stand for not available and not applicable, respectively.

As far as the computational complexity of the proposed algorithm is concerned, we clarify that it operates in two distinct phases: an *offline* evolutionary optimization stage (which is computationally demanding, as are most analogous evolutionary heuristics) and an *online* real-time deployment stage. The *offline* stage is free from real-time operational constraints as it is performed periodically on historical SCADA logs to extract fuzzy rules. Once extracted, the resulting rules require only a basic logical, min-max-principle-based fuzzy inference mechanism to operate in the *online* stage. This ensures sub-millisecond operation times, enabling lightweight, line-rate anomaly detection on smart-grid communication buses. Furthermore, sharing fuzzy sets across uniform physical domains for all 128 input attributes drastically narrows the search space. Consequently, our approach achieves faster convergence and lower memory usage than traditional evolutionary rule learners, which optimize independent fuzzy partitions for each attribute.

Table 7. Results of our approach and comparison with alternative methods (three-class classification scheme).

Source	Method	Learn-to-Test Ratio	Number of Runs	Average Accuracy Measures for Learning and Test Data Sets		Average Interpretability Measures			
				$ACC^{(lrm)}$	$ACC^{(tst)}$	$\bar{R}$	$\bar{n}_A$	$\bar{n}_F$	$\bar{n}_{AR}$
This paper	Our	9:1	10	78.4%	77.4%	6.8	18.5	13.1	3.9
(Borges Hink et al. (2014)) [16]	RF	9:1	1	n/a	76.1%	n/a	n/a	n/a	n/a
	SVM	9:1	1	n/a	71.7%	-	-	-	-
	OneR	9:1	1	n/a	30.3%	n/a	n/a	n/a	n/a
	NNge	9:1	1	n/a	63.9%	-	-	-	-
	JRipper	9:1	1	n/a	87.8%	n/a	n/a	n/a	n/a
	NB	9:1	1	n/a	34.6%	-	-	-	-
	AB+JRipper	9:1	1	n/a	95.2%	n/a	n/a	n/a	n/a

n/a and “-” stand for not available and not applicable, respectively.

6. Conclusions

This paper utilizes the collection of 15 open-source simulated smart-grid data sets provided by Mississippi State University, USA, in collaboration with Oak Ridge National Laboratories, USA [15,16], as the experimental basis. The main objective and contribution of this work is to design—from those data sets—accurate as well as interpretable and transparent systems for detection and classification of cyber-attacks, natural-event-related disturbances, and regular operation of the smart grid. In order to achieve that goal, we use our knowledge-based DM/ML approach implemented as FRBCs characterized by a genetically optimized accuracy-interpretability trade-off. Due to the complementary/contradictory nature of the system’s interpretability and accuracy, we carry out the genetic learning and optimization process using our generalization (characterized by better performance) of one of the well-known MOEOAs. The transparency and interpretability of our approach lie in its ability to generate compact, explicit, and intuitive explanations of the mechanisms governing various smart-grid contingencies. These explanations are delivered as sets of fuzzy linguistic classification rules that are inherently easy to comprehend. Conversely, the accuracy of the methodology reflects its capability to correctly identify and classify these diverse contingency events.

The contribution of this work also involves a broad cross-validation-based experiment. It shows that our approach outperforms 31 alternative methods in terms of transparency and interpretability of the generated classification decisions while still providing acceptable accuracy for those decisions. It is worth emphasizing that the overwhelming majority of the existing smart-grid data classification methods, especially inherently accuracy-oriented “black boxes”, concentrate exclusively on classification accuracy, without attempting to discover—from those data—an insight into mechanisms governing different smart-grid contingencies. Although our approach is not as accurate as “black boxes”, it surpasses them in terms of transparency and interpretability of the generated classification decisions, while ensuring acceptable accuracy of these decisions. Therefore, it can act as a second cybersecurity layer that uses transparent rules to help SCADA operators instantly verify and explain alerts triggered by accurate, fast “black-box” detectors. This paper also provides a contribution to a more general and important area comprising the automatic and effective selection of input variables (features), which are essential in a given decision-making problem (such as smart-grid contingency classification considered in this work).

When considering the deployment of our approach in a real-world physical grid, we must highlight that the extracted rule bases cannot and should not be treated as a direct source of ready-to-use practical knowledge or field-ready operational rules. Since the framework was trained and evaluated on synthetic data (which is a standard practice due to the severe confidentiality of real-world critical infrastructure SCADA logs), the rules are inherently bound to the statistical properties of the benchmark data set. They serve

to validate the methodology's capability to extract logic, rather than acting as a definitive manual for grid operators.

Finally, we want to clarify that our manuscript introduces a high-level architectural and algorithmic concept rather than a fully deployed, commercialized security system. In the critical infrastructure domain, where black-box deep learning currently reigns, we view this work as a valuable “seed” for serious, practical, interpretable systems of the future. By proving that fuzzy multi-objective rule induction can maintain competitive performance under strict semantic constraints, we hope to inspire the community to advance from purely accuracy-driven models to operator-aligned security architectures.

Author Contributions: Conceptualization, M.B.G. and F.R.; Methodology, M.B.G. and F.R.; Software, M.B.G. and F.R.; Validation, M.B.G. and F.R.; Formal analysis, M.B.G. and F.R.; Investigation, M.B.G. and F.R.; Writing—original draft, M.B.G. and F.R.; Writing—review & editing, M.B.G. and F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in 15 open-source simulated smart-grid data sets provided by the Mississippi State University, USA 4 in collaboration with the Oak Ridge National Laboratories, USA at <https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets>, accessed on 1 January 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pan, S.; Morris, T.; Adhikari, U. Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. *IEEE Trans. Smart Grid* **2015**, *6*, 3104–3113.
2. Pan, S.; Morris, T.; Adhikari, U. A specification-based intrusion detection framework for cyber-physical environment in electric power system. *Int. J. Netw. Secur.* **2015**, *17*, 174–188.
3. North American Electric Reliability Corporation, Critical Infrastructure Protection (NERC-CIP). NERC Standards CIP-002-4 Through CIP-009-4. 2012. Available online: www.federalregister.gov/ (accessed on 1 January 2024).
4. The Smart Grid Interoperability Panel—Cyber Security Working Group. *NISTIR 7628 Guidelines for Smart Grid Cyber Security*; National Institute of Standards and Technology, U.S. Department of Commerce: Gaithersburg, MD, USA, 2010. Available online: <https://www.nist.gov/publications/guidelines-smart-grid-cyber-security> (accessed on 1 January 2024).
5. U.S. Government Accountability Office. *GAO-11-117; Electricity Grid Modernization: Progress Being Made on Cybersecurity Guidelines, but Key Challenges Remain to Be Addressed*. National Institute of Standards and Technology's (NIST): Gaithersburg, MD, USA, 2011. Available online: <https://www.gao.gov/products/gao-11-117> (accessed on 1 January 2024).
6. Falliere, N.; O'Murchu, L.; Chien, E. W32. Stuxnet Dossier, v1.4, Technical Report MS10-046. 2011. Available online: http://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf (accessed on 1 January 2024).
7. Ntalampiras, S. Detection of Integrity Attacks in Cyber-Physical Critical Infrastructures Using Ensemble Modeling. *IEEE Trans. Ind. Inform.* **2015**, *11*, 104–111.
8. Lin, H.; Deng, Y.; Shukla, S.; Thorp, J.; Mili, L. Cyber security impacts on all-PMU state estimator—A case study on co-simulation platform GECO. In Proceedings of the 2012 IEEE Third International Conference on Smart Grid Communications, Tainan, Taiwan, 5–8 November 2012; pp. 587–592.
9. Pan, S.; Morris, T.; Adhikari, U. Classification of Disturbances and Cyber-Attacks in Power Systems Using Heterogeneous Time-Synchronized Data. *IEEE Trans. Ind. Inform.* **2015**, *11*, 650–662.
10. Terzija, V.; Valverde, G.; Cai, D.; Regulski, P.; Madani, V.; Fitch, J.; Skok, S.; Begovic, M.M.; Phadke, A. Wide-Area Monitoring, Protection, and Control of Future Electric Power Networks. *Proc. IEEE* **2011**, *99*, 80–93.
11. Madani, V.; Novosel, D.; Horowitz, S.; Adamiak, M.; Amantegui, J.; Karlsson, D.; Imai, S.; Apostolov, A. IEEE PSRC Report on Global Industry Experiences with System Integrity Protection Schemes (SIPS). *IEEE Trans. Power Deliv.* **2010**, *25*, 2143–2155.
12. Buczak, A.; Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1153–1176.
13. Samuel, A.L. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229.

14. Gorzalczy, M.B.; Piekoszewski, J.; Rudziński, F. A Modern Data-Mining Approach Based on Genetically Optimized Fuzzy Systems for Interpretable and Accurate Smart-Grid Stability Prediction. *Energies* **2020**, *13*, 2559.
15. Mississippi State University Critical Infrastructure Protection Center. Industrial Control System Cyber Attack Data Set. 2014. Available online: <https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets> (accessed on 1 January 2024).
16. Borges Hink, R.C.; Beaver, J.M.; Buckner, M.A.; Morris, T.; Adhikari, U.; Pan, S. Machine learning for power system disturbance and cyber-attack discrimination. In Proceedings of the 2014 7th International Symposium on Resilient Control Systems (ISRCS), Denver, CO, USA, 19–21 August 2014; pp. 1–8.
17. Fazzolari, M.; Alcala, R.; Nojima, Y.; Ishibuchi, H.; Herrera, F. A review of the application of multiobjective evolutionary fuzzy systems: Current status and further directions. *IEEE Trans. Fuzzy Syst.* **2013**, *21*, 45–65.
18. Zitzler, E.; Laumanns, M.; Thiele, L. SPEA2: Improving the strength Pareto evolutionary algorithm for multi-objective optimization. In Proceedings of the Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems, Barcelona, Spain, 8–9 May 2001; pp. 95–100.
19. Rudziński, F. Finding sets of non-dominated solutions with high spread and well-balanced distribution using generalized strength Pareto evolutionary algorithm. In Proceedings of the 2015 Conference of the International Fuzzy Systems Association and European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15), Gijón, Spain, 30 June–3 July 2015; Alonso, J.M., Bustince H., Reformat M., Eds.; Atlantis Press: Dordrecht, The Netherlands, 2015; Volume 89, pp. 178–185.
20. Gorzalczy, M.B.; Rudziński, F. An improved multi-objective evolutionary optimization of data-mining-based fuzzy decision support systems. In Proceedings of the 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vancouver, BC, Canada, 25–29 July 2016; pp. 2227–2234.
21. Upendar, J.; Gupta, C.; Singh, G.; Gokaraju, R. PSO and ANN-based fault classification for protective relaying. *Gener. Transm. Distrib. IET* **2010**, *4*, 1197–1212.
22. Vyas, B.Y.; Das, B.; Maheshwari, R.P. Improved Fault Classification in Series Compensated Transmission Line: Comparative Evaluation of Chebyshev Neural Network Training Algorithms. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1631–1642.
23. Li, W.; Monti, A.; Ponci, F. Fault Detection and Classification in Medium Voltage DC Shipboard Power Systems with Wavelets and Artificial Neural Networks. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 2651–2665.
24. Ray, P.K.; Mohanty, S.R.; Kishor, N.; Catalao, J.P.S. Optimal Feature and Decision Tree-Based Classification of Power Quality Disturbances in Distributed Generation Systems. *IEEE Trans. Sustain. Energy* **2014**, *5*, 200–208.
25. Rodriguez, A.; Merino, E.; Aguado, J.; Lopez, J.J.; Munoz, F.; Martin, F.I.; Munoz, J. A Decision Tree and S-transform based approach for power quality disturbances classification. In Proceedings of the 4th International Conference on Power Engineering, Energy and Electrical Drives, Istanbul, Turkey, 13–17 May 2013; pp. 1093–1097.
26. Al Karim, M.; Chenine, M.; Zhu, K.; Nordstrom, L.; Nordström, L. Synchrophasor-based data mining for power system fault analysis. In Proceedings of the 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, Germany, 14–17 October 2012; pp. 1–8.
27. Dahal, N. Synchrophasor Data Mining for Situational Awareness in Power Systems. Ph.D. Thesis, Mississippi State University, Starkville, MS, USA, 2012. Available online: <https://scholarsjunction.msstate.edu/td/4176/> (accessed on 1 January 2024).
28. Ozay, M.; Esnaola, I.; Vural, F.T.Y.; Kulkarni, S.R.; Poor, H.V. Sparse Attack Construction and State Estimation in the Smart Grid: Centralized and Distributed Models. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 1306–1318.
29. Yan, J.; Tang, B.; He, H. Detection of false data attacks in smart grid with supervised learning. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 1395–1402.
30. Singh, V.K.; Govindarasu, M. Decision Tree Based Anomaly Detection for Remedial Action Scheme in Smart Grid using PMU Data. In Proceedings of the 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, USA, 5–10 August 2018; pp. 1–5.
31. Wang, D.; Wang, X.; Zhang, Y.; Jin, L. Detection of power grid disturbances and cyber-attacks based on machine learning. *J. Inf. Secur. Appl.* **2019**, *46*, 42–52.
32. Zhu, N.; Zhu, C.; Zhou, L.; Zhu, Y.; Zhang, X. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm. *Appl. Sci.* **2022**, *12*, 10456.
33. Lian, Y.; Gao, L.; Fang, P.; Lu, P.; Chen, L.; Gao, L.; Xiao, F. A Network Attack Detection Model of Smart Grid Based on XGBoost Algorithm. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing*; Pan, J.S., Li, J., Ryu, K.H., Meng, Z., Klasnja-Milicevic, A., Eds.; Springer: Singapore, 2021; pp. 481–488.
34. Farsi, M.; Alwateer, M.; Alsaedi, S.A.; AlSahafi, A.I.; Balaha, H.M.; Aboelnaga, M.M.; Badawy, M.; Elhosseini, M.A. Detection of disturbances and cyber-attacks in smart grids using explainable machine learning. *Sci. Rep.* **2026**, *16*, 9834.
35. Jeje, M. Cybersecurity Assessment of Smart Grid Exposure Using a Machine Learning Based Approach. In Proceedings of the 2025 IEEE Integrated STEM Education Conference (ISEC), Princeton, NJ, USA, 15 March 2025; pp. 1–7.
36. 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.* **2025**, *42*, e13556.

37. Nwakanma, C.I.; Ahakonye, L.A.C.; Jun, T.; Lee, J.M.; Kim, D.S. Explainable SCADA-edge network intrusion detection system: Tree-LIME approach. In Proceedings of the 2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Glasgow, UK, 31 October–3 November 2023; pp. 1–7.
38. Maglaras, L.A.; Jiang, J. Intrusion detection in SCADA systems using machine learning techniques. In Proceedings of the 2014 Science and Information Conference, London, UK, 27–29 August 2014; pp. 626–631.
39. Maglaras, L.A.; Jiang, J. OCSVM model combined with k-means recursive clustering for intrusion detection in SCADA systems. In Proceedings of the 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, Rhodes, Greece, 18–20 August 2014; pp. 133–134.
40. Qi, R.; Rasband, C.; Zheng, J.; Longoria, R. Detecting Cyber Attacks in Smart Grids Using Semi-Supervised Anomaly Detection and Deep Representation Learning. *Information* **2021**, *12*, 328.
41. Xu, C.; Liao, Z.; Li, C.; Zhou, X.; Xie, R. Review on Interpretable Machine Learning in Smart Grid. *Energies* **2022**, *15*, 4427.
42. Adewusi, M.; Mustafa, M. Machine learning for power-grid cybersecurity: A review. *Discov. Comput.* **2026**, *29*, 282.
43. Nelufule, N.; Mudau, C.; Nkwe, B.; Chishiri, S.; Mncwango, L.; Mutenwa, M. Cybersecurity in Smart Grids using Machine Learning: A Systematic Literature Review. In Proceedings of the 2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, 15–17 December 2025; pp. 139–146.
44. Amritha, G.; Nair, M.G.; Granelli, F. Machine Learning-Enriched Cybersecurity in Smart Grids. *IEEE Access* **2025**, *13*, 99303–99313.
45. What Is Computational Intelligence? Available online: <https://cis.ieee.org/about/what-is-ci> (accessed on 1 January 2024).
46. Gacto, M.J.; Alcalá, R.; Herrera, F. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inf. Sci.* **2011**, *181*, 4340–4360.
47. Miller, G. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **1956**, *63*, 81–97.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.