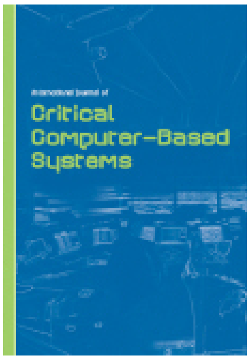
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Critical Computer-Based Systems

ISSN online: 1757-8787 - ISSN print: 1757-8779
https://www.inderscience.com/ijccbs

Detection of cyber-attacks for sensor measurement data using supervised machine learning models for modern power grid system

Manikant Panthi, Tanmoy Kanti Das

DOI: 10.1504/IJCCBS.2022.10054895

Article History:

Received:	23 October 2022
Last revised:	21 January 2023
Accepted:	12 February 2023
Published online:	30 January 2024

Detection of cyber-attacks for sensor measurement data using supervised machine learning models for modern power grid system

Manikant Panthi* and Tanmoy Kanti Das

Department of Computer Application,
National Institute of Technology Raipur,
Chhattisgarh, 492010, India
Email: mani.panthi97@gmail.com
Email: tkdas.mca@nitrr.ac.in
*Corresponding author

Abstract: The smart power grid systems are continually exposed to malicious cyber-attacks that are difficult to detect. If smart power grid attacks are not identified quickly and correctly, they may cause substantial economic losses and damage to the power system. To enhance productivity and improve the security of the smart power grid system against cyber-attacks, real-time detection of smart power grid attacks is still challenging. In recent years, there have been more cyberattacks, which have caused a lot of damage to power systems. This paper presents an experimental investigation of seven different approaches for detecting malicious activities and cyberattacks in the smart power grid system. Further, we employed maximum relevancy and minimum redundancy-hesitant fuzzy set feature selection technique to boost the attack detection performance. The experimental results demonstrate that random forest achieved the highest performance and average accuracy for two-class (95.30%) and three-class (95.33%) classifications, which shows that the presented proposed Model notably outperformed the other cyber-attack detection models.

Keywords: SCADA; MRMR-HFS; cyber-attacks; machine learning.

Reference to this paper should be made as follows: Panthi, M. and Das, T.K. (2023) 'Detection of cyber-attacks for sensor measurement data using supervised machine learning models for modern power grid system', *Int. J. Critical Computer-Based Systems*, Vol. 10, No. 4, pp.330–354.

Biographical notes: Manikant Panthi received his Bachelor of Engineering degree in Information Technology from Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh and Master of Technology from the National Institute of Technology, Rourkela. Currently, he is pursuing a PhD in Cyber-Physical System from the National Institute of Technology, Raipur, India. His research interest lies in the domain of smart grid, cyber security, and machine learning.

Tanmoy Kanti Das received his Bachelor's in Physics in 1993 from Calcutta University, Calcutta, India, and Master of Computer Applications degree from Indira Gandhi National Open University, Delhi, India, in 1999. He received his PhD in Engineering from the Jadavpur University, Kolkata in 2006. Currently, he is a faculty member with the National Institute of Technology Raipur, India. His research interest is in multimedia forensics, digital watermarking, and cyber-physical system security.

1 Introduction

The astonishing advancement of sensors and control systems has resulted in the improvement of the smart power grid (SPG), which outperforms the abilities of a traditional system (Tuballa and Abundo, 2016). The SPG system is the automatic form of energy production, distribution and transmission that combines communication and information technology (Yoldaş et al., 2017). Because of this, the SPG system is more durable, dependable, efficient, and adaptable than a traditional grid. Real-time control and monitoring are essential for smart metres, renewable energy sources, electric vehicles, and many other things (Dileep, 2020; Gungor et al., 2013; Sridhar and Govindarasu, 2014). SPG offers customers a dependable and sustainable power supply through the cyber-physical system. The intelligent grid status may be monitored and controlled by placing remote terminal units, sensors and electronic devices in the SPG field. The measurement data is collected by these electronic devices, sensors, and remote terminal units, which are then transmitted to the 'supervisory control and data acquisition' (SCADA) system (Sayed and Gabbar, 2017; Tawde et al., 2015). The SCADA system monitors and controls the critical power grid system infrastructure. Even though the SCADA system and its technologies were put in place to meet the growing need for reliable and stable energy, they are also very dependent on smart communication technologies. Thus, the modern SPG system, like water treatment plants and oil and gas refineries, that hackers can attack them. The invaders may launch a cyberattack using sophisticated technology to break the system. The adversary compromises of wide area network (WAN), neighbourhood area network (NAN) or home area network (HAN) to compromise the system. An adversary may inflict harm to electrical equipment or a lengthy power loss by hacking the system (Rawat and Bajracharya, 2015). Additionally, they can influence the control centre, offer false meter readings or prices, and fake crucial information via snooping (Han and Xiao, 2016; Zhang et al., 2019). Currently, most intrusion detection systems (IDS) utilized in SCADA in power supply networks are focused on the cyber sector while disregarding the process states in the physical field. Data IDSs are used to find security threats and attacks in a system, but they cannot stop them. However, by correctly training the detection systems, assaults may be detected effectively without any user intervention, reducing the enormous loss incurred in systems without an IDS. These systems will function as a second line of defence in all architectures and play a crucial role in cyber-physical systems detecting various threats. The classification of normal and abnormal activity by IDSs enables the system to identify unknown assaults. The data intrusion assaults are the most common cyber-assaults that risk power grid security. Generally, false data injection, load redistribution, and denial of service are the three main categories of data intrusion assaults (Deng et al., 2017; Mehrdad et al., 2018). Cybercriminals used these assaults to change transmitted data, take control of operations, disturb the safe operation of the SPG system, generate financial benefits, or even damage the physical system. Modern IDSs must identify and differentiate suspicious data from other forms of data. It keeps the data accessible while protecting the network's integrity and privacy from possible threats.On the other hand, IDS that doesn't catch intrusions has terrible effects on both utility companies and their customers. Much attention has been paid to improving feature selection (FS) and machine learning (ML) methods for finding cyberattacks in the power grid system. To identify cyber-attacks in SPG, several ML-based classification algorithms like as random forest (RF), Naive Bayes (NB), One-R, support vector machines (SVMs), decision trees (DTs), k-nearest neighbour (k-NN), and Ada-boost have been used (Borges Hink et al., 2014). This paper presents an experimental investigation of identifying abnormalities, malicious actions, and cyber-attacks in the SPG system. ML techniques are utilized to examine intrusion attacks on the SPG system. The classification and detection of intrusion attacks will assist system administrators and industrial operators in making the crucial decision to protect the SPG system from hackers and detect abnormal activities such as physical component failures and sabotage. We presented a maximum relevancy and minimum redundancy-hesitant fuzzy set (MRMR-HFS) method to select the useful features which help to accurately identify various attacks the abnormal activities in the SPG system.

The main contribution of this work is listed below:

- We presented a scheme using a combination of MRMR and HFS-based FS methods for precisely classifying various cyber-attacks on the SPG system.
- A novel filter-based FS method called MRMR-HFS selects optimal features based on a group of ranking algorithms, which improves classification accuracy.
- Using 15 publicly accessible datasets from the SPG system, we established the efficacy of the suggested approach and compared its performance for 2-and 3-class classification.

The remainder of the paper is structured as follows: In Section 2, the related work is discussed various detection methods for identifying cyber-attack in intelligent power grid systems. Section 3 describes the SPG system and dataset in detail. Section 4 provides brief preliminaries of basic concepts of hesitant fuzzy sets (HFSs) and their definitions. Section 5 defines the proposed methodology in detail. We reported the experimental findings in Section 6, and finally, we presented the conclusion and future work in Section 7.

2 Related work

The SPG system is a technology utilized by many modern power systems to deliver electricity. Smart technologies make the computer-based remote control and automation of the SPG (El-Hawary, 2014) possible, which have made energy more efficient and SPG systems susceptible to cyber-attacks. Many researchers have used different methods for the IDS. A few target systems in the smart grid where the IDS has been thoroughly addressed are the advanced metering infrastructure (AMI), the substation, and the synchro phasor system. It is stated (Radoglou-Grammatikis and Sarigiannidis, 2019) that the IDS for smart grids must quickly and accurately identify a wide range of intrusions. By comparing various patterns or aspects of known attacks, signature-based techniques have historically been used to identify malicious communications or malware. This family of detection methods is appropriate for use in communication networks for smart grids and has a low chance of false positives. For instance, signature-based Snort rules were suggested for detecting cyberattacks in the SCADA system's Modbus (Morris et al., 2013) and DNP3 (Li et al., 2015) networks. Signatures cannot identify new cyberattacks, and the current information base about cyberattacks in the smart grid is inadequate.

On the other hand, anomaly-based intrusion detection creates profiles of the system's usual activities and recognizes aberrant behaviours as intrusions. Heuristics, statistical analysis, or ML techniques may be used to create the profiles. For each instance, Blum latches the generalized likelihood test with locally optimal hypothesis testing to increase the probability that smart grid incursions and failures would be found (He and Blum, 2011). For a synchro phasor-specific IDS, Khan et al. (2018) established signature-based, heuristic, and state-full criteria to recognize both well-known and fresh cyber-attacks in the IEEE C37.118 communication architecture. Seven data stream mining techniques for the AMI were evaluated by Faisal et al., who provided suggestions for their application in the suggested distributed IDS system (Faisal et al., 2015). While signature-based IDPs are more likely to be able to identify known attacks, anomaly-based IDSs may be less likely to do so. Pan et al. suggested using heterogeneous data and a common path mining method to make an IDS that can find power system events and cyber-attack (Pan et al., 2015b). The state tracking and extraction method (STEM) technique prepares data for common route mining. Then, frequent item set mining is employed to identify common routes connected to certain system behaviours (Adhikari et al., 2018a). A common route is an ordered collection of key states for a specific cyberattack or power system incident. Things may be sorted into many categories using signatures that are common pathways. The examination of common path mining demonstrates the relationship between various events and behaviours that fit several similar routes. The IDS for the electricity system can be scaled up using this approach, although it isn't particularly precise.

Adhikari et al. (2018a, 2018b) developed the Hoeffding adaptive trees (HAT) and non-nested generalized exemplars (NNGE) techniques for creating offline and online event intrusion detection systems (EIDSs). Processing the two IDS power system security datasets using STEM is acceptable. For datasets on binary and multi-class power systems, these identification algorithms have an accuracy of greater than 90%. Although the findings of these algorithms were encouraging, their false positive rates remained above 1%. To protect the SPG from cyber attacks, Camana Acosta et al. (2020) developed an extremely randomized tree (ERT)-based strategy. KPCA was utilized to reduce the dimensions. The KPCA is a two-step procedure that entails both doing regular PCA and I translating the data into a higher-dimensional space. Then, the ERT-based approach, which utilizes many DTs, is employed for categorization. These DTs consist of nodes representing children, leaves, and roots. The testing samples are sent to the DT in the last stage of the testing phase, and the sample is guided to the leaf node by the best splits. Compared to the RF, this ET method effectively lowers bias and variance. Additionally, it quickly and accurately pinpoints the attacks. An ML approach was recommended by Ahmed et al. (2018) to protect the power system from online threats. To extract traits, the genetic algorithm (GA) is also employed. Here, the fitness function is used to evaluate the effectiveness of each option. GA helps choose the best subset with less classification error and can distinguish between two groups in this way. The SVM is then given the attributes that the GA has chosen, and it assists in identifying inaccurate data. Data are either classified as compromised or uncompromised by the system. This lessens the complexity of the detection and increases its precision. But as the system gets bigger, so does the effectiveness of the detection. An isolation forest (IF)-based technique for SPG integrity preservation was put out by Ahmed et al. (2019). They employed the IF technique to identify the assault and the PCA strategy to decrease the data's dimensionality. While PCA extracts the features, FS removes the discriminating features. In this instance, the IF separates the zero or two child nodes after they have grown fully; they are referred to as 'leaf nodes'. This IF method lowers the chance of a breach of covert data integrity. Karimipour et al. (2019) created a method for locating cyber-attacks in the SPG that uses a dynamic Bayesian Network (BN). The system is broken down into several smaller systems under this framework, and the smaller system properties are learnt using a technique called 'symbolic dynamic filtering' (SDF). The series data is first converted into an SDF symbol sequence and compressed. The 'Boltzmann machine' is then used to find the attacks. The subset selection decreases this framework's complexity by selecting a feature and the methods for filtering. The concurrent pre-processing of data and breakdown performed by the subsystems further speed up computing.

In Gururaj et al. (2023), the cybersecurity vulnerabilities to smart grids and the approaches for mitigating a denial-of-service attack are examined. This work aims to create a surrogate artificial neural network-based approach for forecasting the compressive strength of concrete, which is the most important element in the life service of concrete and its durability in civil engineering building projects (Chen, 2022). Based on machine vision recognition, this study builds an online flaw detection system for pharmaceuticals and spirits caps using an LED light source, an industrial camera, an industrial computer, and a PLC (Yang et al., 2023). Yin et al. (2021) presents a bagging strategy-based kernel extreme learning machine for complicated network intrusion detection. The bagging algorithm is used in this method to train many sub-kernel extreme learning machines separately. In Goutham et al. (2022), the author provides information on smart grids and energy storage units. Several challenging challenges connected to the integrity and dependability of rapid, scattered, and sophisticated energy transactions and data transfers might be resolved using this technology. Utilization of blockchain might reduce energy transaction costs while simultaneously enhancing the security and long-term sustainability of distributed energy resource integration, reducing obstacles to developing a more decentralised and resilient power system. Flammini et al. (2022) discussed the work on the Internet of Things, its mechanism, and its relevance concerning current developments. Sensors will serve as the channel for exchanging information from one point to another. Smart grids play a significant role compared to traditional networks, where energy use and regulation are crucial. The role of blockchain in the Internet of Things is examined. Various benefits and drawbacks, as well as obstacles encountered while using blockchain in IoT, are discussed. In this study, a blockchain-based, IoT-enabled, secure SG system is built to establish more secure metre reading communication inside the AMI system. The work utilised the blockchain method in both HAN and NANs to prevent intruders' manipulation of electric metre reading so that the government's billing system may be created effectively without incurring any financial loss. Periodically, the smart metres upload the metre reading and digital signature to a nearby server. Using a consensus technique, the distributed server checks the legitimacy of the received metre reading. Using smart contract codes running on the server, the metre reading is validated against tampering. The validated metre reading is subsequently included into the distributed ledger. In this manner, the proposed paradigm prohibits an adversary from gaining unauthorised access to private data and transmitting false metering data. Thus, inaccurate demand estimate and incorrect invoicing are prevented (Gururaj et al., 2022).

The methods currently employed to identify cyberattacks in the SPG system are mostly centred on the system's selection of features and simplicity. Nevertheless, as the scale of the system rises, so does the detection effectiveness, and the likelihood of more potent cyber attacks being detected may decline. Only a few ML-based methods have been disclosed for the detection phase. However, the ML techniques deployed have extremely weak Accuracy in assault detection. More strategies have been developed, including 'neural networks, BNs, and evolutionary algorithms'. It is hard to discern between actual assaults and false alarms due to the false positives and false negatives produced by the existing intrusion detection technologies. Evolutionary algorithms also struggle to find the optimal weights because of their slow convergence. Inaccuracy is increased by incorrect weight value setting, which lowers precision and system performance. The author introduced the MRMR-HFS-based system for cyber-attack detection to address these issues.

3 Description of the SPG system

This section describes the SPG architecture which is utilised in this investigation of the proposed scheme.

Figure 1 SPG architecture (see online version for colours)

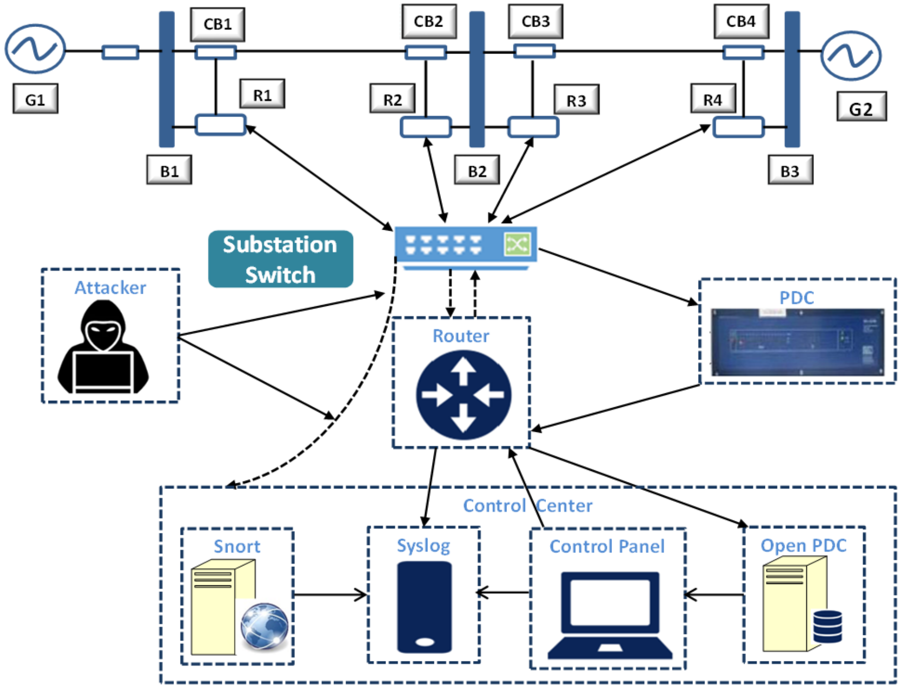


Figure 1 shows the benchmark power system architecture. The SPG system is divided into two power generators (G1, G2), four intelligent electronic devices (IEDs) (R1-R4), and four breakers (BR1-BR4). The IEDs switch the breakers on and off, and two lines are formed by connecting pairs of breakers (BR1-BR2 and BR3-BR4). The control centre may monitor and control the whole network, located at the bottom of the diagram alongside several intelligent devices. The relay employs the distance protection mechanism to trigger the brakes upon detecting mistakes and malfunctions. Since there is no validation procedure to distinguish between genuine and fake errors, the trip breaks will be executed anyway. The operators also send a manual command to the intelligent relays to turn off the breakers. As seen in Figure 1, during an attack, it is considered that the intruder is already inside the system and has authority over it and that they may issue a false order from the substation switch.

Table 1 Problem types and attack scenario

Problem types	*No. of datasets*	*No. of scenario*				
		Attack	*Normal*	*Natural*	*No-event*	*Total*
2-class	15	28	9	-	-	37
3-class	15	28	-	8	1	37
Multi-class	15	Each event has its distinct class.				37

The split scenario and problem types are listed in Table 1. The 45 different dataset types are considered here and can be split into 2-class, 3-class and multi-class. There is nothing common between all the datasets. Each dataset has 5000 samples and one sample related to any one of the 37 event circumstances. A total of 5,226 observations are a part of the 3-class datasets. These data samples comprise 1,221 natural occurrences, 3,711 attack, and 294 no-event occurrences. The used scenarios, as in Buczak and Guven (2016), are consistent. Hence, we considered various attack scenarios in Buczak and Guven (2016) line maintenance, false data injection, short circuit, relay setting change, and remote command input. The whole event scenarios (37) in binary datasets are divided into normal operation scenarios (9) and different attack scenarios (28). The 3-class real scenarios are classified into realistic scenarios (8), attack scenarios (28), and no event scenarios (1). The entire scenarios in multi-class datasets are considered as a single class. A detailed overview of the SPG dataset is given in Pan et al. (2015a) and Zaharie et al. (2011). The precise distribution for 2-class and 3-class for 15 datasets is given in Table 2.

Table 2 The number of classes and instances distribution for 15 power grid datasets

Dataset	*Total instances*	*2-class data set*		*3-class data set*		
		Attack (%)	*Natural (%)*	*Attack*	*Natural*	*No event*
D-1	4,966	3,866 (77.85%)	1,100 (22.15%)	3,866 (77.85%)	927 (22.15%)	173 (3.48%)
D-2	5,069	3,525 (69.54%)	1,544 (30.46%)	3,525 (69.54%)	1,222 (30.46%)	322 (6.35%)
D-3	5,415	3,811 (70.38%)	1,604 (29.62%)	3,811 (70.38%)	1,250 (29.62%)	354 (6.54%)
D-4	5,202	3,402 (65.4%)	1,800 (34.6%)	3,402 (65.4%)	1,397 (34.6%)	403 (7.75%)
D-5	5,161	3,680 (71.3%)	1,481 (28.7%)	3,680 (71.3%)	1,211 (28.7%)	270 (5.23%)
D-6	4,967	3,490 (70.26%)	1,477 (29.74%)	3,490 (70.26%)	1,287 (29.74%)	190 (3.83%)
D-7	5,236	3,910 (74.68%)	1,326 (25.32%)	3,910 (74.68%)	1,118 (25.32%)	208 (3.97%)
D-8	5,315	3,771 (70.95%)	1,544 (29.05%)	3,771 (70.95%)	1,188 (29.05%)	356 (6.7%)
D-9	5,340	3,570 (66.85%)	1,770 (33.15%)	3,570 (66.85%)	1,292 (33.15%)	478 (8.95%)

Table 2 The number of classes and instances distribution for 15 power grid datasets (continued)

Dataset	*Total instances*	*2-class data set*		*3-class data set*		
		Attack (%)	*Natural (%)*	*Attack*	*Natural*	*No event*
D-10	5,569	3,921 (70.41%)	1,648 (29.59%)	3,921 (70.41%)	1,322 (29.59%)	326 (5.85%)
D-11	5,251	3,669 (69.87%)	1,282 (24.41%)	3,969 (69.87%)	1,137 (24.41%)	145 (2.76%)
D-12	5,224	3,453 (66.1%)	1,771 (33.9%)	3,453 (66.1%)	1,387 (33.9%)	384 (7.35%)
D-13	5,271	4,118 (78.13%)	1,153 (21.87%)	4,118 (78.13%)	950 (21.87%)	203 (3.85%)
D-14	5,115	3,762 (73.55%)	1,353 (26.45%)	3,792 (73.55%)	1,274 (26.45%)	79 (1.54%)
D-15	5,276	3,415 (64.73%)	1,861 (35.27%)	3,415 (64.73%)	1,347 (35.27%)	514 (9.74%)

4 Preliminaries

1 Let X be a reference set, a HFS A on X is defined in terms of the function $h_A(x)$ when applied to X, returns a finite subset of [0,1], where $h_A(x)$ is a set of different values in the interval [0,1], $h_A(x)$ is referred to as the hesitant fuzzy element (HFE) (Rodríguez et al., 2014).

$$A=\left\{\left\langle \left|x, h_A(x)\right\rangle\right|, x\in X\right\} \tag{1}$$

2 Given an HFE h, its lower and upper limits are defined as follows in (Rodríguez et al., 2014):

$$\textit{Lower limit}: h^{-}(x)=\min h(x); \textit{Upper limit}: h^{+}(x)-\max h(x)$$

3 For an HFS, A = $\{<x, h_A(x)>|, x_i \in X, i = 1,2,3\ldots n\}$ the information energy is represented in (Chen et al., 2013).

$$E_{HFS}(A)=\sum\nolimits_{i=1}^{n}\left(\frac{1}{l_i}\sum\nolimits_{j=1}^{l_i} h^2 A\sigma(j)\left(x_i\right)\right) \tag{2}$$

4 The relationship between the two typical HFSs, A and B, is defined as follows in (Rodríguez et al., 2014).

$$C_{HFS}(A, B)=\sum\nolimits_{i=1}^{n}\left(\frac{1}{l_i}\sum\nolimits_{j=1}^{l_i} h_A\sigma(j)\left(x_i\right)h\sigma\left(x_i\right))\right) \tag{3}$$

5 The correlation between the two typical HFSs, A and B, is defined as follows:

$$HFS(A,B)=\frac{C_{HFS}(A,B)}{\left[C_{HFS}(A,A)\right]^{1/2}\left[C_{HFS}(B,B)\right]^{1/2}}$$

$$=\frac{\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i}h_A\sigma(j)(x_i)h_B\sigma(j)(x_i)\right)}{\sum_{i=1}^{n}\left[\begin{array}{c}\left(\frac{1}{l_i}\sum_{j=1}^{l_i}h^2A\sigma(j)(x_i)\right)1/2\\ \left[\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i}h^2B\sigma(j)(x_i)\right)1/2\right]\end{array}\right.} \tag{4}$$

where it must meet the following three conditions:

a $(A, B) = \rho_{HFS}(B, A)$

b $0 \leq \rho_{HFS}(A, B) \leq 1$

c $\rho_{HFS}(A, B) = 1$ of $A = B$.

4.1 Methods used for calculating the ranking

In this section, we define Fisher, ReliefF and information gain (IG) algorithms that are used in the proposed strategy. The ability of each character to be distinguished is measured by the information gained. As a result, it gives each character a unique rating. Additionally, the greatest IG value demonstrates a high level of discrimination. For computing the IG, it is necessary to determine the entropy as shown below.

$$entropy(s)=-p_+\log_2 p_+ - p\_\log_2 p- \tag{5}$$

where $-p_-$ and $-p_+$ are the probabilities of positive and negative labels.

$$Gain\left(s,F_m\right)=entropy(s)-\sum_{v\in F_m}^{1}\frac{\left|s_v\right|}{|s|}entropy\left(s_v\right) \tag{6}$$

where F_m is a set of all possible values for m^{th} features $F_m(m = (1\ldots.d)$ and s_v is the subset of the sample in F_m that has v^s Values and S is the whole sample in F_m.

Relief's iterative ranking system aims to give each character a fair grade. The first phase of the method considers a nil vector based on the number of features. The algorithm then selects two samples for each step, requiring that one sample be the closest neighbour in its class and the second sample be the closest neighbour in its class. At each iteration, these two samples are used to update the vector. When m is less than the sample size, the algorithm will run m times. Fisher attempts to value each feature by examining how well it can distinguish between classes and how widely data from each class are distributed according to that characteristic. The higher score of Fisher, the greater the feature's discriminating ability.

4.2 Measurement of resemblance

The three similarity measurements that are utilized in the suggested approach are described in this section. Assume that p^{th} and q^{th} columns of a dataset contain two sample vectors *X* and *Y*, respectively. These columns are often referred as p^{th} and q^{th} features. The first similarity measure, Euclidean metric (Witten and Tibshirani,), is described in equation (7), which can measure the similarities between the p^{th} and q^{th} attributes.

$$IED_{pq=\frac{1}{(euclidian(X,Y))}} = \frac{1}{|X - Y_2|} \tag{7}$$

where *X* and *Y* are two identical feature vectors, the second similarity measure for the p^{th} and q^{th} features are presented in equation (8) and contain the precise values of correspondence coefficients (Yang et al., 2023).

$$PC_{pq} = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{X})(y_i - \hat{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \hat{Y})^2}} \tag{8}$$

where *X* and *Y* are two feature vectors of the same size, x_i is the i^{th} element of *X*, and *X* is the arithmetic mean of the vector's elements. Similarly, y_i is the i^{th} element of, *Y* and *Y* is the average of the vector's elements.

$$CS_{pq} = \cos ine(X,Y) = \frac{\sum_{i=1}^{n}(x_i)(y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{9}$$

The main cosine similarity (CS) is the 3dr similarity measure (Rodríguez et al., 2014) among the attributes, which is computed similarly to the previously stated similarity measures, where *X* and *Y* are two feature vectors of equal length; also, x_i and y_i are the i^{th} elements of *X* and *Y* are, respectively.

4.3 Merit based on correlation

This section discusses an established advantage of correlation-based FS techniques. This benefit is proposed by Hall (1999). As shown in equation (10), they suggested a formula that would evaluate the relevance among characteristics and their class labels and the redundancy among features by themselves.

$$merit = \frac{k * \overline{RCF}}{\sqrt{k + k(k-1)\overline{RFF}}} \tag{10}$$

The number of features is denoted by k in the formula above, while the connection between the features and their class labels is shown by RCF. The RFF is a good illustration of feature redundancy; for instance, Pearson Correlation Coefficients can be used in equation (10). The equation states that when the relevance of the feature class is maximised, and feature redundancy is reduced, the merit obtains the maximum value. This merit considers both maximum relevance and lowest redundancy.

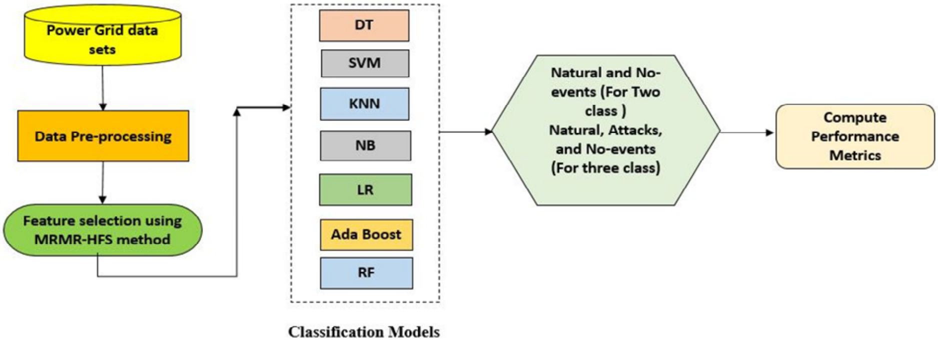
Table 3 The explanation of the symbols and variables used in this paper

Symbols name	*Detailed explanation of the symbols*
d	This variable indicates the dimension of the datasets
R-HFS	Ranking-based algorithm generated by hesitant fuzzy set
S-HFS	A hesitant fuzzy set generates the similarity measure
Rel_cf	Class labels and features relevancy
Red_ff	Redundancy among features
merit	computed value in sf
$IED_{p.q}$	The inverse of Euclidean Distance
$PC_{p.q}$	Pearson Correlation Coefficient
$CS_{p.q}$	Cosine likeness between p and q
HFE	This denotes the data structure that stores a vector of feature membership values
p – pairs	This includes all possible combinations of the features
Features_subset	The final collection of chosen features subset
Temp_sf	The temporary variable that holds the calculation of sf and fc
fc	This variable has a feature that requests to be evaluated
Index	Maximum Merit Archive Index
k	No. of features
i	This is a counter that sums between 1 and d
sf	This variable stores the subset of selected features in each iteration
max_value	The greatest value in the array
feature_max_index	The index of a value in an array that achieves the highest value
n	This represents the counter that counts between 1 and d
total_fs	Data structure that stores each iteration's proposed selected features
merit_archive	Merit values are saved in the Merit Archive

5 Proposed methodology

In this paper, we presented the MRMR-HFS-based FS method; this approach seeks to classify the different assault types seen in the SPG system as either assaults or common occurrences. The four key steps in this method are dataset pre-processing, feature subset selection, classification, and performance evaluation. Initial processing involves replacing all infinity values ('inf') in the SPG dataset with zero. Then, the informative attributes are selected by an MRMR-HFS-based methodology. The classification model is then trained using the various supervised classification approaches on the chosen subset of characteristics. Supervised models are repeatedly trained in this work to utilize split datasets and a 10-fold cross-validation configuration.

Figure 2 Proposed MRMR-HFS scheme (see online version for colours)


Power Grid data sets
Data Pre-processing
Feature selection using MRMR-HFS method
DT
SVM
KNN
NB
LR
Ada Boost
RF
Classification Models
Natural and No-events (For Two class)
Natural, Attacks, and No-events (For three class)
Compute Performance Metrics

For SPG systems, the suggested solution generates an alert message that includes natural and no-event for issues of 2 classes and natural, attack, and no-event for issues of 3 classes. Finally, several performance assessment metrics are used to evaluate the performance of the suggested technique.

In this study, the theoretical presumptions that will serve as the foundation for experimental assessments are as follows:

- By using an appropriate feature subset, the benchmark supervised classification algorithms would perform better and be less susceptible to the things of the curse of dimensionality. Therefore, we used the MRMR-HFS-based FS method.
- The complex power-grid system's non-linear and overlapping feature set is challenging. We postulated that the elegant performance of various classification approaches would aid the correct categorization of distinct attack types in power-grid systems.

5.1 *MRMR-HFS-based FS*

MRMR-HFS method is used for locating and eliminating uninteresting attributes from datasets. This step aims to choose significant characteristics that are firmly linked, instructive, and pertinent to control data events in complex power systems from the 128 features in the SPG datasets. The MRMR-HFS ensemble ranking method was employed in this investigation. This section discusses the merit value in the proposed FS approach, reluctant fuzzy sets, filter methods, and similarity measures. This approach is divided into three sections. The first section considers feature redundancy by creating uncertain fuzzy sets based on similarity measurements. Another section uses ranker algorithms to determine the relevance of characteristics and class labels. Finally, based on the two previously given facts.

MRMR-HFS has the following advantages:

1. compared to previous approaches, it selects fewer subsets of features with greater accuracy, precision, recall, and F-measure values
2. it suggests a FS technique with no parameter adjustment required
3. it does not need any pre-processing before choosing features

4 most importantly, it shows that HFS theory is a good way to combine different feature evaluation criteria.

Figure 3 Work flow diagram of MRMR-HFS

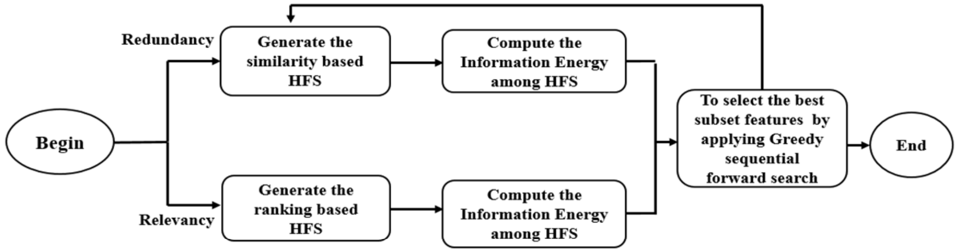

Begin
Redundancy
Relevancy
Generate the similarity based HFS
Generate the ranking based HFS
Compute the Information Energy among HFS
Compute the Information Energy among HFS
To select the best subset features by applying Greedy sequential forward search
End

Figure 3 illustrates the three main parts of the MRMR-HFS. First, ranking HFSs are created to assess the relationships between characteristics and the intended class labels; second, it analyses the similarity hesitant fuzzy sets (S-HFSs) to consider the correlations between attributes. The method then uses a sequential forward search to identify the best characteristics subset. Keeping in mind that the S-HFSs are developed throughout each iteration of the search process based on the dynamic discourse universe is crucial. The recommended MRMR-HFS approach is a FS strategy that is multivariate since this maximizes the correlations between class labels and features and reduces feature redundancy. The algorithm for selecting the feature is shown in algorithm1. Algorithm 2 and Algorithm 3 are shown the calculation of information energies and generating the R-HFS and S-HFS.

Algorithm 1 The algorithm for FS (MRMR-HFS)

```
Start
Output: Subsets of selected features.
        $sf = \{\}$; Let's consider $sf$ is an empty subset.
        Rel_cf//Calling Algorithm 2, Generating the R-HFS and Calculating their
            information energies.
        $\{max\_value, feature\_max\_index\} \leftarrow max(Rel\_cf)$;
        $n \leftarrow 1$;// initialise $n = 1$
        $sf \leftarrow sf \cup feature\_max\_index$ // feature achieves maximum value.
        $merit\_archive(n) \leftarrow max\_value$ // save the selected features.
        $total - fs(n) \leftarrow sf$; // save the selected features from the total features.
        while ($n < d$) // This loop generates d feature subsets candidates.
            for i = 1 to d // This loop find the best features in the feature subset.
                Ref_ff // Calling the Algorithm 3
                calculate the $merit(i) = \frac{k * \overline{Rel_{cff}}}{\sqrt{k + k(k-1)\overline{Rel_{ff}}}}$
            end for
            $merit(sf) \leftarrow \infty$// The merit value avoid redundant feature subset
            $[max\_value, feature\_max\_value] \leftarrow max(merit)$;
            $sf \leftarrow sf \cup \{feature\_max\_index\}$
            $n \leftarrow n + 1$
            $total\_fs(n) \leftarrow sf;$
            $merit\_archive(n) \leftarrow max\_value$
    end while
            $index \leftarrow max(merit\_archive)$
            $feature\ subset \leftarrow total\_sf(index)$
            return_feature subset;// Out put
End
```

Algorithm 2 The algorithm for calculating the Ref_cf (Generating the R-HFS and calculating their information energies)

```
Start
Out put: Ref_cf
    for m = 1 to d
            $h_r - HFS_m(F_m) \leftarrow \{IG(m), Fisher(m), ReliefF(m)\}$
              $Rel\_cf \leftarrow \frac{1}{3} IG(m^2) + ReliefF(m^2) + Fisher(m^2)$
    end for
            return Rel_cf // output
End
```

Algorithm 3 The algorithm for calculating the Red_ff (Calculate the information energies and generate the S-HFS)

```
Start
        Input: sf and candidate feature(fc);
        Output: Red_ff
                $temp\_sf \leftarrow sf \cup fc$;// adding the candidate feature fc to sf.
                $p - pairs \leftarrow$ generate all possible pairs and combinations of temp_sf.
                $cnt \leftarrow 1;$
                for each pair (p,q) in P_pairs
                        $h_{s\to(x_{pq})\leftarrow\{IED_{pq}, CS_{pq}, PC_{pq}\}}$
                        $Red\_ff(cnt) \leftarrow \frac{1}{3}(IED(p.q)^2, PC(p.q)^2, CS(p.q)^2);$
                        //calculating the information energies of the above HFS.
                        $cnt = cnt + 1;$
                end for
                return Red_ff // output
End
```

6 Experimental results

This segment describes the achieved result and experimentations in detail. All the experiments in this research were performed in Python 3.10 using a 1.60 GHz Intel(R) Core i5 processor and 16 GB of RAM. First, to verify the integrity of the premise that the MRMR-HFS FS approach would significantly contribute to the correct detection of various assaults on the SPG system, we tested different supervised classification techniques on the SPG system.

6.1 Classification algorithms

There is a wide range of classification algorithms since they are not all based on the same assumptions. NB is based on a probability distribution, and a tree-based classifier, which employs knowledge gain for splitting and building branches, are two good examples. Depending on the dataset, the algorithm's performance varies and characteristics since their underlying mechanics are different. Our work proposes a wide range of algorithms using ML to classify several attacks in the SPG systems. We chose various classification algorithms, such as DTs, RFs, NB, SVM, Adaboost, logistic regression, and KNNs (Agrawal et al., 2022; Ahirwal and Kose, 2020), that are utilized for malicious activities in the SPG system.

6.2 Evaluation parameters

The classification performance of the proposed MRMR-HFS scheme is estimated using different evaluation parameters, as described in equation (11) to equation (15) (Chandra et al., 2020, 2021; Dewangan et al., 2022), where true positive (TP) and true negative (TN) denotes the correctly predicted instances, whereas the false positive (FP) and false-negative (FN) represents the misclassification; $C_P = TP + FN$ and $C_N = TN + FP$.

$$Accuracy = \frac{TP + TN}{C_P + C_N} \times 100 \tag{11}$$

$$Specificity = \frac{TN}{C_N} \times 100 \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \times 100 \tag{13}$$

$$Recall = \frac{TP}{C_P} \times 100 \tag{14}$$

$$F_1\text{-}Score = \frac{2 \times precision \times recall}{precision + recall} \times 100 \tag{15}$$

6.3 Result analysis and discussion

From the achieved experimental results shown in Figure 4 (2-class data) and Figure 5 (3-class data), it is observed that RF and DT classifiers considerably outperformed each other for all the fifteen original feature sets. The remarkable performance of the RF and DT classifiers can also be validated using various evaluation metrics such as Accuracy, precision, recall, and F1-score using 10-fold cross-validation, as depicted in Table 4 and Table 5 for 2-class and 3-class datasets, respectively.

Figure 4 Accuracy comparison between the different classification's models on fifteen datasets for 2-class data (128 features) (see online version for colours)

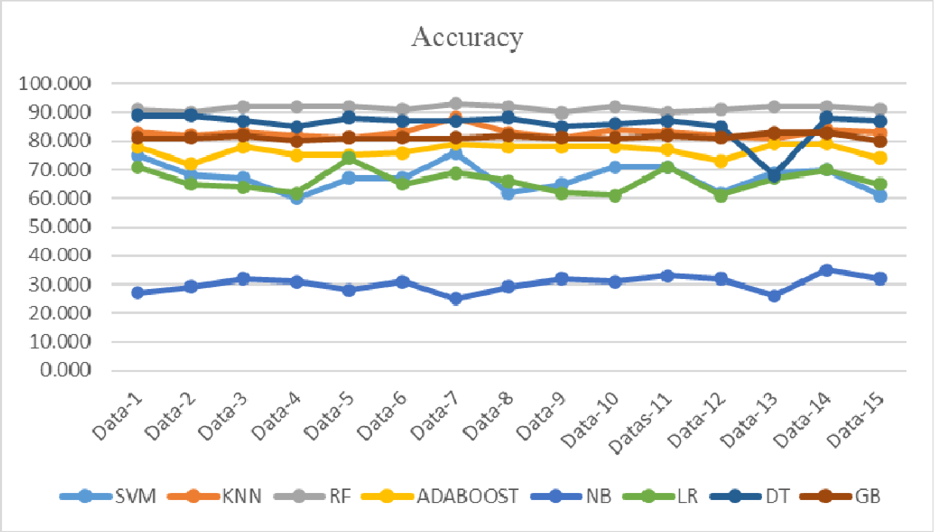


Figure 5 Accuracy comparison between the different classification's models on fifteen datasets for 3-class data (128 features) (see online version for colours)

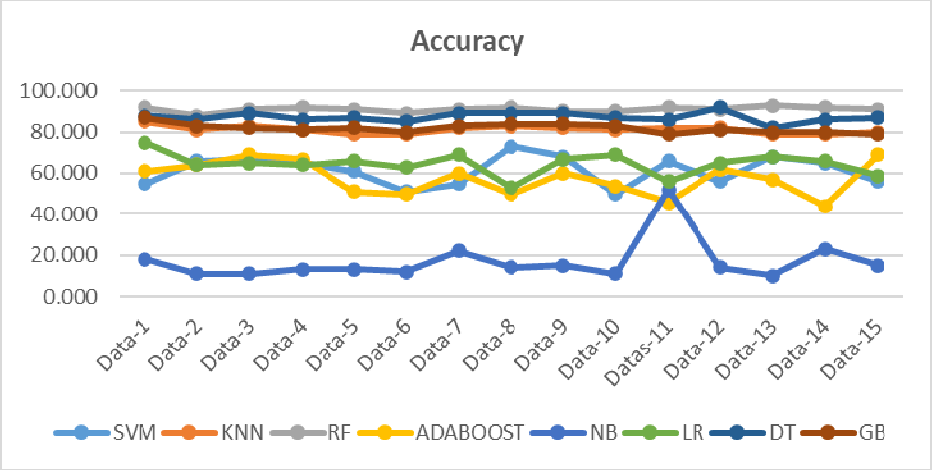


Table 4 Performance evaluation metrics of fifteen datasets for 2-class data (128 features) using RF

Dataset	*Accuracy*	*Precision*	*Recall*	*F1-score*
D-1	91.00	92.00	90.00	91.00
D-2	90.00	90.00	90.00	90.00
D-3	92.00	92.00	89.00	91.00
D-4	92.00	92.00	92.00	92.00
D-5	92.00	92.00	90.00	91.00
D-6	91.00	91.00	89.00	90.00
D-7	93.00	93.00	93.00	94.00
D-8	92.00	93.00	92.00	92.00
D-9	90.00	92.00	90.00	90.00
D-10	92.00	90.00	91.00	91.00
D-11	90.00	92.00	88.00	89.00
D-12	91.00	91.00	90.00	90.00
D-13	92.00	91.00	90.00	91.00
D-14	92.00	91.00	91.00	90.00
D-15	91.00	91.00	90.00	90.00

Table 5 Performance evaluation metrics of fifteen datasets for 3-class data (128 features) using RF

Dataset	*Accuracy*	*Precision*	*Recall*	*F-score*
D-1	92.00	92.00	91.00	91.00
D-2	88.00	89.00	88.00	89.00
D-3	91.00	92.00	90.00	91.00
D-4	92.00	93.00	92.00	93.00
D-5	91.00	90.00	90.00	90.00
D-6	89.00	87.00	85.00	85.00
D-7	91.00	92.00	91.00	92.00
D-8	92.00	93.00	90.00	91.00
D-9	90.00	91.00	92.00	90.00
D-10	90.00	89.00	88.00	88.00
D-11	92.00	91.00	90.00	91.00
D-12	91.00	92.00	90.00	91.00
D-13	93.00	91.00	92.00	91.00
D-14	92.00	91.00	91.00	91.00
D-15	91.00	91.00	89.00	90.00

However, we observed numerous characteristics had an overlapping range for various class labels when we examined the initial 128-feature input from the SPG system. Due to the obvious similarities between attacks and natural occurrences in the SPG system, this sensitive collection of characteristics has been developed. The generalization ability of ML systems is hampered by this sensitive set of characteristics, which brings about the curse of dimensionality. FS, which substitutes more discriminating qualities for less informative ones, is one potential remedy for this problem. The MRMR-HFS method was used in the second experiment to choose the ideal subset of characteristics. Table 6 summarizes the number of selected feature for 2-class and 3-class data using MRMR-HFS. The intuition is to choose those characteristics that may effectively define assaults and natural occurrences on SPG systems. Furthermore, the MRMR-HFS approach is widely employed in various fields and has been shown to be effective.

Table 6 No. of chosen feature for 2-class and 3-class data using MRMR-HFS

Data-set	*Original features*	*No. of chosen features*	
		2-class	*3-class*
D-1	128	8	9
D-2	128	14	11
D-3	128	23	13
D-4	128	13	11
D-5	128	18	12
D-6	128	15	6
D-7	128	18	8
D-8	128	22	18
D-9	128	18	8
D-10	128	13	16
D-11	128	14	2
D-12	128	16	15
D-13	128	17	15
D-14	128	14	18
D-15	128	18	17

As a result, it was noted that the suggested MRMR-HFS scheme and 10-fold cross-validation were both used to evaluate classification models. Figure 6 depicts the outcome achieved for 2-class data utilizing fifteen power-grid datasets. According to Figure 6, the recommended MRMR-HFS approach achieved the best accurate classification performance, while the DT classifier ranked second. Additionally, the performance of SVM and NB classifiers is worse than average when using the provided feature set compared to the others. Figure 7 demonstrates that the suggested method beat the others in the three-class issue, earning the best classification accuracy, whilst the DT stays in second place. Table 7 and Table 8 show the thorough performance evaluation using the given subset of features for the 2-class and 3-class tasks, respectively.

Figure 6 Accuracy comparison between the different classification models on 15 datasets for 2-class data (selected features) (see online version for colours)

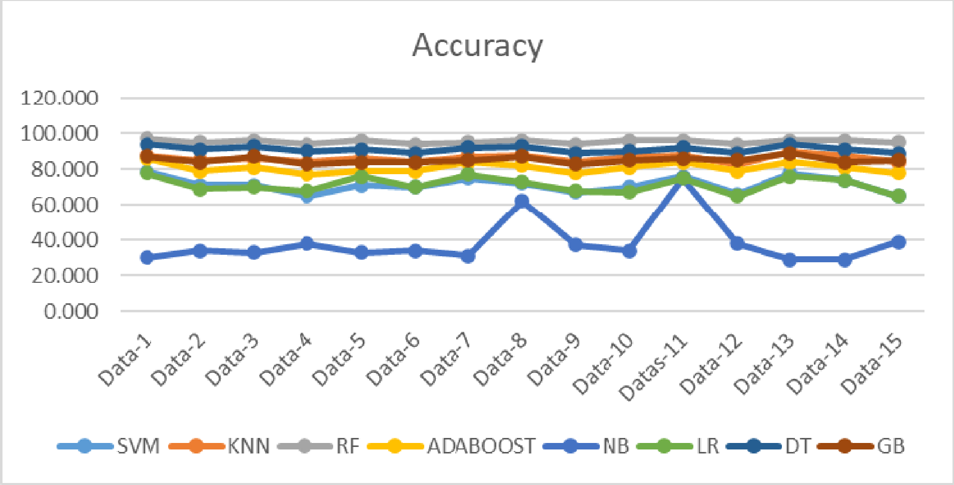


Figure 7 Accuracy comparison between the different classification models on 15 datasets for 3-class data (selected features) (see online version for colours)

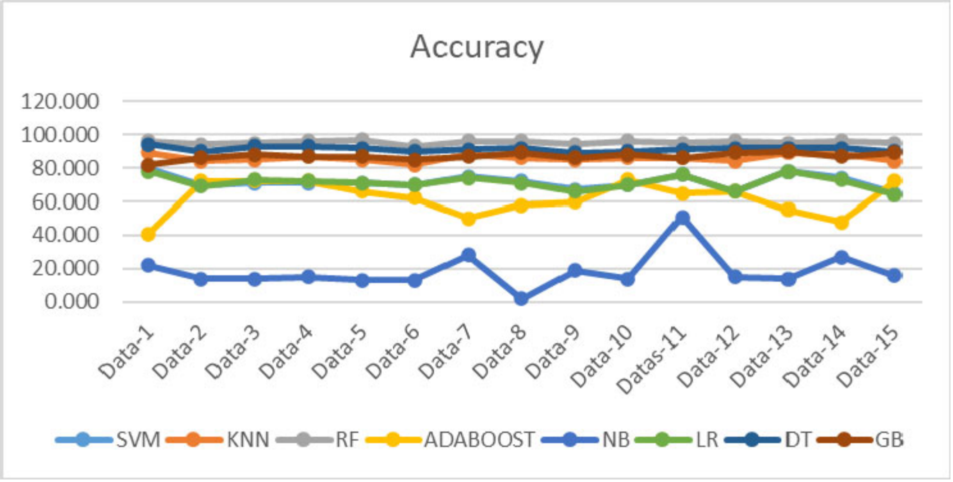


Figure 6 and Figure 7 show that the performance of the NB and SVM algorithms for 2-class and 3-class issues utilizing the specified feature subset is significantly worse than that of the other classifiers. Due to its preliminary design for binary classification issues, their performance is worse.

Table 7 Performance evaluation metrics of fifteen datasets for 2-class data (128 features) using RF

Dataset	*Accuracy*	*Precision*	*Recall*	*F-score*
D-1	97.00	97.00	97.00	97.00
D-2	95.00	95.00	95.00	95.00
D-3	96.00	96.00	96.00	96.00
D-4	94.00	94.00	94.00	94.00
D-5	96.00	96.00	96.00	96.00
D-6	94.00	94.00	94.00	94.00
D-7	95.00	95.00	95.00	95.00
D-8	96.00	96.00	96.00	96.00
D-9	94.00	94.00	94.00	94.00
D-10	96.00	96.00	96.00	96.00
D-11	96.00	96.00	96.00	96.00
D-12	94.00	94.00	94.00	94.00
D-13	96.00	96.00	96.00	96.00
D-14	96.00	96.00	96.00	96.00
D-15	95.00	95.00	95.00	95.00

Table 8 Performance evaluation metrics of fifteen datasets for 3-class data (128 features) using RF

Dataset	*Accuracy*	*Precision*	*Recall*	*F-score*
D-1	96.00	95.00	95.00	95.00
D-2	94.00	94.00	94.00	94.00
D-3	95.00	95.00	95.00	95.00
D-4	96.00	96.00	96.00	96.00
D-5	97.00	96.00	97.00	97.00
D-6	93.00	93.00	93.00	93.00
D-7	96.00	96.00	96.00	96.00
D-8	96.00	96.00	96.00	96.00
D-9	94.00	94.00	94.00	94.00
D-10	96.00	96.00	96.00	96.00
D-11	95.00	95.00	95.00	95.00
D-12	96.00	96.00	96.00	96.00
D-13	95.00	96.00	96.00	96.00
D-14	96.00	96.00	96.00	96.00
D-15	95.00	95.00	95.00	95.00

To demonstrate the robustness and superior performance of the proposed method, we compared the experimental findings obtained with the existing benchmark approaches. The comparison results presented in Tables 9 and 10 show that the proposed method has achieved better accuracy for 2-class and 3-class issues in each of the fifteen (15) datasets used in this study. Tables 11 and 12 show the comparative analysis of the proposed scheme with existing methods for two and three-class data sets. It demonstrates that the suggested average Accuracy for 15 data sets is superior to current approaches.

Table 9 Comparative analysis (% accuracy) of MRMR-HFS approach with existing classification methods for (2-class data)

Datasets	*Study-1 (Gumaei et al., 2020)*	*Study-2 (Haghnegahdar and Wang, 2020)*	*Study-3 (Borges Hink et al., 2014)*	*Study-4 (Panthi, 2021)*	*Proposed method*
D-1	96.00	94.00	90.00	89.00	97.00
D-2	96.00	92.00	89.00	90.00	95.00
D-3	96.00	91.00	90.00	80.00	96.00
D-4	96.00	92.00	89.00	89.00	94.00
D-5	96.00	97.00	88.00	90.00	96.00
D-6	95.00	94.00	90.00	91.00	94.00
D-7	96.00	93.00	91.00	92.00	95.00
D-8	96.00	94.00	89.00	96.00	96.00
D-9	95.00	91.00	90.00	89.00	94.00
D-10	96.00	94.00	91.00	89.00	96.00
D-11	96.00	97.00	90.00	95.00	96.00
D-12	97.00	93.00	91.00	96.00	94.00
D-13	95.00	95.00	91.00	85.00	96.00
D-14	95.00	96.00	88.00	88.00	96.00
D-15	96.00	97.00	90.00	95.00	95.00

Table 10 Comparative analysis (% accuracy) of MRMR-HFS approach with existing classification methods for (3-class data)

Datasets	*Study-1 (Haghnegahdar and Wang, 2020)*	*Study-2 (Borges Hink et al., 2014)*	*Study-3 (Panthi, 2021)*	*Proposed method*
D-1	99.00	94.00	90.00	96.00
D-2	95.00	93.00	91.00	94.00
D-3	98.00	95.00	87.00	95.00
D-4	99.00	96.00	88.00	96.00
D-5	99.00	94.00	84.00	97.00
D-6	96.00	95.00	91.00	93.00
D-7	97.00	93.00	92.00	96.00
D-8	92.00	92.00	91.00	96.00
D-9	96.00	94.00	91.00	94.00
D-10	99.00	95.00	88.00	96.00
D-11	92.00	93.00	89.00	95.00
D-12	99.00	94.00	89.00	96.00
D-13	94.00	95.00	90.00	95.00
D-14	95.00	96.00	89.00	96.00
D-15	96.00	94.00	92.00	95.00

Table 11 Comparative analysis of the proposed scheme with existing methods for 2-class data

Author name and year	*Model*	*Dataset*	*Average accuracy (15 datasets)*
Gumaei et al. (2020)	SVM-CFS	2-class	79.46
Haghnegahdar and Wang (2020)	WOA-ANN	2-class	95.00
Borges Hink et al. (2014)	ADA-JRIP	2-class	95.00
Kholidy and Erradi (2019)	NNGE-CFS	2-class	94.68
Panthi (2021)	RANDOMFOREST+ADABOOST	2-class	91.62
Alimiet et al. (2021)	GA-RBF SVM	2-class	91.90
Proposed scheme	*Random forest-MRMR-HFS*	*2-class*	*95.30*

Table 12 Comparative analysis of the proposed scheme with existing methods for 3-class data

Author name and year	*Model*	*Dataset*	*Average accuracy (15-datasets)*
Pan et al. (2015c)	CPM	3-class	90.04
Haghnegahdar and Wang (2020)	WOA-ANN	3-class	95.00
Borges Hink et al. (2014)	ADABOOST + JRIPPER	3-class	94.00
Panthi (2021)	RANDOMFOREST+ADABOOST	3-class	90.04
Alimi et al. (2021)	GA-RBF-SVM	3-class	90.09
Kholidy and Erradi (2019)	NNGE-CFS	3-class	94.62
Proposed scheme	*Random forest-MRMR-HFS*	*3-class*	*95.33*

7 Conclusions and future work

Cyberinfrastructure is crucial to cybersecurity, creating various security issues in modern power grid systems. System operation in real-time, the cyberinfrastructure must collect and analyse a vast volume of generated data to solve problems like classification of attack and failure of power system exposure on power system datasets with various levels of complexity. The primary objective of this study is to analyse and classify anomalous occurrences in modern power grid systems using the MRMR-HFS scheme offers an efficient and robust security solution. Additionally, we employed a 10-fold cross-validation setup for testing the proposed method's effectiveness, using well-known measures, including Accuracy, precision, recall, and F-score. The results demonstrated in this paper show that the RF and DT give the best results for two-class and three-class data set.

The proposed method is useful for dealing with high-dimensional datasets despite the existence of other FS techniques. In the future, researchers should use more diverse and relevant power grid data sets to make their findings more general. Lastly, it is important to create a new public benchmark data set that can greatly affect how ML algorithms are judged. Therefore, we can investigate other strategies, such as hybrids of the wrapper and filter-based FS techniques and parallel versions of hybrid approaches, to increase the detection engine's detection precision and processing efficiency.

References

Adhikari, U., Morris, T.H. and Pan, S. (2018a) 'Applying Hoeffding adaptive trees for real-time cyber-power event and intrusion classification', *IEEE Trans. Smart Grid*, September, Vol. 9, No. 5, pp.4049–4060, doi: 10.1109/TSG.2017.2647778.

Adhikari, U., Morris, T.H. and Pan, S. (2018b) 'Applying non-nested generalized exemplars classification for cyber-power event and intrusion detection', *IEEE Trans. Smart Grid*, Vol. 9, No. 5, pp.3928–3941, doi: 10.1109/TSG.2016.2642787.

Agrawal, N., Govil, H., Chatterjee, S., Mishra, G. and Mukherjee, S. (2022) 'Evaluation of machine learning techniques with AVIRIS-NG dataset in the identification and mapping of minerals', *Adv. Sp. Res.*, September, doi: 10.1016/j.asr.2022.09.018.

Ahirwal, M.K. and Kose, M.R. (2020) 'Audio-visual stimulation based emotion classification by correlated EEG channels', *Health Technol. (Berl).*, January, Vol. 10, No. 1, pp.7–23, doi: 10.1007/s12553-019-00394-5.

Ahmed, S., Lee, Y., Hyun, S.H. and Koo, I. (2018) 'Feature selection-based detection of covert cyber deception assaults in smart grid communications networks using machine learning', *IEEE Access*, Vol. 6, pp.27518–27529, doi: 10.1109/ACCESS.2018.2835527.

Ahmed, S., Lee, Y., Hyun, S.H. and Koo, I. (2019) 'Unsupervised machine learning-based detection of covert data integrity assault in smart grid networks utilizing isolation forest', *IEEE Trans. Inf. Forensics Secur.*, October, Vol. 14, No. 10, pp.2765–2777, doi: 10.1109/TIFS.2019.2902822.

Alimi, O.A., Ouahada, K., Abu-Mahfouz, A.M. and Rimer, S. (2021) 'Power system events classification using genetic algorithm based feature weighting technique for support vector machine', *Heliyon*, January, Vol. 7, No. 1, p.e05936, doi: 10.1016/j.heliyon.2021.e05936.

Borges Hink, R.C., Beaver, J.M., Buckner, M.A., Morris, T., Adhikari, U. and Pan, S. (2014) 'Machine learning for power system disturbance and cyber-attack discrimination', *7th Int. Symp. Resilient Control Syst. ISRCS 2014*, doi: 10.1109/ISRCS.2014.6900095.

Buczak, A.L. and Guven, E. (2016) 'A survey of data mining and machine learning methods for cyber security intrusion detection', *IEEE Commun. Surv. Tutorials*, Vol. 18, No. 2, pp.1153–1176, doi: 10.1109/COMST.2015.2494502.

Camana Acosta, M.R., Ahmed, S., Garcia, C.E. and Koo, I. (2020) 'Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks', *IEEE Access*, Vol. 8, pp.19921–19933, doi: 10.1109/ACCESS.2020.2968934.

Chandra, T.B., Verma, K., Singh, B.K., Jain, D. and Netam, S.S. (2020) 'Automatic detection of tuberculosis related abnormalities in Chest X-ray images using hierarchical feature extraction scheme', *Expert Syst. Appl.*, November, Vol. 158, p.113514, doi: 10.1016/j.eswa.2020.113514.

Chandra, T.B., Verma, K., Singh, B.K., Jain, D. and Netam, S.S. (2021) 'Coronavirus disease (COVID-19) detection in chest X-Ray images using majority voting based classifier ensemble', *Expert Syst. Appl.*, March, Vol. 165, p.113909, doi: 10.1016/j.eswa.2020.113909.

Chen, L. (2022) 'Hybrid structured artificial network for compressive strength prediction of HPC concrete', *J. Appl. Sci. Eng.*, Vol. 26, No. 7, pp.991–1001.

Chen, N., Xu, Z. and Xia, M. (2013) 'Correlation coefficients of hesitant fuzzy sets and their applications to clustering analysis', *Appl. Math. Model.*, February, Vol. 37, No. 4, pp.2197–2211, doi: 10.1016/j.apm.2012.04.031.

Deng, R., Xiao, G., Lu, R., Liang, H. and Vasilakos, A.V. (2017) 'False data injection on state estimation in power systems—attacks, impacts, and defense: a survey', *IEEE Trans. Ind. Informatics*, April, Vol. 13, No. 2, pp.411–423, doi: 10.1109/TII.2016.2614396.

Dewangan, A.K., Kumar, S. and Chandra, T.B. (2022) 'Leaf-rust and nitrogen deficient wheat plant disease classification using combined features and optimized ensemble learning', *Res. J. Pharm. Technol.*, June, Vol. 15, No. 6, pp.2531–2538, doi: 10.52711/0974-360X.2022.00423.

Dileep, G. (2020) 'A survey on smart grid technologies and applications', *Renew. Energy*, February, Vol. 146, pp.2589–2625, doi: 10.1016/j.renene.2019.08.092.

El-Hawary, M.E. (2014) 'The smart grid – state-of-the-art and future trends', *Electr. Power Components Syst.*, Vol. 42, Nos. 3–4, pp.239–250, doi: 10.1080/15325008.2013.868558.

Faisal, M.A., Aung, Z., Williams, J.R. and Sanchez, A. (2015) 'Data-stream-based intrusion detection system for advanced metering infrastructure in smart grid: a feasibility study', *IEEE Syst. J.*, Vol. 9, No. 1, pp.31–44, doi: 10.1109/JSYST.2013.2294120.

Flammini, F., Rakesh, K.R. and Jayanna, S.S. (2022) 'Integration of blockchain with IoT-enabled sensor networks for smart grids', in *Blockchain Technology for Smart Grids: Implementation, Management and Security*, pp.25–51, Institution of Engineering and Technology, doi: 10.1049/PBPO211E_ch2.

Goutham, B., Gururaj, H.L., Sunil Kumar, B.R. and Ravikumar, V. (2022) 'Smart grid: energy storage and transaction', in *Blockchain Technology for Smart Grids: Implementation, Management and Security*, pp.253–272, Institution of Engineering and Technology, doi: 10.1049/PBPO211E_ch10.

Gumaei, A. et al. (2020) 'A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids', *Appl. Soft Comput.*, November, Vol. 96, p.106658, doi: 10.1016/j.asoc.2020.106658.

Gungor, V.C. et al. (2013) 'A survey on smart grid potential applications and communication requirements', *IEEE Trans. Ind. Informatics*, Vol. 9, No. 1, pp.28–42, doi: 10.1109/TII.2012.2218253.

Gururaj, H.L. et al. (2022) *Blockchain Technology for Smart Grids: Implementation, Management and Security*, IET.

Gururaj, H.L., Swathi, B.H., Trupti, R., Darshan, U.R., Rajendra, A.B. and Paramesha, K. (2023) 'Analysis of preventive measures against DDoS attacks in smart grid', *J. Inst. Eng. Ser. B*, January, doi: 10.1007/s40031-022-00844-1.

Haghnegahdar, L. and Wang, Y. (2020) 'A whale optimization algorithm-trained artificial neural network for smart grid cyber intrusion detection', *Neural Comput. Appl.*, July, Vol. 32, No. 13, pp.9427–9441, doi: 10.1007/s00521-019-04453-w.

Hall, M.A. (1999) *Correlation-based Feature Selection for Machine Learning*, April.

Han, W. and Xiao, Y. (2016) 'Non-technical loss fraud in advanced metering infrastructure in smart grid', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10040, pp.163–172, doi: 10.1007/978-3-319-48674-1_15.

He, Q. and Blum, R.S. (2011) 'Smart grid monitoring for intrusion and fault detection with new locally optimum testing procedures', in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*, May, pp.3852–3855, doi: 10.1109/ICASSP.2011.5947192.

Karimipour, H., Dehghantanha, A., Parizi, R.M., Choo, K.K.R. and Leung, H. (2019) 'A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids', *IEEE Access*, Vol. 7, pp.80778–80788, doi: 10.1109/ACCESS.2019.2920326.

Khan, R., Albalushi, A., McLaughlin, K., Laverty, D. and Sezer, S. (2018) 'Model based intrusion detection system for synchrophasor applications in smart grid', in *IEEE Power and Energy Society General Meeting*, January, July, pp.1–5, doi: 10.1109/PESGM.2017.8274687.

Kholidy, H.A. and Erradi, A. (2019) 'VHDRA: a vertical and horizontal intelligent dataset reduction approach for cyber-physical power aware intrusion detection systems', *Secur. Commun. Networks*, June, Vol. 2019, pp.1–15, doi: 10.1155/2019/6816943.

Li, H., Liu, G., Jiang, W. and Dai, Y. (2015) 'Designing snort rules to detect abnormal DNP3 network data', in *ICCAIS 2015 – 4th International Conference on Control, Automation and Information Sciences*, October, pp.343–348, doi: 10.1109/ICCAIS.2015.7338690.

Mehrdad, S., Mousavian, S., Madraki, G. and Dvorkin, Y. (2018) 'Cyber-physical resilience of electrical power systems against malicious attacks: a review', *Curr. Sustain. Energy Reports*, March, Vol. 5, No. 1, pp.14–22, doi: 10.1007/s40518-018-0094-8.

Morris, T.H., Jones, B.A., Vaughn, R.B. and Dandass, Y.S. (2013) 'Deterministic intrusion detection rules for MODBUS protocols', in *Proceedings of the Annual Hawaii International Conference on System Sciences*, January, pp.1773–1781, doi: 10.1109/HICSS.2013.174.

Pan, S., Morris, T. and Adhikari, U. (2015a) 'A specification-based intrusion detection framework for cyber-physical environment in electric power system', *Int. J. Netw. Secur.*, Vol. 17, No. 2, pp.174–188.

Pan, S., Morris, T. and Adhikari, U. (2015b) 'Classification of disturbances and cyber-attacks in power systems using heterogeneous time-synchronized data', *IEEE Trans. Ind. Informatics*, Vol. 11, No. 3, pp.650–662, doi: 10.1109/TII.2015.2420951.

Pan, S., Morris, T. and Adhikari, U. (2015c) 'Developing a hybrid intrusion detection system using data mining for power systems', *IEEE Trans. Smart Grid*, November, Vol. 6, No. 6, pp.3104–3113, doi: 10.1109/TSG.2015.2409775.

Panthi, M. (2021) 'Identification of disturbances in power system and DDoS attacks using machine learning', *IOP Conf. Ser. Mater. Sci. Eng.*, January, Vol. 1022, No. 1, p.012096, doi: 10.1088/1757-899X/1022/1/012096.

Radoglou-Grammatikis, P.I. and Sarigiannidis, P.G. (2019) 'Securing the smart grid: a comprehensive compilation of intrusion detection and prevention systems', *IEEE Access*, Vol. 7, pp.46595–46620, doi: 10.1109/ACCESS.2019.2909807.

Rawat, D.B. and Bajracharya, C. (2015) 'Cyber security for smart grid systems: Status, challenges and perspectives', in *Conference Proceedings – IEEE SOUTHEASTCON*, April, Vol. 2015, June, pp.1–6, doi: 10.1109/SECON.2015.7132891.

Rodríguez, R.M., Martínez, L., Torra, V., Xu, Z.S. and Herrera, F. (2014) 'Hesitant fuzzy sets: state of the art and future directions', *Int. J. Intell. Syst.*, June, Vol. 29, No. 6, pp.495–524, doi: 10.1002/int.21654.

Sayed, K. and Gabbar, H.A. (2017) 'Scada and smart energy grid control automation', in *Smart Energy Grid Engineering*, pp.481–514, Elsevier, doi: 10.1016/B978-0-12-805343-0.00018-8.

Sridhar, S. and Govindarasu, M. (2014) 'Model-based attack detection and mitigation for automatic generation control', *IEEE Trans. Smart Grid*, March, Vol. 5, No. 2, pp.580–591, doi: 10.1109/TSG.2014.2298195.

Tawde, R., Nivangune, A. and Sankhe, M. (2015) 'Cyber security in smart grid SCADA automation systems', in *ICIIECS 2015 – 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems*, March, pp.1–5, doi: 10.1109/ICIIECS.2015.7192918.

Tuballa, M.L. and Abundo, M.L. (2016) 'A review of the development of Smart Grid technologies', *Renew. Sustain. Energy Rev.*, Vol. 59, pp.710–725, doi: 10.1016/j.rser.2016.01.011.

Witten, D.M. and Tibshirani, R. (2010) 'A framework for feature selection in clustering', *J. Am. Stat. Assoc.*, June, Vol. 105, No. 490, pp.713–726, doi: 10.1198/jasa.2010.tm09415.

Yang, Q., Yu, X. and Chen, Q. (2023) 'Design of drug and wine bottlecap defect detection system based on machine vision', *J. Appl. Sci. Eng.*, Vol. 26, No. 4, pp.489–500, doi: 10.6180/jase.202304_26(4).0005.

Yin, S., Li, H., Laghari, A., Karim, S. and Jumani, A. (2021) 'A bagging strategy-based kernel extreme learning machine for complex network intrusion detection', *ICST Trans. Scalable Inf. Syst.*, October, Vol. 8, No. 33, p.171247, doi: 10.4108/eai.6-10-2021.171247.

Yoldaş, Y., Önen, A., Muyeen, S.M., Vasilakos, A.V. and Alan, İ. (2017) 'Enhancing smart grid with microgrids: challenges and opportunities', *Renew. Sustain. Energy Rev.*, May, Vol. 72, pp.205–214, doi: 10.1016/j.rser.2017.01.064.

Zaharie, D., Perian, L., Negru, V. and Zamfirache, F. (2011) 'Evolutionary pruning of non-nested generalized exemplars', in *SACI 2011 - 6th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, May, pp.57–62, doi: 10.1109/SACI.2011.5872973.

Zhang, M. et al. (2019) 'False data injection attacks against smart gird state estimation: construction, detection and defense', *Sci. China Technol. Sci.*, Vol. 62, No. 12, pp.2077–2087, December, doi: 10.1007/s11431-019-9544-7.