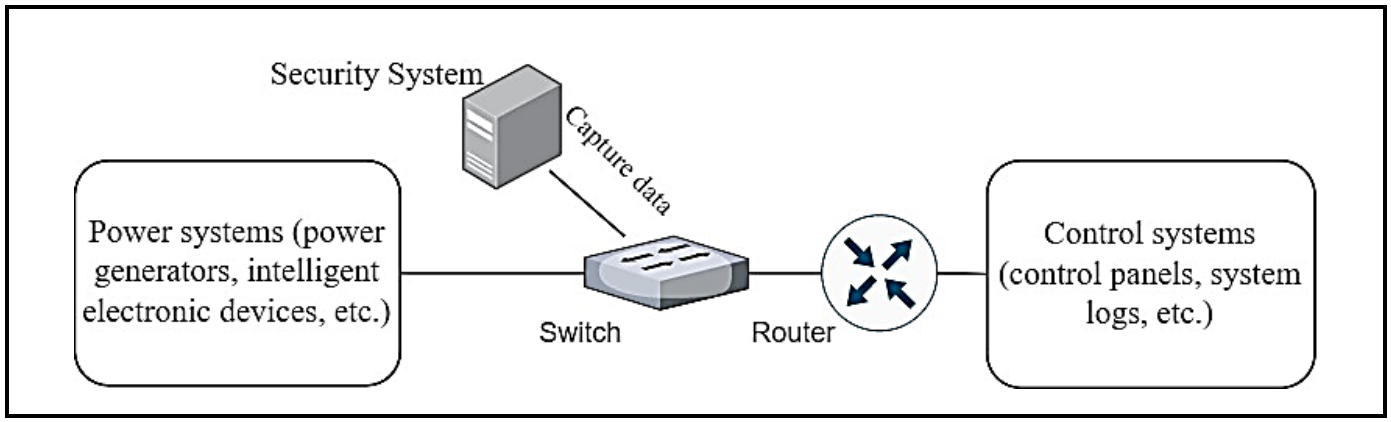
Data Mining for Enhanced Security: A Transformative Framework for Smart Grid Protection

Rabie A. Ramadan [1 2], Muataz S. Al-Daweri [1, *], Ismail S. Al Muniri [1]
[1] *Department of Information Systems, College of Economics, Management and Information Systems, Nizwa University, Nizwa, Oman*
[2] *Computer Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt*



ABSTRACT: Smart grids fall at the intersection of conventional energy systems and modern informatics in the present digitalized energy environment. The growing number of linked devices and sensors in these networks leads to the generation of complex structures and vast quantities of data, presenting benefits and challenges. Safeguarding these complex structures against malicious intrusions and illegal activities is an important problem. The paper's main objective is to enhance smart grid security by utilizing the data mining and Artificial Intelligence (AI) approaches. As huge amounts of data are collected from the smart grids based on tiny and smart internet of things (IoT) devices, this data poses challenges as well as provides opportunities. The challenges come from analyzing this huge data, especially in real-time. At the same time, it provides opportunities to enhance the smart grid services and protection. Therefore, to overcome these challenges, this paper proposes a feedforward deep learning approach for data mining to secure the smart grid from different anomalies and allow the system to adapt to any risk it might face. Deep learning will allow the system to adjust dynamically to emerging risks. The proposed system has been examined using Power System Attack Datasets sourced from the Mississippi State University and Oak Ridge National Laboratory. The results show a detection accuracy of 91% just using 50% of the dataset features. Different percentages of the features are examined as well. However, we concluded that 50% of the features are enough for identifying the smart grid risks based on the given dataset.

Keywords: Smart Grids; Sustainable Energy; Security; Digital Age; Big Data.

استخراج البيانات لتعزيز الأمان: إطار تحويلي لحماية الشبكات الذكية

ربيع أ. رمضان ,معتز س. الداوري ،إسماعيل س. المنيري

الملخص: تناقش الورقة تقاطع نظم الطاقة التقليدية والمعلوماتية الحديثة في سياق الشبكات الذكية. مع زيادة عدد الأجهزة والمستشعرات المتصلة، تولد الشبكات الذكية هياكل معقدة وكميات هائلة من البيانات، مما يقدم فوائد وتحديات في نفس الوقت. الهدف الرئيسى هو تعزيز أمان الشبكات الذكية باستخدام تقنيات استخراج البيانات والذكاء الاصطناعي (AI). تشكل البيانات الناتجة عن العديد من أجهزة إنترنت الأشياء (IoT) في الشبكات الذكية تحديات تحليلية، خاصة في الوقت الفعلى، ولكنها توفر أيضًا فرصًا لتحسين الخدمات والحماية. لمعالجة هذه التحديات، تقترح الورقة نهج التعلم العميق التقدمى لاستخراج البيانات لتأمين الشبكة الذكية من مختلف الشذوذات وتمكين النظام من التكيف مع المخاطر المحتملة. يتيح هذا النهج للنظام التكيف ديناميكيًا مع التهديدات الناشئة. تم اختبار النظام المقترح باستخدام مجموعات بيانات هجمات نظام الطاقة من جامعة ولاية ميسيسيبي ومختبر أوك ريدج الوطني، وحقق دقة اكتشاف بلغت 91٪ باستخدام 50٪ فقط من ميزات مجموعة البيانات. كما تم اختبار نسب مختلفة من الميزات، وخلصت الدراسة إلى أن 50٪ من الميزات كافية لتحديد مخاطر الشبكة الذكية بناءً على مجموعة البيانات.

الكلمات المفتاحية: الشبكات الذكية؛ الطاقة المستدامة؛ الأمان؛ العصر الرقمي؛ البيانات الضخمة.

Corresponding author's e-mail: muataz@unizwa.edu.om

DOI:10.53540/tjer.vol.21iss1pp23-32

1. INTRODUCTION

In our modern world, most traditional power grids have been replaced by smart power grids, smart grids for simplicity. These smart grids facilitate the inclusion of information systems and data collection utilizing many sensors and devices (Borlase, 2017). This integration between the smart grids of information systems makes it possible to advise energy reliability and sustainability. It offers valuable insights into grid status, operation, and user activities. It allowed researchers to investigate many of the grid issues, as well as industries, to expand their utilization of the grids. Researchers were able to identify patterns, trends, and anomalies that allow a deep understanding of the grid dynamics. So, researchers identified the factors that influence grid reliability, efficiency, and overall performance, enabling policymakers and industry to be more informed in decision-making. They allow them to implement effective grid operation strategies. However, integrating smart grids and information systems technology raises new challenges, such as security issues. It also raises many of the smart grid's vulnerabilities that could possibly disrupt energy delivery systems, attach customer privacy, and manipulate the energy market.

At the same time, the sectors and organizations are facing significant transformation from traditional systems to new smart grids. In fact, most of the time, they combine traditional systems and the latest advancements in modern information systems (Rohde & Hielscher, 2021). These advancements include involving customers in the process of energy management and enhancing the efficiency of power grids. This has also raised another level of security where many tiny sensors are involved in the smart grid management operations (Tuballa & Abudno, 2016). Therefore, based on recent research, security has become a serious issue of newly developed smart grid networks (Clastres, 2011).

Thus, it is obvious that the smart grids' security, management, and operations are critical issues in this context. In fact, many cyberattacks pose huge risks to the smart grid infrastructure and its ecosystem (Kim et al., 2023), which is of great importance on the national and international levels. This has attracted many researchers and scholars investigating the topic. There is also a large number of features found in some datasets related to smart grids (Morris & Gao, 2014), which points to the importance of finding a smaller size of features to give good detection accuracy. There are not many studies that explore this type of experimentation in this field.

At the same base, Artificial Intelligence (AI) techniques are emerging. It allows knowledge to be acquired from raw data and can be efficiently adapted to detect any possible variabilities. AI techniques are also effective in the early detection of risks in real-time systems such as smart grid systems, ensuring data integrity.

The challenge in the field of smart grid security is the timely identification and mitigation of security incidents. The current security mechanisms tend to be more responsive than anticipated since they face challenges in keeping up with the constantly evolving environment of cyber threats and the complex architecture of the smart grid system. In addition, the vast amount and rapid rate at which smart grids produce data provide a challenge to conventional security processes, hindering the detection of significant patterns that may indicate an imminent security breach. The delay in identifying abnormalities often leads to substantial reaction time delays, during which the potential threat may become severe and permanent. Identifying anomalies and possible security breaches is complicated by distinguishing between normal irregular patterns that arise from abnormal but approved use and those that signify security risks. Figuring this out is hard because the grid's usage pattern constantly changes. This is because of things like weather, customer behaviour, and the way the energy source works.

The primary concern examined in this study is the insufficiency of traditional security systems in effectively identifying and reacting to evolving and intricate threats inside the information-abundant setting of intelligent power grids. There exists a pressing need for the development of a security framework that has the capability to:

- Effectively analyze and interpret vast amounts of operational data to identify subtle, complex patterns indicative of security anomalies.
- Operate in real-time to provide timely detection and mitigation to counter potential threats before they escalate into full-scale security incidents.
- It can be rebuilt to learn and evolve, adapting its detection capabilities to the ever-changing cyber threats and smart grid operations landscape.

This paper aims to explore the possibility of data mining and AI techniques for reshaping the security of smart grids. Certainly, this research will be of interest to many entities, including the smart grid industry, policymakers, and customers. The paper proposes a novel framework for smart grid security that is possible to affect energy future security. The study also focuses on which features are the most important for smart grid security to achieve the highest accuracy possible.

The method used in this paper utilized AI for data mining to the huge amount of data collected and transferred from the smart grid devices. It aims to identify the early anomalies, the behaviour of the data, and security breaches. It detects early induction to unauthorized access and patterns of the smart grids and prevents the problem from escalating.

The significance of this study is to provide the following:

- A new method design for smart grid protection. This method involves the use of a feed-forward deep learning approach.
- An analysis that supports employing feature reduction methods and careful feature selection to boost the efficacy of machine learning models. The method utilizes the random forest (RF) approach to rank the features for selection.

- A strategy that can be employed to make a machine learning model adaptive for rebuilding to detect new security threats. This strategy suggests the use of ensemble techniques for adapting to new security threats.

2. RELATED WORK

Smart grid cybersecurity has been extensively investigated by many researchers in recent years. The first paper explored here is the one that tried to integrate blockchain technology with machine learning for the purpose of protecting smart grids (Tuballa & Abundo, 2016). It targets the security of peer-to-peer energy transfer in some of the applications. Also, the authors reviewed some of the collaborative state-of-the-art technology for smart grids. Furthermore, the authors of (Huda et al., 2024) discuss the importance of smart home technology and their automation in developing smart cities. Part of the smart cities is the smart grid where advanced technologies could be used. The research showed the importance of modern technologies in smart cities such as Digital Twin and Federated Learning. Those technologies facilitate the development of urban areas as well as smart grids. As an extension, the authors of (Hasan et al., 2023) investigated the cybersecurity and complex variabilities of smart grids through different types of analysis. Furthermore, the authors (Bouramdane, 2023) focus on different smart grid attacks that have significant effects on smart grid power transfer as well as the devices. They demonstrated the efficiency of artificial intelligence techniques in detecting some of these attacks.

Also, extensive research was conducted by Bouramdane, (2023), examining the cyber-physical characteristics of smart grids and showed that there are a number of complexities and variabilities associated with smart grids. They examined the overall systems of the smart grid to reach this conclusion. In relation to the cybersecurity of smart grids, the authors (Bouramdane, 2023) investigated some of the cybersecurity attacks that could be conducted on smart grids and concluded that AI algorithms could be beneficial in protecting smart grids from such attacks, especially deep learning. Furthermore, (Sifat et al., 2023), the authors introduced a Digital Twin grid for data analytics and stressed that blockchain technology could be beneficial to smart grid security.

In (Wasumwa, 2023) the authors investigated the utilization of collaborative strategies and explored emerging technologies in smart grid security. Again, the focus is on the combination of blockchain technology and deep learning for the benefit of smart grid security. They also, recommend them to the future of smart grids to avoid the current and future risks. One of the possible attacks on the smart grids is the face data injection attack which affects the reliability and security of the collected data and their analysis. This has been studied by (Habib et al., 2023). Similarly, the authors recommended collaborative and advanced technology for handling these security issues.

Once more the authors of (Ding et al., 2022) investigated the different cybersecurity attacks that have been noticed on smart grids. They reported 10 10-year analyses and the threats and risks that faced smart grids during this period. Also, they stressed the importance of blockchain and AI techniques in protecting smart grid data. on the same track, the authors of (Mazhar et al., 2023) studied the utilization of IoT devices for data collection and how they can be beneficial in pattern recognition. On the other hand, (Murugeshwari et al., 2023) presented encryption as a solution to smart grid data security. they proposed an elliptic curve as one of the encryption techniques for privacy protection. their proposal also involves cloud-based data analysis. Simultaneously, the paper (Mirzaee et al., 2022) investigated the many security and privacy challenges that smart grids encounter, emphasizing the increased vulnerabilities posed by improved automation and communication technologies, as well as the inclusion of machine learning.

Sifat et al. (2023) presented the digital twin concept's transformative potential for smart grids in a well-written paper. This study stressed the need for predicting the future and real-time grid monitoring, particularly when applying blockchain technology to increase cybersecurity. In relation to education, the article (Kamsamrong et al., 2022) from the "Cybersecurity Curricula Recommendations for Smart Grids" initiative identified substantial skills deficiencies within the European Union. It endorsed better cybersecurity education for smart grid security by providing practicable, useful programs, online courses, and gamification components to boost participation.

Furthermore, the study by (Bhattacharya et al., 2022) highlighted the importance of incentive mechanisms in smart grids. It addressed the problems and still unsolved issues with data quality, privacy, and security, as well as the application of technologies such as game theory, blockchain, and AI in implementing these incentives. In (Guo et al., 2022), the authors presented a deep-federated-learning architecture. Using real-world datasets, they proved their usefulness in strengthening the security of Point of Interest (POI) microservices in cyber-physical systems.

Even with much of the existing research, insufficient information is known about the potential applications of advanced data mining techniques to enhance the security and real-time monitoring of smart grids. To address this gap, this paper provides a novel approach for detecting potential security breaches that combines decision trees and deep learning. This framework contributes to the development of smart grid cybersecurity research by attempting to develop new paradigms for smart grid security and offering practical alternatives for businesses in data-rich environments.

.

Table 1. Pros and cons of the related work methodologies.

Reference	Technology/Method	Pros	Cons
(Tuballa & Abundo, 2016)	Blockchain + ML	Peer-to-peer energy transfers security enhancement and integration of collaborative technologies.	Integrating blockchain with existing ML models could be complex.
(Huda et al., 2024)	Smart Home Tech	It facilitates urban area development, and it is essential to smart city frameworks.	It may increase dependency on digital technologies, which raises privacy concerns.
(Hasan et al., 2023)	Cyber-Physical Analysis	It addresses variabilities in smart grids it is involves comprehensive system review.	The complexity of the systems may introduce new vulnerabilities.
(Bouramdane, 2023)	AI for Attack Detection	It is effective in detecting smart grid attacks, and it uses advanced AI techniques.	AI models may require extensive data and it may be prone to sophisticated attacks.
(Sifat et al., 2023)	Digital Twin + Blockchain	Blockchain offers real-time monitoring and increased cybersecurity.	The implementation costs are high and need significant infrastructure overhaul.
(Wasumwa, 2023)	Blockchain + Deep Learning	It is a promising security approach for future security measures against advanced threats.	It is still in exploratory stages; it may not be fully practical for current grid systems.
(Habib et al., 2023)	False Data Injection Study	It highlights the critical vulnerability and proposes some of the countermeasures.	Focusing on one type of attack might not be generalized to other threats.
(Ding et al., 2022)	Blockchain + AI	The paper provided a long-term analysis emphasizing blockchain and AI for data protection.	It requires continuous updates and maintenance of AI models and blockchain systems.
(Mazhar et al., 2023)	IoT + AI	It enhances smart grid operations in terms of data collection and pattern recognition.	If not properly secured, IoT devices can be entry points for cyberattacks
(Murugeshwari et al., 2023)	Encryption (Elliptic Curve)	It utilizes advanced encryption techniques and is expected to have strong privacy protection.	Implementation challenges to the complex encryption.
(Mirzaee et al., 2022)	ML for Security	Detailed exploration of ML threats and countermeasures.	High dependency on data quality and availability; potential for overfitting.
(Kamsamron g et al., 2022)	Cybersecurity Education	Promotes essential skills development and innovative teaching methods like gamification.	It may not immediately impact the current workforce's ability to handle cybersecurity threats.
(Bhattacharya et al., 2022)	Game Theory + Blockchain + AI	Addresses incentive mechanisms along with tech integration for enhanced security.	Complex integration of multiple advanced technologies can be challenging.
(Guo et al., 2022)	Deep Federated Learning	Proved effective in enhancing security for microservices in cyber-physical systems.	It still requires further real-world testing and validation for widespread deployment.

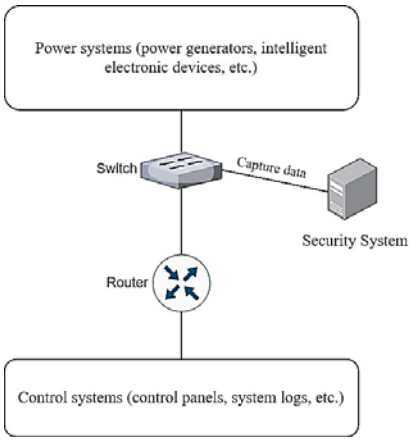


Figure 1. Smart grid proposed analysis and monitoring structure

3. PROPOSED FRAMEWORK AND APPROACHES

This section describes the proposed framework for the security of the smart grids. It considers the mentioned issues in the problem statement section. It also considers distributed analysis instead of centralized analysis due to the huge amount of data collected from the grids. The proposed framework assumes different sensors are already deployed on the smart grid units. Figure 1 shows the proposed security system where the smart grids are connected to a switch, which is used to capture the data for analysis and monitoring. The system is responsible for the heavy processing and overall decisions related to data analysis or security.

The details of the procedures and techniques used in the research are provided in the following subsections. This

includes information on materials, experimental setup, and data collection methods. The information should be detailed enough to allow replication of the study.

3.1 Dataset Description

This paper utilizes a dataset from the "Cybersecurity Curricula Recommendations for Smart Grids" project (Morris & Gao, 2014). It is an in-depth compilation made to examine and improve smart grid security. The dataset is composed of one initial set, which consists of 15 sets, each including 37 power system event scenarios. The datasets are categorized into Binary, Three-class, and Multiclass datasets. They are structured in ARFF and CSV formats to ensure compatibility with different applications.

Two power generators (G1 and G2), four Intelligent Electronic Devices (IEDs, R1 to R4), regulating breakers (BR1 to BR4), and two lines linking these breakers compose the power system architecture upon which the scenarios in the dataset are developed. The two gas turbine generators (G1 and G2) have a capacity of 150 megawatts each. They also operate at 13.8 kilovolts, which are essential for generating electricity. They also work at a frequency of 60 Hz with control features such as automatic voltage regulators and speed control governors. Certainly, this enhances the stability and reliability of power output under different load conditions. The system also integrates four intelligent electronic devices (IEDs) (R1 to R4) for system integrity. The main purpose of those IEDs is to monitor and quickly isolate faults by calculating the impedance to the fault using the distance protection method. Therefore, the tripping of circuit breakers is controlled. Also, the manual control of falls is considered an option for system flexibility. The communication of these IEDs via IEC 61850 protocol. That ensures reliable data exchange across the smart grid. Vacuum circuit breakers (BR1 to BR4) are also associated with IEDs with breaking capacities of 40 KA. They work at a system voltage of 13.8 kV. Those brakes have spring mechanisms for reliable operational responses to fault conditions. Two main transmission lines link these breakers, forming a robust network for reliable electricity distribution with multi-point protection.

The dataset covers the following forms of scenarios: data injection (Attack), relay configuration modifications (Attack), remote tripping command injections (Attack), and short-circuit problems. Every one of these scenarios illustrates a distinct facet of a possible attack vector or vulnerability in a smart grid system. Short-circuit problems, for example, can occur anywhere along a power line. On the other hand, data injection attacks try to change variables, including current-voltage sequence components, to make operators blind by triggering blackouts.

The dataset includes 128 characteristics from measurements made by phasor measuring units (PMUs), which gauge electrical waves on the electrical grid. Each PMU has 29 different kinds of measurements, for a total of 116 measurement columns. Twelve columns are also available for relay logs, Snort alerts, and control panel logs. These characteristics and metrics are essential for evaluating the performance and security of smart grid systems.

3.2 Proposed Method Design

The proposed approach employs a feature selection method to measure the importance of the features in the dataset. Since the dataset includes 128 features, reducing the number of features may be necessary. This step makes the process of detecting attacks/faults faster and, in general, creates a lighter approach. For this task, a random forest (RF) measures the features' importance (Hasan et al., 2016). The feature selection with RF falls within the embedded methods group. The advantages of wrapper and filter techniques are combined in this embedded approach. They are carried out by algorithms with integrated feature selection techniques of their own. The features that are chosen at the top of the decision trees are typically more important than the selected features at the leaf nodes of the trees to provide a better understanding because top splits typically result in larger information benefits. Thus, the features are ranked based on their position in the trees. Furthermore, there are generally two ways to measure their importance besides their locations: the Gini importance index and the permutation importance index. The same procedure is used in this study as in (Hasan et al., 2016) to measure the features' importance. 20%, 50%, 80%, and then all the features are used in a neural network (NN) to test the performance of these features. It will also show how powerful the RF is for measuring and selecting the most important features.

The overall steps of the proposed method are shown in Figure 2. The first step of the method involves cleaning, transforming, and integrating the available sets of the dataset into one. Once the data is integrated, a normalization method called MaxAbsScaler maps the values of the datasets into -1 and 1. For instance, if we have the values 1, -1, and 2, they will be normalized as 0.5, -1, and 1. This normalization method gives the activation function, i.e., the hyperbolic tangent function, better values since this function maps the inputs to values of -1 and 1, which can be more suitable for it.

Once the data is prepared and ready for training, the feature selection approach (based on the RF) will find which features are important to indicate the top 20%, 50%, 80%, and 100%. Further, these sets of top features are used to train a neural network with different configurations to find the final results for discussion and recommendations.

3.3 Classification Model Design

The model used for classifying the attacks/faults in the used dataset consists of using an NN. A typical NN (as shown in Figure 3) includes input, hidden, and output layers, and the more complex the data, the more hidden layers may be added to make the network learn from the data better. That concept is called deep learning. In this study, a network with two hidden layers. As for the

features, different amounts of features are used, i.e., 20%, 50%, 80%, and 100%. In Figure 3, the number of inputs n is equal to 26 (20%), 64 (50%), 102 (80%), and 128 (100%), whereas the number of hidden nodes z is equal to 100 at each hidden layer.

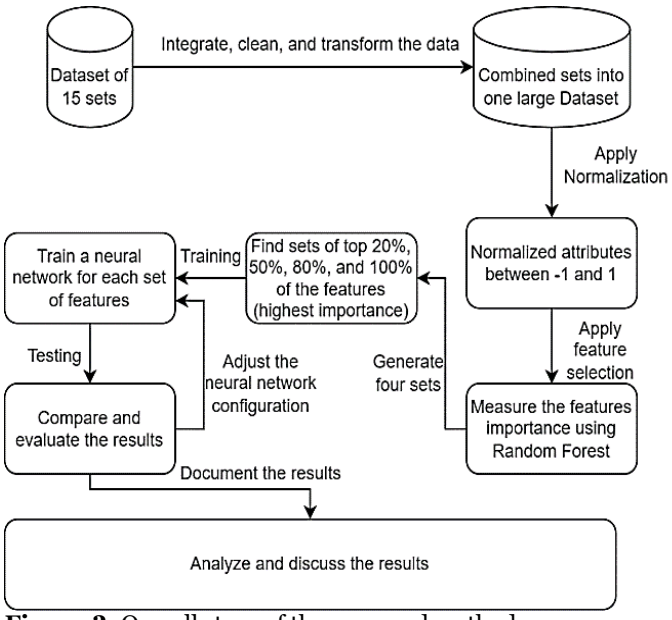


Figure 2. Overall steps of the proposed method.

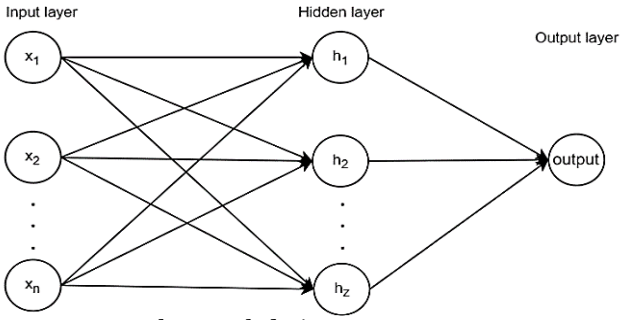


Figure 3. Neural network design.

Table 2. Experiment setup.

Settings	Value/approach
Number of hidden nodes	200 (two hidden layers of each 100 nodes)
Number of inputs	26, 64, 102, and 128
Number of outputs	Binary (either 0 or 1)
Learning rate	0.01
Number of epochs	200 and 400
Optimization algorithm	Adam
Number of instances	78377

4. RESULTS AND DISCUSSION

After running the feature selection approach, the features with the highest importance are listed and sorted, provided in Table 3.

Given the selected settings (see Table 2), the test's classification measurement is calculated based on the confusion matrix (Godbole, 2002). This is a commonly used approach for measuring the performance of a classification task, finding various measurements, such as the detection accuracy and false alarm rate regarding the actual label and predicted label. The confusion matrix is provided in Table 4.

The results (using the confusion matrix) are given in Tables 5, 6, 7, and 8, respectively, for the top 20%, 50%, 80%, and 100% of the features. The number of instances used for this test is a randomized 20%, as the other 80% was used for the training phase.

As shown in Figure 4, the training loss for 20% of the top features illustrates that it can get below the value of 0.4 after approximately 125 epochs, whereas looking at the others, the loss went below 0.4 after less than 75 epochs. The final training loss after 400 epochs for the top 50% is similar to that of the 80% and 100%; thus, it can be concluded from this that the top 50% of the features are sufficient for achieving low training error values.

Based on the data given in the confusion matrices, the detection rate (Drate), accuracy, and false alarm rate (FARate) are provided in Table 9. These measurements are calculated using the following equations:

$$DRate = \frac{TP}{TP+FN} \quad (1)$$

$$FARate = \frac{FP}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

The proposed evaluation questions are critical for assessing the classification model performance. In equation (1), the Detection Rate (DRate) is the Sensitivity or True Positive Rate. It measures the actual positive instances that are correctly identified by the model. The True Positives (TP) are considered as the instances correctly identified as positive while False Negatives (FN) are positive instances incorrectly labelled as negative. Equation (2) defines the False Alarm Rate (FARate), also considered the False Positive Rate. It highlights the proportion of negative instances that are erroneously classified as positive. It is mainly important for false alarm identification. Equation (3) is used to calculate the accuracy of the proposed model. It represents the true results, either positive or negative cases. The accuracy equation has all four components of the confusion matrix.

Table 3. Feature importance (sorted top-down, left-to-right).

Feature	Importance	Feature	Importance	Feature	Importance	Feature	Importance
R2-PM12:I	1.89E-02	R1-PM6:I	1.22E-02	R4-PA3:VH	9.06E-03	R2:F	2.72E-03
R3-PA9:VH	1.59E-02	R4-PM6:I	1.22E-02	R4-PM1:V	9.02E-03	R1-PA2:VH	2.66E-03
R2-PA1:VH	1.58E-02	R2-PA4:IH	1.21E-02	relay2_log	8.89E-03	R1-PM10:I	2.66E-03
R4-PA6:IH	1.57E-02	R4:S	1.21E-02	R1-PA9:VH	8.69E-03	R3:F	2.65E-03
R4-PA9:VH	1.56E-02	R3-PM11:I	1.21E-02	R3-PM12:I	8.65E-03	R4-PA8:VH	2.56E-03
R2-PM7:V	1.56E-02	R1-PM8:V	1.18E-02	R3-PM2:V	8.57E-03	R1-PA3:VH	2.56E-03
R3-PA6:IH	1.54E-02	R3-PA5:IH	1.18E-02	R3-PM6:I	7.78E-03	R1-PA:ZH	2.49E-03
R2-PA:Z	1.53E-02	R3-PM10:I	1.17E-02	R3:DF	7.33E-03	R1-PA8:VH	2.39E-03
R3-PM1:V	1.50E-02	R2-PA:ZH	1.16E-02	R4-PA12:IH	6.19E-03	control_panel_log4	8.27E-04
R2-PA7:VH	1.47E-02	R4-PM5:I	1.15E-02	R3:S	6.18E-03	control_panel_log1	6.97E-04
R1-PM12:I	1.39E-02	R1-PM1:V	1.15E-02	R4-PA:ZH	5.92E-03	R2:S	5.34E-04
R2-PM11:I	1.37E-02	R1-PA:Z	1.14E-02	R1-PA5:IH	5.72E-03	control_panel_log3	4.74E-04
R3-PM5:I	1.37E-02	R1-PM5:I	1.14E-02	R4-PA11:IH	5.50E-03	R1-PA1:VH	4.15E-04
R2-PA6:IH	1.36E-02	R2-PM2:V	1.13E-02	R2-PA2:VH	5.38E-03	control_panel_log2	3.77E-04
R4-PM7:V	1.34E-02	R3-PA3:VH	1.12E-02	R2-PA3:VH	5.34E-03	R1-PM3:V	3.65E-04
R4:DF	1.34E-02	R3-PA:ZH	1.10E-02	R4-PM4:I	5.32E-03	R2:DF	3.31E-04
R2-PM5:I	1.34E-02	R1:F	1.09E-02	relay4_log	5.22E-03	R4:F	3.25E-04
R4-PM11:I	1.34E-02	R4-PM12:I	1.08E-02	R1-PA11:IH	5.12E-03	R3-PA:Z	2.29E-04
R1-PM11:I	1.33E-02	R3-PA12:IH	1.08E-02	relay1_log	4.86E-03	R2-PA8:VH	1.89E-04
R3-PM3:V	1.33E-02	R1-PA6:IH	1.07E-02	R3-PM4:I	4.85E-03	R3-PA4:IH	1.77E-04
R3-PA1:VH	1.32E-02	R4-PA5:IH	1.06E-02	R3-PM8:V	4.60E-03	relay3_log	1.39E-04
R1-PM2:V	1.29E-02	R4-PM8:V	1.06E-02	R2-PA12:IH	4.46E-03	R4-PA2:VH	1.12E-04
R2-PM3:V	1.28E-02	R3-PM7:V	1.06E-02	R2-PA10:IH	4.41E-03	R1-PM7:V	9.30E-05
R2-PM6:I	1.28E-02	R2-PM1:V	1.04E-02	R4-PA7:VH	4.35E-03	R1-PA12:IH	7.46E-05
R4-PA1:VH	1.27E-02	R1-PA4:IH	1.01E-02	R3-PA2:VH	3.63E-03	R1-PM9:V	9.29E-06
R3-PA8:VH	1.27E-02	R1-PA10:IH	9.87E-03	R1:S	3.61E-03	R4-PM9:V	3.95E-06
R4-PA:Z	1.27E-02	R2-PA9:VH	9.83E-03	R2-PM4:I	3.48E-03	R3-PM9:V	2.88E-06
R2-PA5:IH	1.26E-02	R4-PM10:I	9.72E-03	R1-PA7:VH	3.46E-03	R2-PM9:V	2.32E-06
R4-PM2:V	1.26E-02	R3-PA7:VH	9.51E-03	R3-PA10:IH	3.46E-03	snort_log1	1.20E-06
R4-PA10:IH	1.24E-02	R4-PA4:IH	9.51E-03	R3-PA11:IH	3.43E-03	snort_log3	0.00E+00
R1-PM4:I	1.24E-02	R2-PM10:I	9.37E-03	R2-PM8:V	3.43E-03	snort_log2	0.00E+00
R1:DF	1.23E-02	R4-PM3:V	9.14E-03	R2-PA11:IH	3.36E-03	snort_log4	0.00E+00

Table 4. Confusion matrix of the test.

Actual label	Predicted label	
	Normal	Faults/attacks
Normal	True negative (TN)	False positive (FP)
Faults/attacks	False negative (FN)	True positive (TP)

Table 5. Confusion matrix results for the top 20% of the features.

Actual label	Predicted label	
	Normal	Faults/attacks
Normal	3664	931
Faults/attacks	943	10138

Table 6. Confusion matrix results for the top 50% of the features.

Actual label	Predicted label	
	Normal	Faults/attacks
Normal	3854	751
Faults/attacks	641	10430

Table 7. Confusion matrix results for the top 80% of the features.

Actual label	Predicted label	
	Normal	Faults/attacks
Normal	3646	922
Faults/attacks	634	10474

Table 8. Confusion matrix results for the top 100% of the features.

Actual label	Predicted label	
	Normal	Faults/attacks
Normal	3568	862
Faults/attacks	758	10488

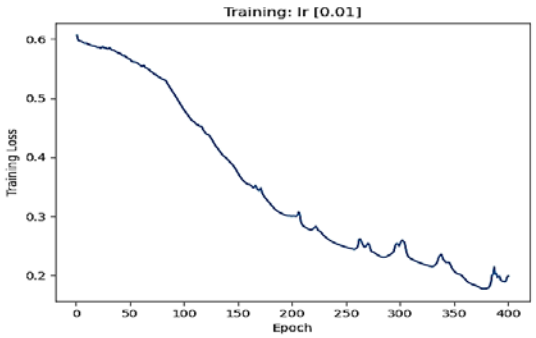
Table 9. Model evaluation.

Set of features	Detection rate (Drate)	Accuracy	False alarm rate (FARate)
20% of top features	91.48%	88.04%	19.45%
50% of top features	94.21%	91.12%	16.30%
80% of top features	94.29%	90.07%	20.18%
All features	93.25%	89.66%	19.45%

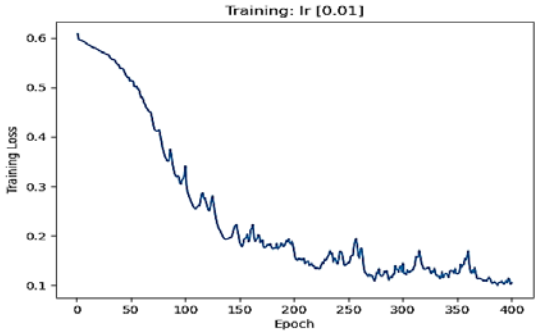
Furthermore, the behaviour of the training is captured by measuring the error at each epoch. This is demonstrated in Figure 4.

The results given provide valuable insights into the effect of feature selection on the efficacy of a machine learning model, as demonstrated by the metrics above—FARate, Accuracy, and Drate. Substantial observations can be drawn from a nuanced examination of these results regarding the correlation between the number of features applied and the model's effectiveness.

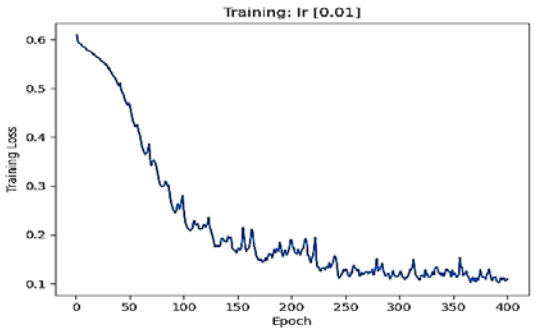
Beginning with the Drate, an insightful trend becomes obvious. As the proportion of the best features in the feature set increases from 20% to 80%, the Drate improves, achieving a maximum of 94.29% with 80% of the features. When every feature is taken into account, however, the Drate decreases marginally to 93.25%. This conclusion suggests that adding additional features improves the model's ability to detect true positives. However, there is a threshold beyond which additional features may not necessarily result in enhanced detection or may even have the opposite effect. The observed outcome may be due to unnecessary or redundant data, limiting the model's capacity.



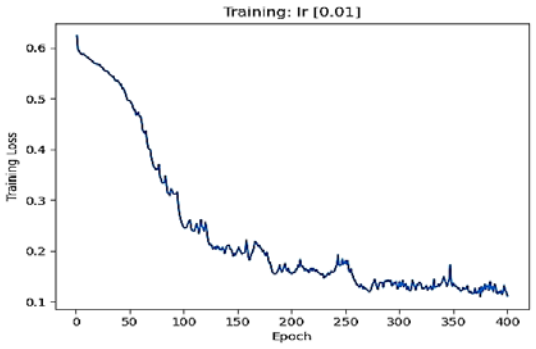
(a)



(b)



(c)



(d)

Figure. 4. Training loss at each epoch where (a) using 20% of the top features, (b) using 50% of the top features, (c) using 80% of the top features, and (d) using all features.

When examining the accuracy of the model, a similar trajectory becomes obvious. The model's accuracy increases from 88.04% when 20% of the features are utilized to 91.12% when half of the top features are implemented. This enhancement suggests that the inclusion of the supplementary features improves the model's overall performance in accurately categorizing positive and negative instances. However, with 80% of the features, the accuracy reduces significantly to 90.07%, and to 89.66% with every feature. The observed reduction as the level of feature inclusion increases could signify overfitting, a phenomenon in which the model becomes overly intricate and begins to attribute learning to noise rather than the true underlying patterns in the data.

The False Alarm Rate provides further clarification. The minimum FARate achieved by the model is 16.30% when 50% of the best features are utilized. The observed increase in FARate at higher feature levels (20.18% with 80% of the features and 19.45% with all features) suggests that the model may be erroneously classifying negative cases as positive due to the inclusion of an excessive number of features. Once more, this increase in false positives at higher feature levels suggests that the model may be inundated with irrelevant or noisy data.

Therefore, these results underscore the significance of optimal feature selection in machine learning models. Metrics indicate optimal performance is achieved when a balanced subset of features (50 per cent of the top features in this case) is implemented. This equilibrium permits the inclusion of sufficient informative data in the model while preventing the introduction of noise or overfitting that occurs with an excessive number of features. The experience of experiencing a decrease in model performance when 80% or all of the features were included should serve as a dismal reminder not to assume that greater data quality results in improved performance automatically. Frequently, the significance and calibre of the data surpass its mere volume. This analysis provides compelling support for utilizing feature reduction methodologies and meticulous feature selection to maximize the efficacy of machine learning models.

It is important to note that the classification method used in this study is a single NN model. In order to make the system adaptive, a homogenous ensemble approach can be utilized. Each time a new threat is introduced, a model can be trained and added to the system. Adding or removing models from an ensemble model allows the system to be more dynamic.

5. CONCLUSION

The importance of the smart grid and the huge data generated by their sensors and devices lead to the urgency of new security frameworks and approaches. Therefore, this paper introduced a new framework that proposes distributed vulnerability detection where close grides are clustered to report their data to the nearest fog node. The paper also presented a deep-learning approach for real-time attack detection. The proposed approach showed that more than 94% of the attacks can be detected using 50% of the features, which indicates the proposed algorithm's efficiency in handling huge data with reasonable performance. The future work involves practically implementing the proposed framework on real smart grid networks.

Although the results presented in this paper are good and with high-accuracy, there is still room for improvement. For instance, more deep learning algorithms could be used for attack detection. Also, due to the large number of features used, more feature reduction is required to reduce real-time processing and computation.

CONFLICT OF INTEREST

The authors have no conflict of Interest.

FUNDING

The authors receive no funding for this paper.

REFERENCES

Bhattacharya, S., Chengoden, R., Srivastava, G., Alazab, M., Javed, A.R., Victor, N., Maddikunta, P.K.R., & Gadekallu, T.R. (2022). Incentive mechanisms for smart grid: State of the art, challenges, open issues, future directions. *Big Data and Cognitive computing*, 6(2), 47.

Borlase, S. (Ed.). (2017). Smart grids: infrastructure, technology, and solutions. CRC press.

Bouramdane, A. A. (2023). Cyberattacks in Smart Grids: Challenges and solving the Multi-Criteria Decision-Making for cybersecurity options, including ones that incorporate artificial intelligence, using an analytical hierarchy process. *Journal of Cybersecurity and Privacy*, 3(4), 662-705.

Clastres, C. (2011). Smart grids: Another step towards competition, energy security and climate change objectives. *Energy policy*, 39(9), 5399-5408.

Ding, J., Qammar, A., Zhang, Z., Karim, A., & Ning, H. (2022). Cyber threats to smart grids: Review, taxonomy, potential solutions, and future directions. *Energies*, 15(18), 6799.

Godbole, S. 2002, "Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers," Annual Progress Report, Indian Institute of Technology–Bombay, India.Guo, Z., Yu, K., Lv, Z., Choo, K. K. R., Shi, P., & Rodrigues, J. J. (2022). Deep federated learning enhanced secure POI microservices for cyber-physical systems. *IEEE Wireless Communications*, 29(2), 22-29.

Habib, A. A., Hasan, M. K., Alkhayyat, A., Islam, S., Sharma, R., & Alkwai, L. M. (2023). False data injection attack in smart grid cyber physical system: Issues, challenges, and future direction. *Computers and Electrical Engineering*, 107, 108638.

Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. *Journal of information security*, 7(3), 129-140.

Hasan, M. K., Habib, A. A., Shukur, Z., Ibrahim, F., Islam, S., & Razzaque, M. A. (2023). Review on cyber-physical and cyber-security system in smart grid: Standards, protocols, constraints, and recommendations. *Journal of Network and Computer Applications*, 209, 103540.

Huda, N. U., Ahmed, I., Adnan, M., Ali, M., & Naeem, F. (2024). Experts and intelligent systems for smart homes' Transformation to Sustainable Smart Cities: A comprehensive review. *Expert Systems with Applications*, 238, 122380.

Kamsamrong, J., Siemers, B., Attarha, S., Lehnhoff, S., Valliou, M., Romanovs, A., Bikovska, J., Peksa, J., Pirta-Dreimane, R., Grabis, J., Kunicina, N., Srebko, J., Vartiainen, T., Eltahawy, B., & Mekkanen, M. (2022). State of the Art Trends and Skill-gaps in Cybersecurity in Smart Grids. *Erasmus+ Strategic Partnership Project*, 2022-04.

Kim, Y., Hakak, S., & Ghorbani, A. (2023). Smart grid security: Attacks and defence techniques. *IET Smart Grid*, 6 (2), 103-123.

Mazhar, T., Irfan, H. M., Haq, I., Ullah, I., Ashraf, M., Shloul, T. A., Ghadi, Y.Y., & Elkamchouchi, D. H. (2023). Analysis of Challenges and Solutions of IoT in Smart Grids Using AI and Machine Learning Techniques: A Review. *Electronics*, 12(1), 242.

Mirzaee, P. H., Shojafar, M., Cruickshank, H., & Tafazolli, R. (2022). Smart grid security and privacy: From conventional to machine learning issues (threats and countermeasures). *IEEE access*, 10, 52922-52954.

Morris, T., & Gao, W. (2014), "Industrial control system traffic data sets for intrusion detection research," *In Critical Infrastructure Protection VIII: 8th IFIP WG 11.10 International Conference*, ICCIP 2014, Arlington, VA, USA, March 17-19, 2014, Revised Selected Papers 8 (pp. 65-78). Springer Berlin Heidelberg.

Murugeshwari, B., Selvaraj, D., Sudharson, K., & Radhika, S. (2023). Data Mining with Privacy Protection Using Precise Elliptical Curve Cryptography. *Intelligent Automation & Soft Computing*, 35(1).

Rohde, F., & Hielscher, S. (2021). Smart grids and institutional change: Emerging contestations between organisations over smart energy transitions. *Energy Research & Social Science*, 74, 101974.

Sifat, M. M. H., Choudhury, S. M., Das, S. K., Ahamed, M. H., Muyeen, S. M., Hasan, M. M., Ali, M., Tasneem, Z., Islam, M., Islam, M., Badal, F., Abhi, S. H., Sarker, S. K., & Das, P. (2023). Towards electric digital twin grid: Technology and framework review. Energy and AI, 11, 100213.

Tuballa, M. L., & Abundo, M. L. (2016). A review of the development of Smart Grid technologies. *Renewable and Sustainable Energy Reviews*, 59, 710-725.

Wasumwa, S. A. (2023). Safeguarding the future: A comprehensive analysis of security measures for smart grids. *World Journal of Advanced Research and Reviews*, 19(1), 847-871.