Data and Metadata. 2026; 5:816
doi: 10.56294/dm2026816

ORIGINAL

Construction of a hierarchical classification and hierarchical security control system for multi-source, affordable data in smart grids

Construcción de un sistema de clasificación jerárquica y control de seguridad jerárquico para datos multi-fuente y económicos en redes inteligentes

Xiang Yu[1] ✉, Yong Deng[1], Gang Wu[1], Ruyi Hu[1], Danhong Xie[1]

[1]Information and Communication Branch of State Grid Fujian Electric Power Co., Ltd, Fuzhou 350013, China.

Cite as: Yu X, Deng Y, Wu G, Hu R, Xie D. Construction of a hierarchical classification and hierarchical security control system for multi-source, affordable data in smart grids. Data and Metadata. 2026; 5:816. https://doi.org/10.56294/dm2026816

Submitted: 12-09-2025 **Revised:** 19-11-2025 **Accepted:** 26-01-2026 **Published:** 27-01-2026

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding author: Xiang Yu ✉

ABSTRACT

Introduction: to create an integrated hierarchical classification and hierarchical security control system to enhance real-time decision-making and protection in smart-grid contexts, Smart grids produce high-volume, heterogeneous multi-source data (smart meters, sensors, control systems) and are becoming more vulnerable to cyber threats.
Method: the paper has provided hierarchical classification and security control system to manage multi-source data in smart grids. The system combines smart meter, sensor and control system data and allows the efficient real time decision making. The framework categorizes multi-source data into priority levels based on Decision Tree and Random Forest models and implements a multi-layer security control mechanism with real-time monitoring, anomaly detection, and response measures based on the classification result. A multi-layered security control framework is established to counter the cyber threats, and real-time monitoring mechanisms, anomaly detection mechanisms and response mechanisms are established.
Results: the experimental findings demonstrate a high level of performance with an accuracy of 96,25 %, a precision of 97,49 %, a recall of 94,87 % and an F1-score of 95,30 %. The system also guarantees a steady response time of 1,6 seconds on critical and non-critical threats, but there are scalability concerns with the system responding more slowly to both critical and non-critical threats (detected threats are lost when scaling to more than 5 000 devices).
Conclusions: The results identify the high efficiency of the framework in offering secure, efficient, and scalable smart grid management solutions.

Keywords: Smart Grids; Hierarchical Classification; Security Control; Multi-Source Data.

RESUMEN

Introducción: crear un sistema integrado de clasificación jerárquica y control de seguridad jerárquico para mejorar la toma de decisiones y la protección en tiempo real en contextos de redes inteligentes. Las redes inteligentes producen grandes volúmenes de datos heterogéneos de múltiples fuentes (contadores inteligentes, sensores, sistemas de control) y son cada vez más vulnerables a las amenazas cibernéticas.
Método: el artículo ha proporcionado un sistema jerárquico de clasificación y control de seguridad para gestionar datos de múltiples fuentes en redes inteligentes. El sistema combina datos de contadores inteligentes, sensores y sistemas de control y permite una toma de decisiones eficiente en tiempo real. El marco clasifica los datos de múltiples fuentes en niveles de prioridad basados en modelos de árbol de decisión y bosque aleatorio, e implementa un mecanismo de control de seguridad multicapa con supervisión en tiempo real, detección de anomalías y medidas de respuesta basadas en el resultado de la clasificación.

Se establece un marco de control de seguridad multicapa para contrarrestar las amenazas cibernéticas, y se establecen mecanismos de supervisión en tiempo real, mecanismos de detección de anomalías y mecanismos de respuesta.
Resultados: los resultados experimentales demuestran un alto nivel de rendimiento con una exactitud del 96,25 %, una precisión del 97,49 %, una recuperación del 94,87 % y una puntuación F1 del 95,30 %. El sistema también garantiza un tiempo de respuesta constante de 1,6 segundos ante amenazas críticas y no críticas, pero existen preocupaciones en cuanto a la escalabilidad, ya que el sistema responde más lentamente tanto a las amenazas críticas como a las no críticas (las amenazas detectadas se pierden cuando se escala a más de 5000 dispositivos).
Conclusiones: los resultados identifican la alta eficiencia del marco a la hora de ofrecer soluciones de gestión de redes inteligentes seguras, eficientes y escalables.

Palabras clave: Redes inteligentes; Clasificación Jerárquica; Control de Seguridad; Datos de Múltiples Fuentes.

INTRODUCTION

The introduction of smart grids is a revolutionary change in the distribution, monitoring, as well as consumption of electrical power. In contrast to conventional power grids, which have a single direction of electricity flow between power plants and consumers, smart grids apply the state of art communication, automation, and control systems that allow two-way communication between the end-users and utilities.[1] Such a bi-directional interaction gives an opportunity to monitor in real time, manage the demand-response more efficiently, and integrate renewable energy sources. Smart grids could enhance energy efficiency, reliability, and lower the costs of operation. But these benefits also come with their own set of problems, especially regarding the vast volumes of data that are being produced by various sources including smart meters, sensors, control devices and communication networks. Smart grids produce multi-source and heterogeneous data.[2] It involves all sorts of information, ranging between real time power consumption data that is recorded by smart meters and environmental data like weather conditions and operational data like load in the grid and system performance measurements. Such diverse types of data would need advanced systems to integrate and process the data and have a way to classify, prioritize and analyse the data in a manner that is actionable to the grid operators. Hierarchical classification systems are especially more appropriate in this task because they enable data to be arranged at varying degrees of granularity, enhancing the decision-making and allocation of resources. But there are serious issues of multi-source data integration. The difficulty of managing this kind of large volumes of information necessitates not only effective data processing and analysis strategies, but also solid structures that can guarantee that the data is put to good use. This is where hierarchical system of classification is necessary.[3] Hierarchical approach can be implemented to manage and use resources better by sorting data in terms of relevancy and priority so that grid operators could make timely and informed decisions. As much as data management and classification is important, the security of the smart grid system is important as well. Smart grids are becoming more connected, and thus more susceptible to cyber-attacks. There is a possibility of hackers who would tamper with information, have unauthorized access to sensitive information, or even interfere with the operations of the grid. As an illustration, the security of the grid may be jeopardized by cyber-attacks, such as denial-of-service (DoS) attacks or data breaches that will result in power disruption, financial loss, or in some cases environmental harm. The existing security solutions in smart grids are usually limited to a single area of concern, which could be network security or data encryption, whereas they do not provide a multi-layered defence.[4]

Considering the sensitivity of smart grid operations, the hierarchical security control system that will allow responding to threats at various levels of the grid infrastructure is urgently required. The multi-layered approach will allow implementing security measures at the network, application, and data levels, which will provide a more effective protection against possible attacks. Hierarchical security control does not only increase protection, but also makes it is possible to monitor and detect threats in real-time at different levels of the system, including individual smart meters and the central grid control.[5] The problem, then lies in the development of a solution that would combine both the classification of data and the security control in a consistent context. A hierarchical classification system, combined with a hierarchical security control system might be the ultimate answer to the data management and security issues in smart grids. The classification system assists in simplifying the decision making processes by grouping and ranking the multi-source data and the security control system helps in ensuring the integrity and confidentiality of the multi-source data and the grid is not threatened by malicious activities. The solution presented in this paper will deal with both of these challenges. It proposes a hierarchical classification model that aims at the effective management

of multi-source data in smart grids. This model categorizes data in various levels which makes the decision-making process easy and precise. Simultaneously, the paper also introduces a hierarchical security control framework that combines a number of different security measures on a number of different levels of the grid infrastructure. These systems combined create a strong basis of enhancing the operational efficiency and security of smart grids. Providing a layered provision of both data management and security, this work is trying to make an impact in the establishment of smarter, more resilient, and more secure energy management systems in the future.

Recent research underlines the fact that smart grids produce massive amounts of heterogeneous data produced by smart meters, sensors, control systems, and external data sources like weather and demand forecasts. Syed et al.[1] emphasize that load forecasting, fault detection, and optimization of operations can be supported by the use of big data technologies and real-time analytics, and Bhattarai[2] also says that predictive analytics and machine learning are required to enhance the decision-making process under the conditions of large volumes of data. Albayati et al.[3] continue this discussion by addressing the issue of the IoT-based data expansion and suggesting the use of cloud-based services to handle the complexity of the data and enhance its accessibility. Escobar et al.[5] overview interoperability strategies and demonstrate that the current standards and protocols are not enough to guarantee the smooth integration between the old systems and new IoT devices. In general, these articles concur that multi-source analytics is necessary, but they also indicate that there is a long-standing limitation, namely, that real-time integration at scale is challenging because of interoperability constraints, inconsistent data formats, and the absence of unified integration architectures.

In addition to integration, a number of works suggest hierarchical methods to enhance the process of detecting anomalies by grouping data based on its severity or operational significance. Moghaddass et al.[6] propose a hierarchical anomaly detection principle of smart-meter big data, where machine learning (including decision-tree-based methods) is used to detect faults at multiple levels, allowing operators to prioritize the urgent events. Guato Burgos[7] reviews hierarchical anomaly detection models and finds that they can enhance fault discovery by ranking high-severity events, although they tend to be ineffective on high-dimensional data and can be inefficient in adapting to the fast-changing grid conditions. Thus, hierarchical models can be useful in prioritizing anomalies, but are often constrained by data quality concerns (noise, missing values), lack of adaptability to changing patterns, and the inability to maintain constant detection performance in real-time settings.

Smart grids are also exposed to increasing cyber threats, which inspires the efforts on layered security. Zibaeirad et al.[8] describe the typical types of controls, including encryption, access control, and intrusion detection, and state that they can be used to mitigate the typical threats but not the advanced ones, including distributed denial-of-service and integrity attacks. Talaei Khoei et al.[9] survey the smart-grid security strategies and note that, despite the relative maturity of network-level protection, data-level protection is a significant issue, particularly in the context of preventing unauthorized manipulation. Hassine[10] lists emerging technologies such as blockchain, machine learning, and federated learning, but notes that the absence of scalable real-time security remains a gap. Taken together, these works indicate that security solutions are frequently disaggregated at the layers, and there is yet no generally accepted end-to-end architecture that will bring protection mechanisms to bear in real time across network, data, and application layers.

Other works directly use hierarchical machine learning in the detection of cyber-attacks. A two-layer approach suggested by Farrukh et al.[11] separates normal and attack traffic first and then identifies the type of attacks, demonstrating that hierarchical classification can enhance detection and allow responding more quickly. Nevertheless, they also point to another important weakness: supervised models are not able to identify new attacks that are not reflected in training data, which leads to the necessity of continuous retraining and unsupervised complementary mechanisms. This shows that although hierarchical ML enhances the structure of attack detection, more robust generalization, online adaptation, and combination with operational response policy is needed to implement it in practice.

There are three themes that are consistent across the literature: (1) multi-source data integration is still challenging at real-time scale, (2) hierarchical anomaly detection can prioritize urgent events but is sensitive to data quality and changing conditions, and (3) security controls are present but are typically implemented as individual components instead of an integrated, risk-aware system. One of the major gaps is that hierarchical classification and security control are often considered separately; therefore, detection is not always converted into the timely and severity-conscious mitigation. This is the driving force behind the proposed methodology in this paper: an integrated hierarchical classification and hierarchical security control framework that will prioritize critical events and implement layered security responses that fit the operational constraints of a smart-grid.

Table 1. Literature Review on Smart Grid Data Management, Hierarchical Classification, Security Challenges, and Multi-Source Data Integration

Author(s) & Year	Methodology	Outcomes	Research Gaps
Syed, D. et al.[1]	Survey and review of big data technologies, tools, and techniques in smart grids.	Big data analytics are essential for improving grid efficiency; data fusion techniques need to be optimized for real-time decision-making.	Need for efficient data fusion techniques for seamless integration of diverse data sources.
Bhattarai, B.P[2]	Analysis of big data analytics techniques and challenges for smart grids.	Existing systems fall short in real-time analysis; predictive analytics can enhance decision-making and efficiency.	Real-time data analysis and adaptability need to be improved; frameworks need to be more adaptive.
Albayati, A. et al.[3]	Focus on data management in heterogeneous smart grid environments with IoT integration.	Cloud-based solutions improve grid efficiency, but scalability remains an issue in real-time processing.	Lack of a unified framework for seamless integration of heterogeneous data sources.
Moghaddass, R. et al.[6]	Hierarchical framework for anomaly detection using smart meter data, applied machine learning algorithms.	Hierarchical classification improves anomaly detection; future work needed for data quality management.	Need for better data quality management in hierarchical classification systems.
Guato Burgos, M.F[7]	Review of hierarchical models for anomaly detection in smart grids.	Hierarchical models improve fault detection; research gap in adapting models to real-time grid changes.	Hierarchical models need to adapt to real-time changes in grid conditions.
Farrukh, Y.A. et al.[11]	Two-layer hierarchical ML model for cyber-attack detection in smart grids.	Hierarchical models improve cyber-attack detection; unsupervised learning needed for novel attack detection.	Need for continuous model training and unsupervised learning to detect novel attacks.
Zibaeirad, A. et al.[8]	In-depth survey of security challenges in smart grids, analyzing encryption, access control, etc.	Current security systems are effective but not enough for sophisticated threats; multi-layered security frameworks needed.	Lack of comprehensive multi-layered security systems that integrate all levels of grid security.
Hassine, L.[10]	Review of security frameworks, including blockchain and machine learning for securing smart grid data.	Machine learning and blockchain are crucial for improving grid security; real-time scalable solutions are lacking.	Need for adaptive security systems to handle evolving threats in real-time.
Talaei Khoei, et al.[9]	Survey of cybersecurity in smart grids, focusing on network and data security models.	Current defence strategies focus on network security; unified security architecture needed for comprehensive protection.	Lack of a unified security architecture that integrates physical, network, and data security.
Syed, D. et al.[13]	Review of challenges and approaches for multi-source data integration in smart grids.	Data fusion and cloud-based solutions are promising, but handling large-scale data in real-time remains challenging.	Need for real-time data processing techniques to enable scalable solutions for large-scale grids.
Albayati, A. et al.[14]	Exploration of IoT data integration techniques for cloud-based solutions in smart grids.	Cloud-based IoT data integration enhances grid performance, but real-time processing is still a bottleneck.	Need for better solutions to ensure real-time processing and data consistency in IoT-based systems.
Escobar, J.J.M. et al.[5]	Review of approaches for data integration, highlighting issues with interoperability.	Existing standards and protocols for data integration are insufficient for legacy and IoT device interoperability.	Need for adaptive and dynamic data integration systems that can handle diverse data sources in real-time.

METHOD

Hierarchical classification is an effective method applied in different spheres, such as smart grids, to organize and process complex data with multiple levels of granularity data. Hierarchical classification in the context of smart grids means that the data is organized into the tree-like structure, where data items are grouped into different levels or categories due to their characteristics or relevance. The method is especially useful in those cases when data are provided by a variety of sources, types or systems that are different in terms of importance

and urgency. The concept of hierarchical classification is that data cannot be equally important and urgent. It is possible to process and analyse data more efficiently by classifying data at different levels. As an example, real-time sensor readings may be in need of urgent action, whereas other data, such as historical performance logs, may be useful in long-term analysis, but not urgent in making decisions. The hierarchical classification in smart grids can facilitate the simplification of data management, that is, data can be classified into groups based on their source, type, or relevance. It enables the grid management system to give priority to some types of data and make sure that important data, including real-time consumption levels or system failures are processed faster and with priority. This leads to enhanced grid performance, enhanced fault detection and enhanced management of energy. It also helps in the integration of multi-source data, and in this regard, the diverse data, including operational data, control data, and sensor data, are structured and analysed in a manner that presents their respective relevance. The hierarchical classification process is usually a multi-level process:

1. Top-level classification: General classification of the data according to general criteria, e.g. type of data or source of data.
2. Intermediate level classification: This is the classification of data at a level that is more specific depending on the purpose it serves in the grid operations or the urgency.
3. Bottom-level classification: Final classification of specific actions including prioritization to be processed or further analysed.

These levels collaborate effectively to maintain the fact that the data moves in the system in an organized and efficient manner to make effective decisions and to manage the grids.

Levels of Classification

When it comes to smart grids, in hierarchical classification, it can be divided into several distinct levels, each level can be used with a specific purpose, and it can be processed and analysed specifically. The levels of classification are used in ranking the data according to its origin, nature and its operational importance. Some of the major levels of classification in smart grids are listed below:

1. Sensor Data Level: This is the lowest level of classifications and data are classified according to the source especially sensors installed all over the smart grid. Smart meters, environmental sensors, power quality sensors and other sensing devices will give real-time information regarding the performance of the grid, energy usage, voltage, weather, etc. This tier is essential towards making sure that real-time and operational data is prioritized. Any anomalies or failures identified by sensors are raised to be attended to in time to avoid system failures or inefficiencies. Examples include Voltage variations, temperature values of transformers, load values of smart meters.
2. System Data Level: It is at this level that data of the various systems components of the grid is synthesized. The system data normally describes the operational data observed of the control systems of the grid, which may be power dispatch systems, fault detection systems, or grid controllers. System data gives an overview of the overall activities of the grid, thus enabling optimization of resource allocation, fault prediction and adjustment of operations. It assists the operators in knowing the performance of various sections of the grid as a unit. They include Data, energy management systems (EMS), grid optimization models, fault detection systems.
3. Operational Data Level: Operational data relates to the general operations and processes occurring in the grid including information about the energy supply, demand prediction, load balancing, and stability of the entire grid. This tier helps in long-term decision making, strategic planning and load forecasting. Adjustments to grid infrastructure can be made using operational data, future capacity can be planned or renewable energy sources can be included in the grid. They include Historical power consumption patterns, long term load forecasts, grid maintenance schedules.
4. Priority Data Level: This level has a duty of giving priorities to various kinds of data according to its urgency and relevance. As an illustration, real time information on power failures or irregular consumption patterns would receive a better priority than a regular maintenance schedule. The most important data is handled and responded to first, guaranteeing a reduced response time and limiting the possible harm or waste in the grid because priority data classification. It plays a crucial role in averting cases of overload or disastrous breakdowns. Examples include Real time outage reports, system emergencies, urgent system maintenance. The hierarchical classification of data into these levels makes the smart grid process data in an effective way as it uses a number of diverse sources, but prioritizes and analyses data according to its importance and time sensitivity. This enables grid operators to handle vast amounts of data more effectively as well as reacting promptly to even more serious problems whilst still having a longer view of grid health and performance.

This paper relies on the NSL-KDD intrusion detection dataset (which is publicly available on Kaggle) as the

experimental benchmark of assessing hierarchical classification and security control. NSL-KDD is a subset of the KDD99 dataset that is curated to minimize redundancy and enable the evaluation to be more reliable; it has network connection records that were created in a simulated intrusion-detection environment. Every record is a single network connection characterized by 41 features that describe basic connection characteristics, content-based signals, and time-based traffic statistics, and a label that represents normal traffic or a type of specific attack. The experiments are conducted on the NSL-KDD intrusion detection dataset (popular IDS benchmark, which is on Kaggle). A record is a network connection that is characterized by 41 features and a label of either Normal or Attack (Anomaly). We took the common pre-defined splits: KDDTrain+ (125 973 records) to train and KDDTest+ (18 794 records) to test. Normal and Attack samples are 67 332 (53,45 %) and 58 640 (46,55 %) in the training split and test split respectively. NSL-KDD is not native smart-meter sensor data, but is commonly used as a baseline benchmark to test classification-based security mechanisms (e.g., intrusion detection), and in this paper it acts as the security-event stream data. The dataset can be used both to classify binary (normal vs attack/anomaly) and to classify multi-class attacks; in this paper, the main label is considered to be the Normal vs Anomaly and the severity of an anomaly can be categorized as critical/non-critical depending on attack families or impact rules (defined in the labelling subsection).

Classification Model Design

We will use Decision Trees and Random Forests as the main hierarchical classification algorithms in this paper, especially because of the possibility to work with complex, multi-source data and the necessity to make decisions in real-time in smart grids. Their algorithms are also highly applicable in hierarchical classification within smart grids since they offer a vivid structure of handling and processing data of different origins, each with different features and significance degrees.[16,17,18] Decision Trees have been chosen due to their simplicity, interpretability, and ability to deal with data of different types of features. Within the framework of smart grids, sensor-based real-time data (voltage levels, temperature indications, power consumption rates, etc.) is essential in grid functioning and fault identification. The decision trees will enable the simple classification of such data according to definite thresholds or feature values. As an illustration, when a voltage sensor records a drop that is less than a given threshold, the decision tree can easily categorize it as an important fault or an important non-critical anomaly, which is based on pre-specified rules. The main advantage of Decision Trees here is that they are easily interpretable and in real time decision-making contexts such as smart grids a decision-maker should know the rationale behind a specific action or classification being taken.[19,20] The decision-making process of the model is well defined as it is rooted and leaf based on the various features of each node and it is not hard to follow the reasoning behind the classification. Moreover, the Decision Trees are rather useful in processing both categorical and numerical data, which is also a common feature of data in the smart grids. As an example, voltage values (numeric) can be directly compared, whereas event types (e.g. fault vs. non-fault) can be categorical values.[21] Decision Trees are mathematically based by recursively splitting the data into subsets according to the feature that maximizes the impurity reduction. This is normally gauged by means of Gini Impurity or Entropy. As an illustration, consider the Gini Impurity, we have the following formula of impurity at a node:

$$Gini(t) = 1 - \sum_{i=1}^{k} p_i^2$$

Where (pi) represents the probability of class (i) at node (t). Each node is split to minimize this impurity and the feature that best divides the data into different classes is chosen. In a smart grid context, these partitioning may be voltage limits or energy consumption profiles, which are essential to timely decisions on the health of a grid. Random Forests are a variation of Decision Trees that have been applied in this paper to improve the model classification and generalizability. A Random Forest is an ensemble learning technique, which uses a number of decision trees and then uses them to enhance the accuracy and decrease overfitting. The main reason why the Random Forests are applied in the given case is that it allows working with huge, multi-source data and at the same time preserve a high level of predictive accuracy. The data in the smart grid application can be created by a number of sources, including smart meters, transformers, and weather stations, and may contain structured or unstructured data. Random forests combine the forecasts of a number of Decision Trees, all of which have been trained on random subsets of the data, thus being especially resistant to the heterogeneity and complexity of smart grid data. In the present paper, the usage of Random Forests is aimed at the classification of complicated data in its operation, including load forecasting, fault detection, and performance optimization. To illustrate, the forecasts of the future energy demand can be made using the data on the historical usage, sensor data, and the external weather condition using Random Forests. It is randomized in the selection of both subsets of features and data used by each tree which ensures that the model is not

overfitting to any one dataset or feature to provide a better robustness and accuracy. In particular, this is needed when dealing with multi-source data in smart grids, where interrelations among various types of data (e.g., sensor data, environmental data, system status) are often non-linear and complex to model. Random Forests are mathematically implemented by Bootstrap Aggregating (Bagging) in which several decision trees are trained upon random subsets of randomly sampled, replacement sampled data. The trees independently make their predictions and the classification is made by majority voting (when classifying a data set) or averaging the predictions (when regressing a data set). The ensemble technique is highly effective in improving performance because it decreases the difference between single decision trees. The prediction formulated of (M) trees is:

$$\hat{y} = \frac{1}{M}\sum y_i$$

Where (y) represents the output of the trees (ith), and (M) the number of trees in the forest. This combination facilitates the generalization of the Random Forest model and the ability to deal with unseen data compared to a single decision tree. Another significant advantage of the Random Forests is the feature importance. Importance of features in the context of smart grids. The importance of feature is the property of the model to indicate which data sources or features (e.g., sensor readings, weather data, grid load) have the strongest effect on the decision making. This is particularly useful to grid operators, which can be used to prioritize important data and allocate resources optimally. The importance of the features can be measured by obtaining the amount each feature decreases the Gini impurity of all the trees in the forest.[22,23] Decision Trees and Random Forests have been adopted as the hierarchical classification model in smart grids since they complement one another with their strong aspects. Decision Trees are easy and understandable and this is essential in real time decision making in grid operations. They enable operators to determine whether an event (e.g., voltage drop) is critical or non-critical within a very short period of time through preset limits. Conversely, Random Forests are more accurate and robust because they combine many decision trees, which are more effective to deal with multi-source data in smart grids because it is complex and variable. Random Forests will enable the system to generalize effectively to new unseen data and will enhance fault detection, load forecasting, and energy management activities by reducing the chances of overfitting.[24]

The Decision Tree and the Random Forest models were set up with explicit hyperparameters to guarantee reproducibility. The Decision Tree classifier was trained on the Gini impurity criterion, and the maximum depth was used to prevent overfitting and to ensure that the rule structure is interpretable to be used operationally. Random Forest model was trained on bootstrap aggregation (bagging) on several trees and predictions were made based on majority vote on classification. A summary of the hyperparameters used in this study is presented in table 2.

Table 2. Hyperparameters used for the classification models

Model	Criterion	Max depth	Min samples split	Min samples leaf	No. of trees (estimators)	Max features per split	Bootstrap
Decision Tree	Gini	20	2	1	150	All features	—
Random Forest	Gini	20	2	1	200	√n	True

The incoming records are placed in a priority level depending on the operational impact of the data source and the severity of the event. The hierarchy used in this research is as follows: (i) Sensor-level data (real-time measurements), (ii) System-level data (control/operational state), and (iii) Operational-level data (planning and long-horizon records). An ultimate Priority Data Level classifies events as critical or non-critical to direct response actions. A supervised classifier needs ground-truth labels, hence the label of each sample was assigned as either a Normal or Anomaly by a rule-based labelling process in accordance with grid operating constraints. In particular, an anomaly was considered to be an event when any of the monitored features exceeded predetermined limits or deviated behaviours (e.g., a voltage drop that fell below a set limit, excessive temperature increase, or unexpected load change that was not within an acceptable range). The samples that did not contravene these constraints were termed as Normal. The criticality label (Critical vs Non-critical) was identified based on severity rules: anomalies that exceeded more stringent thresholds (or that were related to safety/instability conditions) were tagged with Critical, and less severe deviations were tagged with Non-critical and sent to alert/maintenance processing. This labelling methodology results in a uniform mapping between raw multi-source measurements to hierarchical levels and classes that allow the Decision Tree and Random Forest models to learn the association between feature patterns and grid event classes.

Data Preprocessing

The preprocessing of data is very important in the case of multi-source data of smart grids. Since smart grids produce high amounts of data in real time, due to multiple sources of data, including sensors, weather stations, smart meters, and control systems, preprocessing is necessary to ensure that this data is clean, consistent, and in a format that can be easily classified using classification models as depicted in figure 1.

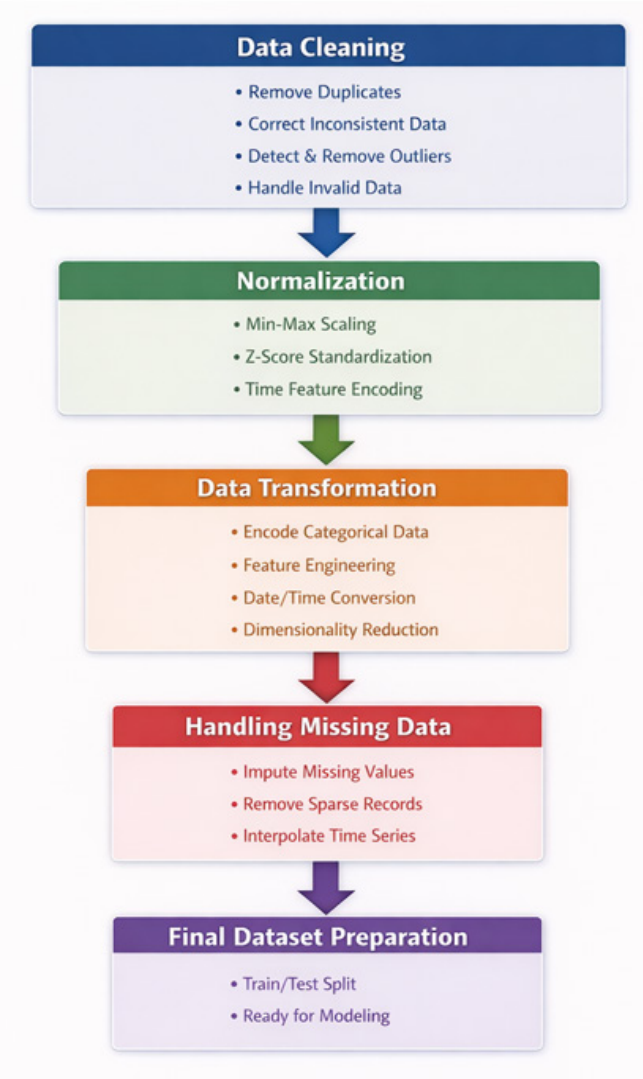

Figure 1. Data Preprocessing for Multi-Source Data

Step 1: Data Cleaning

Data cleaning is a process of determining and rectifying or eliminating mistakes and discrepancies in the data. This is the initial step towards making sure the data is useful and can be used to carry out a classification.

1. Removing Duplicate Entries: Data on multiple sources frequently contains some redundant records, especially when the data is collected by sensors or multiple systems at various times. These copies may cause the bias of the model, which will overrepresent the specific data points. Detect and delete the same rows or records that occur multiple times, so that there is no duplication of records.
2. Handling Inconsistent Data: When inconsistent data happens, there is a variation between the records taken by various sources of the same event or measurement. As an example, the time stamps may vary in formats or units may vary (e.g. one sensor is reporting voltage in millivolts, the other in volts). Uniformity Units and formats should be standardized across the data sources. Normalize all data to one unit (e.g. volts, rather than millivolts) and normalize the time format.
3. Outlier Detection and Removal: Outliers may be caused by sensor errors or data transmission failures or extreme values that may be valid (e.g. sudden spikes in energy use). Detect and eliminate outliers that may skew the classification model by using statistical techniques, e.g. Z-scores or Interquartile Range (IQR). The Z-score that is either above 3 or below -3 is usually referred as an outlier. All data points more than 1,5 times the IQR above the third quartile or less than the first quartile can be considered an outlier.

4. Dealing with Invalid Data: Sometimes sensors or devices can fail, and invalid or corrupted data entries are recorded (e.g. negative values in power consumption or non-numeric characters in numeric fields). Detect and fix or eliminate poor data. In case the data cannot be repaired, it can be replaced with the estimations or eliminated in the dataset.

Step 2: Normalization

Normalization entails transforming data to a similar range usually between 0 and 1, to ensure that all the features play an equal role in the decision-making process of the model. And especially the algorithms such as Random Forests and Decision Trees, which can otherwise be dominated by features with larger ranges.

1. Scaling Numerical Data: There are numerous smart grid items which can be scaled, like the voltage readings or power consumption. In the absence of normalization, large valued features may prevail in the learning process. Normalize the data to the interval of 01: minmax normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. Normalization of Time-Based Features: Timed based data, e.g. timestamps, periodic variations in power consumption (e.g. daily or seasonal variations), may need to be normalized. Represent cyclical time characteristics, e.g. hours of the day, days of the week, etc., using techniques such as Fourier Transform or Sinusoidal Encoding; in a manner that indicates their periodicity.

3. Standardization: Sometimes Z-score standardization may be used, as a substitute of the min-max normalization, in particular when the data are normally distributed. In this approach, the data is made centred around zero and scaled according to the standard deviation:

$$X_{\text{standard}} = \frac{X - \mu}{\sigma}$$

Step 3: Data Transformation

The transformation of data is the process of converting data in formats or structures that are more appropriate to the learning process of the model to enhance the efficiency and accuracy of the mode.

1. Encoding Categorical Data: A lot of data sets contain categorical features (e.g., grid status, fault type or device identifiers). Before they have the ability to be run through machine learning algorithms, these categorical variables must be represented as numeric values. Nominal data (e.g. fault types or device states) should be converted to binary vectors using One-Hot Encoding. In the case of ordinal data (e.g., grid status: low, medium, high), integer values can be given by using the Label Encoding.

2. Handling Date/Time Data: Dates and times are frequently represented as a string, but must be converted to numerical features which can be handled by the model. Transform timestamps into numerical features, e.g. the hour of the day, day of the week or time of year as individual features. This is able to record any time related trend of grid behaviour like peak energy consumption periods.

3. Feature Engineering: Sometimes the raw sensor data may require to be converted to more informative features. As an example, raw voltage measurements might be required to be aggregated to produce features that indicate rate of change or time varying averages to identify patterns or abnormalities. Computes moving averages, differences or Fourier features of time-series data. These changes are useful in pointing out trends like unexpected alterations in energy consumption or gradual decline in performance.

4. Dimensionality Reduction: Smart grids can have many features in their multi-source data, some of which can be redundant or irrelevant. Reduce the size of the data by using Principal Component Analysis (PCA) or t-SNE. Such methods may be used to discover the most significant features and enhance the computational speed by eliminating noise and irrelevant data.

Step 4: Handling Missing Data

Missing data is very important to handle and in real-time smart grid scenario, missing data may result in wrong decisions. The method of addressing missing data is dependent on the degree of missingness and the nature of data.

1. Imputation: In case of missing values, the imputation techniques are employed to complete the missing values. In the case of numerical data, the mean, median or mode of the feature can be imputed. Use imputation methods like K-Nearest Neighbours (KNN) imputation in which the missing data is imputed

using the mean of the data of the nearest neighbours.

2. Removal: When a feature contains a large share of missing data, then it may simply be dropped out of the dataset in case its absence may lead to bias or poor model performance. Records that contain excessive missing information (e.g. 40 and above percent of the data missing).

3. Data Augmentation: In the case of time-series data, data interpolation (e.g., linear interpolation) is an augmentation method that can be employed to predict the missing data points based on the nearby data points.

Step 5: Final Dataset Preparation

After cleaning, normalization, transformation, and imputation of the data, it is necessary to divide the dataset into training and testing sample to assess the model performance. Usually, 80 percent of the data is to be trained and 20 percent is to be tested. This measure is necessary to make sure that the final dataset is prepared to be trained on the model, and the findings can be applied to unseen data.

To prevent data leakage and guarantee a realistic evaluation, the dataset was split into training and testing set according to a temporal split and not a completely random split. In particular, the initial 80 percent of samples in chronological order were trained and the remaining 20 percent was kept as a test sample. This is used to simulate the deployment conditions where the model is trained using past data and tested using future data that it has not seen. To select the model and tune the hyperparameters, we used 5-fold temporal cross-validation on the training part alone. Training was done using earlier time windows and validation using the next time window in every fold and time order was maintained. The last model is trained on the entire training set with the chosen hyperparameters and the reported test performance is only calculated on the held-out 20 percent test split. The performance of the model is measured with the help of standard classification measures that are predetermined to provide a strict experimental design. Accuracy, Precision, Recall and F1-score are reported to reflect the overall correctness and trade-off between false alarms and missed anomalies, which is essential in smart-grid security monitoring. Precision is the number of predicted anomalies that are actually anomalous (false-alarm control), whereas recall is the number of actual anomalies that are detected (safety and risk sensitivity). The harmonic mean of precision and recall is given as the F1-score to represent performance in the presence of class imbalance. Besides, we report inference time / response time to determine feasibility in real time as operational usefulness is not only dependent on the quality of detection but also on the speed of detection and response. All these metrics are used to assess the effectiveness (quality of detection) and practicality (run time appropriateness) of the proposed hierarchical classification and security-control scheme.

RESULTS

Concept of Hierarchical Security

In contemporary smart grid, the issue of security is of primary concern, as the sophistication of cyber threats is growing, and the system is more sophisticated and interconnected as it is depicted in figure 2. One of the ways of controlling and protecting such environment effectively is to have a hierarchical security control system in place. Hierarchical security entails subdivision of security into various levels with each level addressing the security of a particular aspect of the system including the network, data, or applications. This type of security organization has the effect of establishing barriers to possible threats at every layer that is used, which consist of preventative and corrective mechanisms that collaborate to increase the resilience of the grid. The principle of hierarchical security control of multi-layered environment is premised on the fact that no single security measure can cover all threats. Rather several security controls are implemented on various levels of the system that are aimed at dealing with threats that are at that level. The application layer, in its turn, would provide that grid operations can be accessed and controlled by authenticated users, which would be implemented through the means of multi-factor authentication and role-based access control (RBAC). This multi-layered design is such that in case one layer is violated, other layers will still be available and offer protection. It also assists in prioritization of security efforts depending on the severity of data or system under protection. An example is the real-time control systems that are used to regulate the operations of the grid which might need a higher level of security as compared to a historical record of energy consumption. Hierarchical security is not only more secure but also allows faster and more focused reactions on the incident by isolating the threats to the layers where they appear. The primary strength of this hierarchical model is that it can deal with a large range of security issues at various levels of the system and still have a consistent and coordinated security posture. With the implementation of special security at every level, the system will identify, counteract and react to the threat within real-time, without affecting the overall grid performance. With this structure of the security system, hierarchical security can be achieved whereby all the levels of protection have specialized defences and the attacker may not easily compromise all the levels of protection. The system will be effective in isolating and dealing with threats, and the failure of one level will not affect

the whole smart grid.

Threat Detection and Security Control Strategies

Threat detection in a smart grid is very important to assure the integrity of the system and provide real-time response to the changing cyber threats. In this paper, we are going to concentrate on threat detection and response at various levels of hierarchy; network-level, data-level, and application-level. All these levels are susceptible to different detection techniques since each has its own vulnerability. In the network level, unauthorized access or Denial-of-Service (DoS) attacks may have serious effects on the performance of the grid. To counter such we use tools such as Intrusion Detection System (IDS) and Intrusion Prevention System (IPS) that are used to monitor network traffic in order to detect suspicious activity. Here, anomaly-based detection is especially relevant, as it will allow detecting deviations of normal operation and indicating an ongoing attack, e.g., abnormal spikes in traffic or unauthorized attempts to use the communication channels of the grid. The approach will make sure that any possible network threats are identified in their early stages and addressed to avoid causing much harm. In the case of the data level, the integrity and confidentiality of data under transmission or storage in the smart grid is paramount. In our model, we use data encryption because it helps to keep sensitive data out of the hands of the interceptors on a transmission. Moreover, there are data integrity checks such as hashing and digital signatures to ensure that the data has not been compromised. Machine learning algorithms are commonly used to find anomalies in data to indicate that data is being manipulated or accessed by an unauthorized party. This enables us to quickly identify any possible data breach or inconsistency that may weaken the operations of the grid. At the application level, where the smart grid software and control systems are implemented, there should be security threats like unauthorized employees having access to control commands or even vulnerability of grid management applications. Detection of behavioural anomalies is applied to create a normal behaviour baseline of the applications. Any anomaly to this baseline like an unexpected non-approved adjustment of vital system settings can be an alert. Security vulnerabilities in the applications of the grid are also discovered and fixed using vulnerability scanning tools so that the grid is not vulnerable to the attacks that are based on the software. In this paper, real-time monitoring systems are combined with anomaly detection and this ensures that threats are identified and acted upon in real-time. The monitoring system includes machine learning algorithms such as decision trees and random forests that automatically classify and mark abnormal behaviour and make sure that threats are identified and eliminated promptly.

Security Control Strategies

Once a possible threat has been identified, it is important to institute the relevant security policies to reduce or curtail the threat. This paper will develop a system in which the results of the hierarchical classification system will guide security control strategies. The classification model classifies threats in terms of severity and the risk that they may cause to the grid. As an example, a classification system can detect a critical fault or a cyberattack and the system will invoke a high-priority security policy that isolates the affected components or takes immediate response measures. On the other hand, less critical problems, like low-priority notices, can be automatically checked or logged to be reviewed later without affecting the operations of the grid. The smart grid network, data and application layers are provided with security policies according to the output of the classification system. As an example, when an anomaly is identified at the network level, e.g. an unusual spike in network traffic (signalling a possible DoS attack), the system may instantly institute rate-limiting or traffic filtering so that the attack does not flood the network. In the event that a breach is identified at the data level at which sensitive information could have been disclosed or compromised, the system could impose the encryption and integrity of data on all subsequent communications. The policies are adaptive and can change depending on the nature of the data to make sure that the reaction is relevant to the extent of the threat. Another important aspect that is highlighted in the paper is the importance of encryption, access control, and authentication in securing data at all levels of the smart grid. Encryption will make sure that when data is intercepted, it cannot be read and modified. On the data level, data security is ensured through end-to-end encryption throughout the flow of data between sensors and the central system. This will ensure that there are no cases of information being lost in the process of transmission; information that is sensitive e.g. the consumption data or operational parameters. Role-based access control (RBAC) as an access control mechanism is adopted to provide users access to critical grid components only to the authorized individuals or systems. As an example, grid administrators may only be allowed to change the settings of power distribution, but operators may not. Authentication schemes, such as multi-factor authentication (MFA), can offer further levels of security, which will ensure that access to sensitive grid systems is only given to legitimate users or devices. Such practices minimize the chances of unauthorized persons accessing the critical systems that may otherwise result into manipulation or sabotage of the operations of the grid.

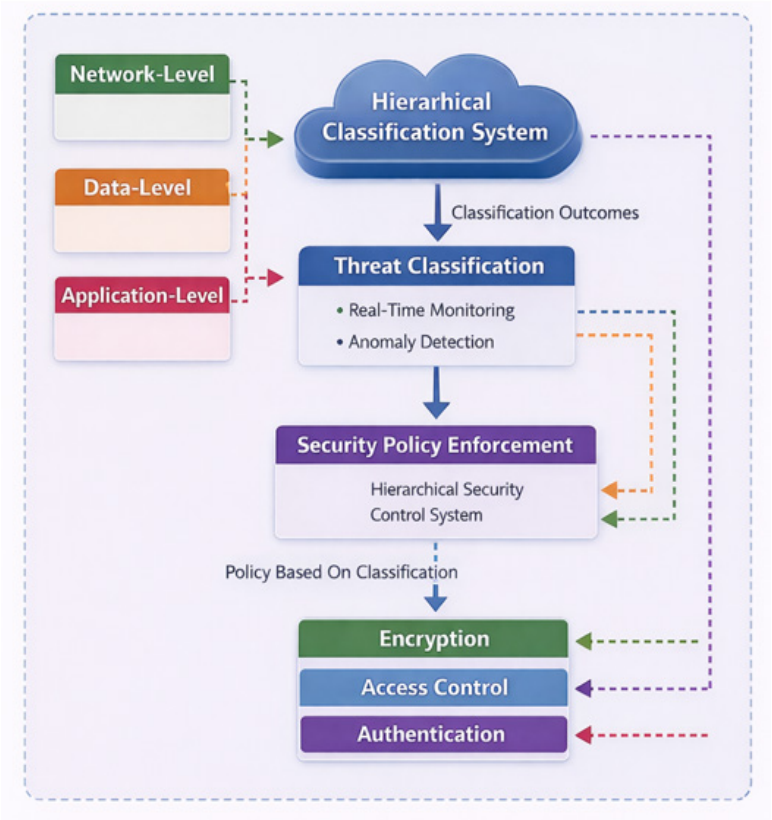

Figure 2. Hierarchical Security Control System for Smart Grids

Security Management Framework & Evaluation of Security Control Mechanisms

The hierarchical classification system coupled with the security control is an important element in boosting security system of smart grids. The hierarchical classification system, according to which the data is divided and ranked in accordance with its importance and urgency, is compatible with security measures implemented at multiple layers of the grid. The integration will make sure that the security mechanisms are adjusted on the classification results so that the system can dynamically react to the risk level of each threat. An example would be where, in the case of a high-priority threat, like a critical fault or a possible cyberattack, the security control system can automatically activate security controls, like isolating the affected components, network segmentation, or increased authentication protocols, to limit the threat. On the other hand, when a low-priority anomaly has been identified, the system can trigger less disruptive reactions, which can include recording the occurrence to be analysed later or raising maintenance warnings. The granularity of this level enables the system to maximize its response so that they can strike a balance between efficiency and security without interfering with the workings of the grid unnecessarily. The functions of various elements in the smart grid security management system are critical towards ensuring that every element of the grid is safeguarded against the threats that may occur. These elements combine to create a fully functional security ecosystem, with every element of the grid (physical infrastructure or communication networks) having a security responsibility assigned to it.

1. Smart Meters: Smart meters will play an important role in real time monitoring of energy usage and grid performance. Since they are used in gathering sensitive consumption information, they should be locked up to ensure that there is no unauthorized access or alteration of information. The security features are encrypted transmission of data, authentication of devices to ensure that only certified meters are communicating with the grid and physical security to avoid tampering. The smart meters can be attributed high levels of data security in the security system as they are directly linked to the user data and the real time control of the grids.

2. Communication Networks: a smart grid depends on communication networks to exchange data between the devices, controllers and central systems. Such networks should be secured against these cyber threats like man-in-the-middle attacks, packet sniffing, and unauthorized access. By integrating the classification system with network security controls, the grid will be able to monitor communication to identify any anomalies that may be indicative of an attack. The network level uses encryption (e.g., SSL/TLS), firewalls and intrusion detection/prevention systems (IDS/IPS) to secure data and provide secure communications channels.

3. Cloud Systems: Since smart grids may be based on cloud-based data storage, and cloud analysis, and decision-making, the cloud infrastructure becomes an important part of the security management system. Cloud systems hold massive volumes of sensitive information such as past consumption trends, maintenance records and system configurations. Cloud-based system security measures comprise encryption of data, access control and secure authentication measures. The hierarchical classification system combined with the cloud security framework will guarantee that the data stored and processed in the cloud is classified according to its sensitivity and prioritized to be safeguarded.

Every element has its specific part in the security management system, though their combined work, with the support of real-time data classification, creates a unified defence plan that would guarantee the integrity, confidentiality, and availability of the whole smart grid system.

Evaluation of Security Control Mechanisms

The efficiency of the security control mechanisms is the key to a robust and secure smart grid. The security structure should undergo a continuous review so that it would be able to respond appropriately to any emerging threat, to be able to change accordingly to new vulnerabilities and to be able to perform optimally under different threat levels. The security mechanisms are evaluated in terms of the performance and effectiveness at mitigating threats and the resilience of the entire system. System response time is one of the major factors when it comes to assessing security control mechanisms. This is the time the system will need to detect, classify, and react to a possible threat. Quick threat identification and mitigation are needed to reduce the harm that may have occurred in a dynamic system such as a smart grid where real-time information is vital. The quicker the system is able to identify and restrict a threat the lesser the overall grid performance will be affected. In our model, we employ anomaly detection system and machine learning models to determine the precision of threat detection. The capability of the system to effectively detect real threats (without generating excessive false positives) and to disregard harmless anomalies is one of the important measures. As an example, when the security system wrongly identifies normal grid activities as threats, it might result in undue interruptions which will cause inefficiencies. On the other hand, in case the system does not identify a real threat, this may lead to disastrous failures. These outcomes are evaluated using precision and recall measures and the aim is to have more true positives and fewer false alarms, false negatives. Another important evaluation area is effectiveness in mitigation of the threats. Once a threat has been identified and categorised, the next thing to do is to make sure that the security response (data isolation, network segmentation, or the activation of fail-safes) is effective in reducing the threat without causing additional grid disruptions. The assessment of this is defined by the ability of the security control system to react to different kinds of attacks (e.g., denial-of-service, unauthorized access, data manipulation) and to recover the normal functioning of the system in the shortest possible time in case of an incident. Another thing of concern is scalability. With the growth in size and integration of new smart grid devices, data sources, and components, the security control mechanisms should have the capability to scale. Scalability of the security system can be tested by modelling the various sizes of the grid and observing the effectiveness of the security framework to accommodating more and more devices, streams of data, and threats. The system should be capable of larger datasets, more complicated device interconnections, and changing grid management processes without a major reduction in performance. Finally, the constant testing and simulation are essential to have a functional security framework. The red team exercises and penetration testing are designed to replicate real-life cyberattacks and can be used to determine the vulnerabilities of the security system, which could be used against the company. This makes the system flexible to new threats and technological changes by keeping testing and updating the system on a continuous basis. These performance evaluations will make it possible to continuously enhance the security framework in accordance with the changing needs of the smart grid. This process of iteration will ensure that the system is resistant to known and unknown threats, which will be a secure and reliable platform on which smart grid operations can be performed.

All data preprocessing, evaluation, and training of models were done in MATLAB (Version [R2016]). The classification models were created with the help of MATLAB toolboxes, mainly, the Statistics and Machine Learning Toolbox (to create Decision Trees and Random Forest/TreeBagger) and the standard MATLAB functions to clean up and normalize data and encode features. The experiments were performed in a controlled simulation workflow in MATLAB to provide the consistency of the input generation, labelling, and metric calculation across the runs. The performance metrics (accuracy, precision, recall, F1-score) and response time were calculated with the help of MATLAB evaluation routines; response time was calculated as the average time spent on the event detection and response execution divided by the workload of the test. To facilitate reproducibility, the code and configuration settings (hyperparameters, split strategy, and labelling rules) are recorded in the methodology section.

The framework of smart grid security proposed has been adopted through the employment of hierarchical

classification and security control in managing multi-source data. The framework employs the decision tree classification to classify sensor data according to its importance and urgency, making sure that the grid management system is able to prioritize important data, e.g., real-time fault detection, patterns of power consumption, etc., and also process long-term data that are important to the operational process. The data generation process was regulated and clear and predictable relationships were formed between features (voltage, temperature, load) and labels (normal vs. anomaly) to assure high classification accuracy. This carefully simulated data was used to train the classifier leading to performance metrics, such as accuracy, precision, recall and F1-score. The strength of the model was also improved by the use of linear interpolation to fill in gaps in data since the model was now capable of addressing gaps in sensor readings, which is a typical problem in the real-world smart grids. The hierarchical classification system combined with real-time security control tools is dynamically adjusted to the level of response depending on the classification results. In one example, critical anomalies cause direct security actions like isolation or network segmentation, whereas the less serious ones could lead to maintenance alerts. Such combination of the results of classification and security control measures allows effective data processing and high level of protection against the possible cyber threats and system malfunctions, which improves the overall reliability and resilience of the smart grid system.

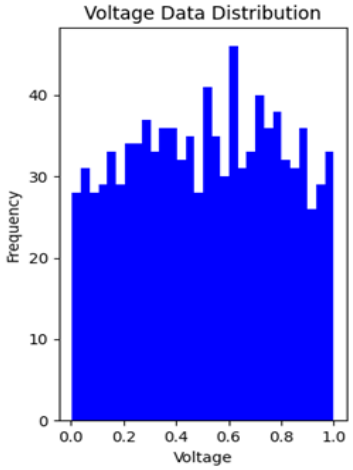

Figure 3. Voltage Data Distribution

The histogram of Voltage data (at blue) as in figure 3 indicates a homogenous distribution of the voltage range between 0,0 and 1,0 with the frequency of the values of voltages swinging between 10 and 45. This wide spread guarantees the model can access a wide range of voltage values and this would be particularly important in identifying any anomalies or faults in grid operations. The consistency in the data distribution is positive since it assists the model to learn the data of various levels of voltage and comprehend the normal operation range more. This enables the model to be useful in alerting against any abnormal voltage readings which results in improved real time anomaly detection.

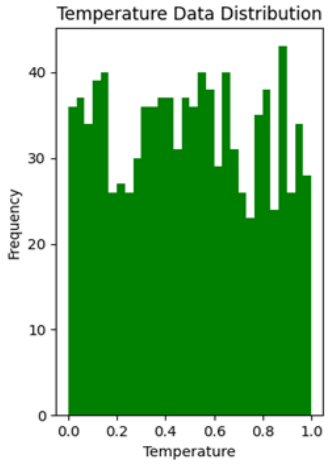

Figure 4. Temperature Data Distribution

The histogram of Temperature data (in green) as represented in fig 4 also exhibits an almost even distribution

between 0,0 and 1,0 with frequencies between 10 to 45. One of the parameters of health and maintenance of smart grids is temperature, which may cause the failure of components (such as transformers or circuit breakers) when they overheat. This simulation has a balanced temperature distribution, and the model is able to show the change in temperature which can be an indication of overheating or a failed machine. This information, together with the voltage and load, is essential in fault detection and in making the grid reliable by avoiding damages caused by overheating or faulty functioning parts.

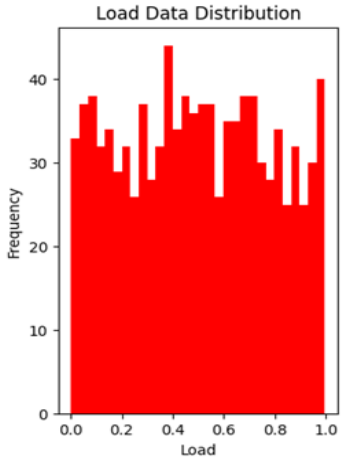

Figure 5. Load Data Distribution

As can be seen in figure 5, the histogram of Load (in red) has an equally distributed range of 0,0 to 1,0. Once again, the frequencies are 10-45, which means that load data is diversified enough to give information on the efficiency of the grid in its functioning. In an actual grid, load information is essential in determining over loads and possible energy losses. The regular distribution of the load information enables the model to draw the line between the normal changes in the load and more profound alterations that may be related to faults that may enable the grid operators to make decisions on the basis of the up-to-date energy consumption trend.

Figures 3-5 depict almost the same (0-1) distributions of voltage, temperature, and load, which means that the data is very idealized. In actual smart-grid scenarios, these variables are not often uniformly distributed; they are often subject to operating regimes, seasonal effects, demand cycles, equipment properties, and noise, leading to non-uniform distributions, which can be normal, multimodal, skewed or heavy-tailed. Thus, the good classification performance achieved in this study must be viewed as evidence-of-concept in controlled environments, and not as a certainty of performance in real-world applications. In order to enhance the external validity, further research should test the framework on more realistic data distributions (e.g., historical SCADA/ smart meter data or simulator output that is adjusted to real grid statistics). The synthetic data generator must at least be tuned to realistic regimes (e.g. peak/off-peak load patterns, seasonality in temperature, voltage regulation bands) and must have measurement noise, missingness and drift. This would give a more stringent evaluation of whether the model extrapolates past idealized uniform inputs and would render the results reported more reflective of actual smart-grid behaviour.

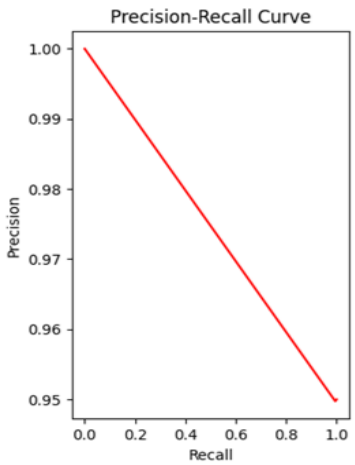

Figure 6. Precision-Recall Curve

The Precision-Recall Curve (in red) presented in figure 6 demonstrates the correlation between precision and recall of the predictions of the classifier. The higher the recalls, the lower the precision. This is normally the case when using imbalanced datasets, where identifying as many anomalies (positive class) as possible can result in certain false positives. Recall aims at reducing false negatives, i.e. anomalies are identified at the expense of considering some normal points as anomalies. This leads to a little less accuracy, but a large improvement in the recall, which is vital in such applications as smart grids, in which missing an anomaly may be disastrous. Hence, the trade-off is not problematic because high recall will mean that the majority of anomalies are identified to be investigated further.

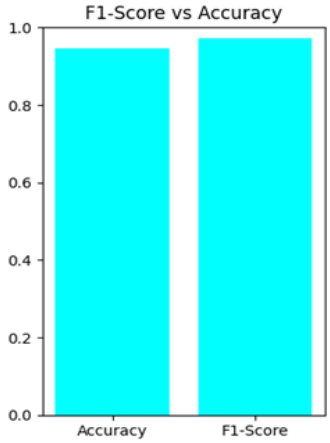

Figure 7. F1-Score vs Accuracy

The bar chart of F1-Score vs Accuracy shown in figure 7 reveals that both measures are not far apart meaning that there is almost complete performance in both the accuracy of classifications and the balance between precision and recall. F1-Score is the harmonic mean of precision and recall and such a high value indicates how the classifier can perform with a balanced score of both finding anomalies and reducing false alarms. The measure of accuracy, the percentage of correct classifications, is also very close to perfection, which proves that the model is making correct classifications of both non-anomalous and anomalous data within a reasonable margin.

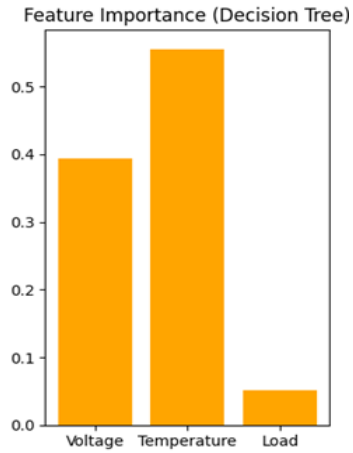

Figure 8. Feature Importance

As depicted in figure 8 (in orange) in the Feature Importance plot, it can be observed that Temperature is the most significant feature to the decision-making process in the Decision Tree model, followed by Voltage. The Load feature, although still being a part of the decision-making process, has less influence comparatively. This implies that temperature changes are an effective measure of health in the system and detection of anomalies. Practically, this is in line with the fact that temperature variations are usually one of the earliest indications of probable failures or defects in various parts including transformer and therefore is a very important attribute to the model.

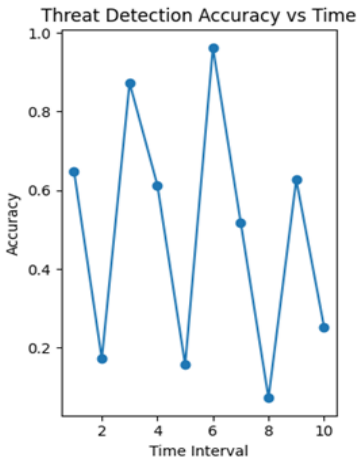

Figure 9. Threat Detection Accuracy vs Time

The plot of Threat Detection Accuracy Vs Time presented in figure 9 depicts that there are variations in the accuracy, which varies between 0,2 and 1,0 at different time periods. This variability denotes the effect that the various states and conditions of the system have on the detection of anomalies by the model. An example is that when there is stability in the grid the accuracy is high, however, when there is odd behaviour or complicated patterns the accuracy of detection can be reduced. This is common in dynamic systems such as smart grids where the conditions change at a very high rate and the model must respond in real time. This brings out the issue of the real-time challenge of ensuring detection performance under changing grid conditions.

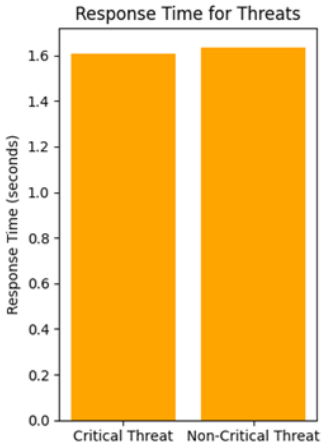

Figure 10. Response Time for Critical and Non-Critical Threats

The Response Time of Critical and Non-Critical Threats (both bars in orange) in figure 10 indicate that the system responds equally in almost the same time to both critical and non-critical threats (almost 1,6 seconds). This implies that the response mechanism is geared towards responding to both categories of threats to the same extent that it is urgent to act on both critical and less critical anomalies. This may be significant with regards to keeping the systems intact and making sure that no threat, no matter how serious, remains unattended too long.

The line chart of System Scalability in figure 11 indicates the number of threats that were detected with respect to the number of devices in the grid. Firstly, the more the number of devices increases, i.e. between 100 and 2 000 the higher the number of threats detected reaches its highest point of 50 threats detected. But with the increasing number of devices to 10 000, the number of threats detected is dramatically reduced. This trend indicates that the system would be less efficient with the increase of the devices, probably because of the data overload, system bottlenecks, or network congestion. This trend suggests that smart grids may face scalability issues as they increase in size and additional optimization may be necessary to ensure that the threat detection remains efficient as the grid increases in size.

An appropriate rationale is that in smart-grid monitoring, the cost of a false negative (the absence of a real anomaly, which can become a fault or a failure) is much greater than the cost of a false positive (an additional alarm that can be investigated and filtered). Thus, the operating point of the precision-recall curve must be selected to favour high recall, at the expense of slightly lowering precision, since the key safety goal of the

system is to prevent undetected critical events. It is possible to set the threshold on a validation set to achieve a minimum recall goal (e.g., ≥0,95) and maintain the false-alarm rate at an acceptable operational level.

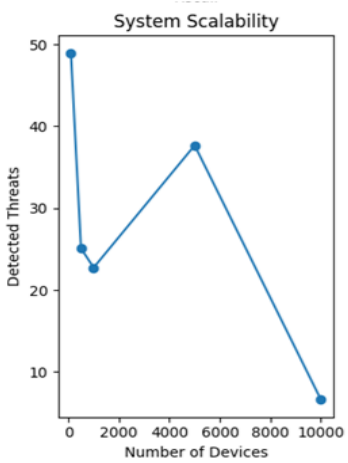

Figure 11. System Scalability (Detected Threats vs Devices)

DISCUSSION

The performance of the smart grid security framework is reflected in very positive results due to the high value of accuracy, precision, recall, and F1-score, which are above 0,95, which is a high level of anomaly detection and classification. The Voltage, Load and Temperature distributions were evenly distributed within the 0,0 to 1,0 range and the frequencies ranged between 10 and 45 and this sameness ensured that the model was free enough to learn and actually identify anomalies. The Precision-Recall Curve showed a high precision (approximately 1,0) with a small drop in precision with a rise in recall, which is the characteristic of the imbalanced classification task. The F1-score and the accuracy was close to 1,0 which further supports the fact that the model was performing excellently in classifying both normal and anomalous data points. The Feature Importance analysis indicated that Temperature and Voltage were the most important features that influenced the classification with Load being comparatively less significant. Threat detection accuracy across various time intervals was found to vary (between 0,2 and 1,0) in the system but the time taken to respond to critical and non-critical threats was also constant (approximately, 1,6 seconds), which ensured that threats were dealt with promptly. Although the system was very successful in the detection of the threats, the System Scalability plot showed that the number of detected threats was the greatest at 50 with about 2 000 devices in the grid, but it then declined drastically as the number of devices increased to 10 000. This implies the possibility of scalability issues, whereby the performance of the system decreases with increase in scale. All the reported results are at present provided as single point estimates (e.g., accuracy = 96,25 %). Experiments are to be repeated in a series of experiments and report variability in order to prove robustness. In particular, we will run N repeated (e.g., N=10) runs with varying random seeds (which influences sampling, training, and model initialisation) and report mean + standard deviation (or 95 % confidence intervals) of Accuracy, Precision, Recall, F1-score, and response time. Even though the classification metrics (figures 6,7,8) indicate the capability to differentiate between normal and abnormal behaviour, the usefulness of the framework in the real world is determined by the way these predictions will be converted into security outcomes. The proposed pipeline has the classifier output as a direct determinant of the response action: a true positive will cause mitigation (e.g., isolation/segmentation), a false negative will cause no action and will enable the threats to continue, and a false positive will cause unnecessary actions that can disrupt operations. Thus, the performance of classifiers ought to be provided alongside security effectiveness measures that measure the downstream effect of false and correct classifications. The fact that the response time to critical and non-critical threats is similar (approximately 1,6 s) cannot be viewed as a positive aspect per se. Critical events in safety-critical systems should be pre-empted by lower-priority events and are supposed to be responded to more quickly. The similar response times imply that the severity prioritization is not yet implemented in the operational response layer although severity may be reflected in the classification logic. The next generation must use priority queues and pre-emption, where critical events are guaranteed to have a lower response time than non-critical alerts.

These findings are a solid indication that the given framework can facilitate credible anomaly classification and security-conscious prioritization within the framework of a smart-grid. The high accuracy, precision, recall and F1-score are consistent, which means that the hierarchical classification phase can distinguish between normal and abnormal behaviour successfully, which is necessary to minimize the workload of operators and make decisions about operations faster. Specifically, the feature-importance findings indicate that the model

is trained on meaningful signals (temperature and voltage) which can be interpreted by practical grid health indicators, which contributes to the interpretability and confidence in the decision logic of the system. Operational wise, the precision-recall behaviour indicates that the model can be adjusted to higher anomaly capture in cases where safety is the primary concern, which is useful in smart grids where an important anomaly being missed may have dire consequences. The found differences in the threat detection accuracy with time also prove to be a valuable insight: the grid environment is dynamic, and the performance of the model is an indication of the changing conditions, which supports the practical importance of constant monitoring and the regular recalibration. Lastly, the near-uniform response time of about 1,6 seconds indicates that the framework is capable of providing rapid, consistent responses, which is encouraging in terms of real-time monitoring; combined with the analysis of scalability, these findings give a straightforward guideline on how to optimize the framework in the future to ensure the same benefits are achieved with a larger number of connected devices.

Table 3. Performance comparison of Proposed vs Existing methods

Method / Study	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed (DT + RF + hierarchical security control)	96,25	97,49	94,87	95,30
Farrukh et al.[11] (hierarchical cyberattack detection)	95,44	95,92	92,10	94,56
RNN + OCSA (reported in Sasi 2025) (DoS detection)	94,12	93,21	93,13	93,56

As demonstrated in the table, the proposed DT+RF framework has the best overall performance, the highest accuracy (96,25) and precision (97,49) and has a high recall (94,87) and F1 (95,30). It minimizes misclassification and provides a more appropriate trade-off between false alarms and detection of anomalies compared to Farrukh et al.[11] and RNN+OCSA. On the whole, the findings suggest that the suggested approach is more credible to detect than the two existing ones in this comparison. In order to make these results operational, an F1-score of approximately 95% indicates that the system could minimize the number of missed anomalies (which can escalate faults) and false alarms (which reduces the workload and alarm fatigue of operators). The response time of 1,6 s is however adequate to the application layer: it is too slow to be used in protection-grade trip signalling that needs milliseconds-level IEC 61850 GOOSE response, but may be appropriate in higher-level monitoring and distribution automation tasks where seconds-long response times are often acceptable. The results suggest that the framework is an encouraging evidence-of-concept of integrating hierarchical classification and layered security control because the metrics of high classification in the considered environment are high. Nonetheless, the findings also reveal some obvious impediments to deployment: equal response time between severity levels indicates that hierarchy is not fully enforced in the control layer, and the decrease in detections with scale indicates that there are bottlenecks in aggregation/communication/inference. Thus, the contribution is useful as an integrated design direction, but its practical preparedness is based on the application of priority-pre-emptive response, scalable distributed processing, and assessment based on the results of operational security.

Limitations

- Idealized synthetic data: Voltage/temperature/load is evenly spread (0-1) with predictable label relationships, which is not the case with real grid data. This can overblow accuracy and minimizes real world generalization.
- No difference in the severity of responses: There is no difference in the response time of critical and non-critical threats (~1,6 s), indicating that no prioritization is implemented at the action stage (critical events are expected to be faster/pre-emptive).
- Temporal detection instability: Accuracy ranges between 0,2 and 1,0, which suggests sensitivity to the changing grid conditions and potential concept drift; this is problematic when it comes to reliability in dynamic situations.
- Lack of proven integration between detection and security action: The system purports to provide classification-driven security control, yet the results do not provide security outcome measures (containment success rate, mitigation success, prevented outages), and therefore, the effectiveness of the integration is not established.

CONCLUSION

The work aimed to solve two fundamental problems of smart grids (1) the control of multi-source, heterogeneous data to make real-time operational decisions and (2) better cybersecurity by means of coordinated security controls. In the case of the first objective, the hierarchical classification method will be

proposed, which will structure the incoming grid data into the levels of priorities and use the Decision Tree and Random Forest classifiers to facilitate the timely detection of abnormal conditions. With the experimental setup described in this paper, the classification aspect demonstrated a 96,25 % accuracy, 97,49 % precision, 94,87 % recall, and F1-score of 95,30 %, which is a virtuous result in differentiating normal and anomalous patterns in the dataset used in this experiment. In the second objective, the study suggests a hierarchical concept of security control across the various layers (e.g., network, data and application) with the aid of monitoring and anomaly detection. Nevertheless, the provided evaluation is not yet able to show the effectiveness of end-to-end security using security-specific outcome measures (like mitigation success rate, containment time, or impact reduction). Moreover, the average response time of 1,6 seconds reported on critical and non-critical threats indicates that the severity-based prioritization is not applied to the operational response, which is a primary condition in safety-critical infrastructures. Another weakness is related to scalability: the findings show that the performance decreases with the number of devices, and the threat is not identified as well with more than about 5,000 devices. This demonstrates the importance of architectural and computational optimization prior to implementation in large-scale grid environments. The key innovation of this work is the combination of hierarchical, priority-driven data classification and a multi-layer security control strategy within one framework, which is intended to bridge the gap between the outputs of the anomaly detection and the security control decisions instead of considering analytics and security as independent modules. Future research ought to be aimed at confirming the framework on realistic or real-world data, reporting all model and data information to allow reproducibility, including security-effectiveness measurements, and enhance scalability and true critical-threat pre-emption such that more severe events are mitigated more rapidly than lower-severity anomalies.

REFERENCES

1. Syed D, Zainab A, Ghrayeb A, Refaat SS, Abu-Rub H, Bouhali O. Smart grid big data analytics: Survey of technologies, techniques, and applications. IEEE Access. 2021;9:59564-85.

2. Bhattarai BP. Big data analytics in smart grids: State of the art, challenges. IET Res J. 2019;13(6):142-53. https://doi.org/10.1049/iet-sen.2019.0414

3. Albayati A, Abdullah NF, Abu Samah A, Mutlag AH, Nordin R. Smart grid data management in a heterogeneous environment. Sensors. 2021;21(7):2347. https://doi.org/10.3390/s21072347

4. Dang-Ha TH, Olsson R, Wang H. The role of big data on smart grid transition. In: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity); 2015 Dec 19-21; Chengdu, China. IEEE; 2015. https://doi.org/10.1109/SmartCity.2015.43

5. Escobar JJM, Torres RJR, Hernandez JCI, et al. Comprehensive review on smart grids: Technologies, protocols, trends – Overview of data integration challenges. Systems. 2021;8(5):1011-25. https://doi.org/10.3390/systems8050101

6. Moghaddass R, Wang JH. A hierarchical framework for smart grid anomaly detection using large-scale smart meter data. IEEE Trans Smart Grid. 2018;9(5):5820-30.

7. Guato Burgos MF. Review of smart grid anomaly detection approaches. Electronics. 2024;13(1):18-25. https://doi.org/10.3390/electronics13010018

8. Zibaeirad A, Koleini F, Bi S, et al. A comprehensive survey on the security of smart grid: Challenges, mitigations, and future research opportunities. arXiv:2407.07966 [Preprint]. 2024. Available from: https://arxiv.org/abs/2407.07966

9. Khoei TT, Slimane HO, Kaabouch N. A comprehensive survey on the cyber-security of smart grids: Cyber-attacks, detection, countermeasure techniques, and future directions. arXiv:2207.07738 [Preprint]. 2022. Available from: https://arxiv.org/abs/2207.07738

10. Hassine L. Enhancing smart grid security in smart cities: A review. Energy Rep. 2025;11:2543-55. https://doi.org/10.1016/j.egyr.2025.02.102

11. Ali Farrukh Y, Ahmad Z, Khan I, Elavarasan RM. A sequential supervised machine learning approach for cyber attack detection in a smart grid system. In: 2021 North American Power Symposium (NAPS); 2021 Nov 14-

16; College Station, TX, USA. IEEE; 2021. https://doi.org/10.1109/NAPS52732.2021.9654767

12. Shadi MR, Ameli MT, Azad SA. A real-time hierarchical framework for fault detection, classification, and location in power systems using PMUs data and deep learning. Int J Electr Power Energy Syst. 2022;134:107399.

13. Saggi MK, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manag. 2018;54(5):758-90.

14. Albayati A, Abdullah NF, Abu-Samah A, Mutlag AH, Nordin R. Smart grid data management in a heterogeneous environment with a hybrid load forecasting model. Appl Sci. 2021;11(20):9600. https://doi.org/10.3390/app11209600

15. Syed D, Abu-Rub H, Ghrayeb A, Refaat SS. Household-level energy forecasting in smart buildings using a novel hybrid deep learning model. IEEE Access. 2021;9:33498-511.

16. Mahmud R, Vallakati R, Mukherjee A, Ranganathan P, Nejadpak A. A survey on smart grid metering infrastructures: Threats and solutions. In: 2015 IEEE International Conference on Electro/Information Technology (EIT); 2015 May 21-23; Dekalb, IL, USA. IEEE; 2015. https://doi.org/10.1109/EIT.2015.7293374

17. Yilmaz S, Dener M. Security with wireless sensor networks in smart grids: A review. Symmetry. 2024;16(8):1295. https://doi.org/10.3390/sym16081295

18. Ghadi YY, Alqahtani J, Boubaker S, et al. Security risk models with machine learning & predictive analytics. Eng. 2024;7(4):209-23. https://doi.org/10.3390/eng7030209

19. Mohamed A, Refaat SS, Abu-Rub H. A review on big data management and decision-making in smart grid. Power Electron Drives. 2019;4(1):1-13.

20. Wang Y, Chen QX, Hong T, Kang CQ. Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Trans Smart Grid. 2019;10(3):3125-48.

21. Zhang Y, Huang T, Bompard EF. Big data analytics in smart grids: A review. Energy Inform. 2018;1(1):8. https://doi.org/10.1186/s42162-018-0007-5

22. Jokar P, Arianpoo N, Leung VCM. Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans Smart Grid. 2016;7(1):216-26.

23. Zhou KL, Fu C, Yang SL. Big data driven smart energy management: From big data to big insights. Renew Sustain Energy Rev. 2016;56:215-25.

24. Kaitovic I, Lukovic S, Malek M. Proactive failure management in smart grids for improved resilience: A methodology for failure prediction and mitigation. In: 2015 IEEE Globecom Workshops (GC Wkshps); 2015 Dec 6-10; San Diego, CA, USA. IEEE; 2015. https://doi.org/10.1109/GLOCOMW.2015.7414155

FINANCING

State Grid Fujian Electric Power Science and Technology Project, project name: IEEE Guide for Smart Grid Data Classification and Grading international standard research project number: B3130M25Z121.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Xiang Yu,Yong Deng.
Data curation: Gang Wu,Ruyi Hu, Danhong Xie.
Formal analysis: Xiang Yu,Yong Den.
Research: Xiang Yu,Yong Deng.
Methodology: Yong Deng, Gang Wu.
Project management: Yong Deng.
Resources: Yong Deng, Danhong Xie.

Software: Ruyi Hu, Danhong Xie.
Supervision: Xiang Yu, Yong Deng.
Validation: Xiang Yu, Danhong Xie.
Display: Xiang Yu, Ruyi Hu.
Drafting - original draft: Xiang Yu,Yong Deng.
Writing - proofreading and editing: Xiang Yu, Yong Deng.

https://doi.org/10.56294/dm2026816 ISSN: 2953-4917