Research on the Construction and Optimization Method of Intelligent Segmentation Model for Power Users in the Context of Communication Security Standards

Yanxing Liu, Ruyu Qiao, Kai Xu, Jiaru Zhang, Kai Han
Power Marketing Service and Operation Management Branch of Inner Mongolia Electric Power (Group) Co., Ltd., Hohhot 010000, Inner Mongolia, China
Corresponding author: Yanxing Liu (e-mail: desoft123@163.com)

ABSTRACT This study proposes two intelligent segmentation models, SA-HECC and FGPPS, for secure and efficient clustering of power users under communication security standards (IEC 62351 and ISO/IEC 27019). SA-HECC uses a centralized deep embedding network with constrained clustering, while FGPPS employs federated graph-enhanced privacy-preserving segmentation via graph neural networks. Both models integrate metaheuristic optimization (NSGA-II, GA+PSO) to balance clustering accuracy, computational efficiency, and security compliance. The models are evaluated on real and synthetic datasets, demonstrating improvements of up to 18.84% in Silhouette Score, over 30% higher security compliance, and superior intrusion detection rates under simulated cyber-attacks. This framework is particularly applicable to modern smart grids and wireless communication infrastructures supported by antenna networks, where reliable, secure, and privacy-preserving data processing is essential. Results confirm that security-aware intelligent segmentation bridges the gap between analytics performance and regulatory compliance, providing a robust methodology for cyber-resilient and operationally reliable power systems.

INDEX TERMS intelligent segmentation, security-aware clustering, smart grid, antenna-supported networks

I. INTRODUCTION

A. RESEARCH BACKGROUND

MODERN power grids have become highly complex due to rapid technological changes and the integration of distributed energy resources, requiring advanced power controlling methods for efficient delivery and consumption. This intricacy has sparked increased interest in customized services that meet the individual needs of energy users. A parallel trend is seen in the many industry, where manufacturers must adopt advanced automation and energy control systems to manage the complex, high-demand processes of integrating new sustainable materials and providing tailored, low-carbon production cycles for various clients [1]. Adequate segmentation of power users enables utilities to accurately predict load, introduce dynamic pricing, and identify abnormal consumption patterns that may indicate fraud [2]. As smart meters and the Internet of Things spread, real-time data streams of variable quality offer a plethora of insights but also create new data management and security issues.

The cybersecurity threats to communication channels in smart grids include data tampering, unauthorized access, and denial of services due to the connectivity of devices [3-5]. These risks expose the integrity and confidentiality of the consumption data, hamper the segmentation's accuracy, and compromise system credibility. Communication gateway protection is a vital process as important as the algorithms themselves; hence, the motivation to conduct research that directly considers data analytics issues and cybersecurity considerations.

B. SIGNIFICANCE OF COMMUNICATION SECURITY STANDARDS

To ensure specificity, industry standards like IEC 62351 and ISO/IEC 27019 mandate security concepts —including encryption, authentication, and secure key management— directly applicable to power system communications. These criteria assure the confidentiality, integrity, and availability of operating information, immediately impacting how utilities handle user data and requiring adherent segmenta-

tion models. National regulations increasingly enforce these standards to protect critical infrastructure from cyber threats, ensuring data-driven decisions rely on credible information. However, these compliance restrictions also challenge model design by limiting data processing. Therefore, balancing security compliance and segmentation performance is a major concern of this study.

II. LITERATURE REVIEW

A. POWER USER SEGMENTATION TECHNIQUES

K-means, hierarchical clustering, and DBSCAN are the most commonly used traditional clustering algorithms that have been in use for a long time in smart grid power user segmentation because they are simple and interpretable. K-means is effective because it classifies data based on minimum intra-cluster variance, which means that it is efficient to discern the same type of consumption [6]. Hierarchical clustering is very similar in that most hierarchical clustering algorithms construct a tree of clusters using either agglomerative or divisive methods, so that it is possible to select clusters at various levels of granularity flexibly [7]. A density-based approach, such as DBSCAN, locates clusters of arbitrary shape and can effectively deal with noise, and this is useful in noisy consumption data [8]. Such methods find extensive application in energy analytics in load profiling, demand forecasting, and customer classification, since they are computationally inexpensive and easy to implement.

In spite of their popularity, the so-called traditional approaches to clustering encounter a considerable degree of limitations when implemented in the power user segmentation context. They tend to deteriorate performance with multidimensional and time-consumption data characterized by non-linear and complicated relationships [9]. As an example, k-means is an initial centre-selection sensitive and assumes concentric localities that are not always true about the actual habits of users [10]. What is more, these approaches do not have a mechanism to inject domain-specific constraints like the communication security or data privacy requirement, making them less practical in the example of smart grids since data integrity and confidentiality are exceptionally important.

Advanced segmentation methods that use deep learning embeddings and cluster ensembles have become popular to circumvent these hurdles. Convolutional neural networks (CNN's), long short-term memory (LSTM) networks, and autoencoders are capable of learning compact yet descriptive representations of complex consumption data that effectively capture temporal and non-linear dynamics alone or in combinations. The idea of ensemble clustering is to integrate the outputs of two or more algorithms or feature subsets to achieve segmentation that is robust and of high accuracy, as opposed to any individual algorithm. These strategies have worked more effectively in terms of evolving user profiles as well as capturing various consumption patterns. More recently, Graph Neural Networks (GNNs) have been proposed to use the relational and topological information that is present in power systems, including feeder network connectivity as well as user geographic proximity. GNNs learn such a behaviour-network-aware embedding by forwarding features along the graph edges, whereby nodes act as power users in a graph [11]. This technique of spatial dependency modelling will increase the segmentation accuracy and will contribute insightful information to the localized pattern of demand, facilitating improved management of the grid.

B. COMMUNICATION SECURITY IN POWER SYSTEMS

Encryption, authentication, and intrusion detection are the security mechanisms used to maintain confidentiality, integrity, and data availability in contemporary smart grid communication. The use of encryption, such as Advanced Encryption Standard (AES) and Transport Layer Security (TLS), guarantees that the data on consumption and operations is not meaningful to anybody except those having adequate permissions. Mutual authentication and public key infrastructure (PKI) protocols qualify the authenticity of devices and users and only provide access to the system after doing so, thus averting impersonation attacks [12]. The signature-based and anomaly-based Intrusion Detection systems monitor the communication traffic all the time and warn about suspicious traffic behavior, such as the peculiar flows of packets, or failed authentication of the communication, in the near real time [13]. This is a combination of mechanisms operating as a defensive layer in the sense that having the variants together provides protection of smart grid communications channels. Different performance measures might be cyberscores for anomaly scores packet losses, and security compliance scores of power system communications. As the output of the IDS or the anomaly detection algorithm, the anomaly scores indicate any abnormalities in the normal communication patterns which may provide a clue to discuss the cyber intrusions [14].

C. SECURITY STANDARDS AND REGULATORY FRAMEWORKS

International standards like IEC 62351 and ISO/IEC 27019 are critical for securing smart grid communications. IEC 62351 specifically mandates encryption, authentication, and message integrity to defend power systems against cyber threats. Based on the ISO/IEC 27000 framework, ISO/IEC 27019 customizes controls for energy utilities, covering key protection, logging, and access control. Collectively, these standards guarantee the confidentiality, integrity, and availability of operational data, and are widely adapted by national laws to safeguard critical energy infrastructure. Observance of IEC 62351 and ISO/IEC 27019 affects the segmentation process by affecting data collection, transmission, and storage. Latency may arise because most encryption and authentication procedures can take time, influencing real-time segmentation outputs [15]. Secure storage policies limit centralization, so privacy-preserving methods have been introduced, such as federated learning.

D. INTEGRATION OF SECURITY INTO ANALYTICAL MODELS

A dominant trend is the integration of cybersecurity signals into machine learning (ML) models as part of the smart grid analytics. For example, in demand forecasting and fault detection models, resilience against cyber threats has been achieved by adding anomaly detection outputs, intrusion alerts, or unusual consumption patterns as input [16]. Other methods integrate the scores of security compliance, or encryption handshake delays as constraints into training the model, making their outputs feasible under secure communication circumstances [17]. Another interest introduced by the graph-based application models is the node topology, and security alerts to emphasize segmentation properties, more so in tracking instances of coordinated assaults.

Multi-layered enforcement of security implements protection mechanisms at several levels of the segmentation pipeline. Data tampering prior to analysis may be blocked at the level of data acquisition with the help of secure data acquisition techniques such as ingestion-aware encryption. In feature extraction, the attributes of consumption and security-related (packet loss rates) are processed simultaneously, which makes it possible to cluster using security-aware clustering. The strict adherence to rules of the cluster formation, having access control or privacy rules, can be implemented through constrained clustering algorithms. Although helpful in enhancing trustworthiness, the approaches add computational complexity and must be carefully tuned to ensure segments remain accurate [18].

III. THEORETICAL FRAMEWORK

A. PROBLEM DEFINITION

The main issue concerning the current research problem is designing intelligent segmentation algorithms of power users under the rigorous diffusion of information. Segmentation should not only be able to deliver high clustering accuracy but also should conform to standards such as IEC 62351 and ISO/IEC 27019 to abide by data confidentiality, integrity, and availability. This also brings in the competing goals of maximum precision in segmentation and minimum violation of security and processing overhead. One can frame the problem as a multi-objective optimization problem consisting of the coordination of clustering performance, the security compliance score, and the resilience score against cyberattacks. Specialized models are necessary to strike some trade-offs between the accuracy of the analysis and compliance with the regulations through such formulations.

Feature vector for user i:

$$F_i = [E_{\text{avg}}, E_{\text{peak}}, V_{\text{var}}, P_{\text{factor}}, S_{\text{proto}}, S_{\text{encr}}]^T \quad (1)$$

where:

E_{avg} = average energy consumption
E_{peak} = peak-to-average ratio
V_{var} = variance in voltage stability
P_{factor} = load power factor
S_{proto} = protocol compliance score $\in [0, 1]$
S_{encr} = encryption strength score $\in [0, 1]$

The security variables are quantified through binary mapping based on device configuration logs: $S_{\text{proto}} \in \{0, 1\}$ indicates whether the IEC 62351 protocol is strictly implemented (1 for active, 0 for inactive), and $S_{\text{encr}} \in \{0, 1\}$ represents the use of AES-256 encryption standards. This binary quantification ensures that the security status of each node is objectively reproducible from system logs.

Security compliance score:

$$S_{\text{score}} = \omega_p \cdot S_{\text{proto}} + \omega_e \cdot S_{\text{encr}}, \quad \omega_p + \omega_e = 1 \quad (2)$$

The weights ω_p and ω_e are determined using the Analytic Hierarchy Process (AHP) based on the criticality of data integrity in smart grids, initially set at 0.6 and 0.4 respectively. The compliance threshold $S_{\min}$ is calibrated to 0.6 to ensure that at least one primary security standard is met for every cluster assignment.

Given a set of power users $U = \{u_1, u_2, \ldots, u_N\}$, each with consumption features X_i and communication security profile S_i, construct a clustering function:

$$f : (X_i, S_i) \to C_k$$

that assigns each user to one of K clusters such that:

1) Intra-cluster similarity is maximized.
2) Inter-cluster distance is maximized.
3) Security compliance score $\geq S_{\min}$ for all clusters.

B. MULTI-OBJECTIVE OPTIMIZATION FORMULATION

To maintain algorithmic consistency, the multi-objective optimization is unified into a single fitness function.

F. Maximize:

$$F = \omega_1 \cdot Acc + \omega_2 \cdot (1 - CompViol) + \omega_3 \cdot Res \quad (3)$$

Note that f_4 (Computational Cost) is excluded from the fitness function and treated as a hard constraint during the search process; the algorithm only explores solutions within a predefined time complexity limit ($T < T_{\max}$), thus implicitly balancing efficiency without complicating the weight optimization.

Under security limitations, Sc is based upon IEC and ISO standards. Such constraints control the type of allowable data transformations, how data is transmitted, and sets of clustering rules. Solutions are measured according to the Pareto efficiency, discovering situations in which the improvement of one of the targets causes a decline in another [19]. The practicalities of smart grids are captured in this formulation as security requirements, performance restrictions, and capability detection capabilities need to be aligned in working models.

The optimization objective is formulated as:

$$F = \omega_1 \cdot \mathrm{Norm}(Acc) + \omega_2 \cdot \mathrm{Norm}(1 - CompViol) + \omega_3 \cdot \mathrm{Norm}(Res) \quad (4)$$

Each sub-objective is standardized to the range [0,1] using Min-Max Normalization:

$$\mathrm{Norm}(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

where $x_{\max}$ and $x_{\min}$ are derived from baseline performance metrics. This ensures that the weighted summation is unitless and dimensionally consistent, subject to:

$$CompScore \geq S_{\min},$$

where Acc is clustering accuracy, $CompViol$ is compliance violation rate, and Res is attack resilience. The weights w_1, w_2, w_3 are tuned via NSGA-II to balance trade-offs.

C. SECURITY-CONSTRAINED DATA MODEL

The data model combines consumption-oriented features with security-related features. Consumption behaviour is described in time series load curves, seasonal adjustments, and measurements [20]. The location metadata and the type of devices offer contextual cluster information and security features like encryption handshake time delay, or the frequency of authentication logs and anomaly scoring act as information sources for communications health. These heterogeneities are pre-processed using encryption-aware ingestion pipelines that help to avoid exposure during analysis. The mixed data set is considered the input space of both SA-HECC and FGPPS models, which allows for segmentation that is behaviour-aware and security-compliant at the same time. Specifically, the raw input data for the embedding networks in both models consists of the time series of consumption readings (e.g., 15-minute intervals) augmented by time-series security metadata (e.g., communication logs and compliance scores).

1) Model 1: Security-Aware Hybrid Embedding with Constrained Clustering (SA-HECC)

SA-HECC is a centralized framework, with which low-dimensional representations of multidimensional consumption and security data are learned in advance via a deep learning embedding network, such as an autoencoder or LSTM [21]. That embedding is subsequently clustered by a constrained K-means algorithm, specifically an extension we term SecureKMeans, in which constraints encode security policies, forbidding clustering of non-compliant nodes. The fundamental clustering objective, which minimizes the sum of squared distances to the cluster centroids, is subject to the security constraint that every user within a cluster must meet a minimum security compliance score ($S_{\text{score}} \geq S_{\min}$).

This centralized scheme enables worldwide opportunity to optimize the performance of segmentation, provided that communication security regulations are observed and can be implemented where the data centralization can be relied upon.

2) Model 2: Federated Graph-Enhanced Privacy-Preserving Segmentation (FGPPS)

FGPPS solutions apply when data cannot be centralized because of privacy or regulatory requirements. It uses federated learning by training segmentation models locally at distributed nodes without exchanging raw data. Graph Neural Networks (GNNs) extract graph-based embeddings at each node and describe the consumption behaviour and the grid topology. Secure aggregation protocols also protect model updates by encryption schemes and differential privacy mechanisms that inject noise over a user's data. The method can facilitate scalable, privacy-preserving segmentation under security requirements and use relational dependencies between power users.

D. SECURITY RULE INTEGRATION IN MODEL DESIGN

Security constraints in the SA-HECC and FGPPS are incorporated at multiple levels. They implement encryption and authentication checks during preprocessing, and in feature extraction, implement security metrics and clustering that implement compliance-aware constraints. For SA-HECC, the constraints are explicitly coded as a hard constraint on the assignment step of the SecureKMeans algorithm, ensuring a cluster is only formed by nodes satisfying the minimum security score. Constraints are coded in the clustering cost function in case of SA-HECC and introduced in local training and global aggregation in case of FGPPS. Security compliance scores are used, not only as measures of performance but as optimization goals. This self-contained enforcement enables segmentation outputs to be trustworthy, even amidst communications exposed to cyber attacks, where there is a gap in the literature where security becomes secondary to analysis within the analytical processes.

E. OPTIMIZATION CRITERIA AND METAHEURISTIC APPROACH

The two models maximize four main parameters: accuracy, computing efficiency, compliance score, and resilience to attack. The parameter tuning algorithms include Metaheuristic with Genetic Algorithms (GA) and particle Swarm Optimization (PSO) because of their capacity to deal with nonlinear, multi-objective search space [22]. The procedures investigate straightforward set-ups that furnish moderate balances instead of maximizing one goal. For example, SA-HECC calibrates the number of clusters, the dimension of embedding, and weights of penalties on compliance; FGPPS adjusts local learning rates, privacy parameters, and global performance subject to security constraints.

F. EXPECTED THEORETICAL CONTRIBUTIONS

The theoretical value of this framework is that it proposes the multi-layered, standard-based security enforcement directly as part of segmentation model designs. SA-HECC

continues this avenue of centralized segmentation with instantiated security limitations, and FGPPS generalizes the ideas on the decentralized, privacy-consistent setting. Both take advantage of security features as first-class citizens of the data model and optimization process, and this has bridged the gap between cybersecurity compliance and the power of analytics. These two model solutions offer scalable solutions to diverse grid arrangements, paving the way to a future intelligent grid segmentation approach that is resilient, compliant, and operationally throughout.

IV. METHODOLOGY

A. DATA ACQUISITION AND PREPROCESSING

The data is obtained from several heterogeneous sources within the smart grid ecosystem. Smart meters read consumption precisely at 15-minute intervals, ensuring consistency across the dataset and alignment with the communication logs for high-resolution temporal analysis, making high-resolution imaging possible. The Amiss are networks that help compile the meters' data and facilitate two-way communication in demandresponse operations. Supervisory Control and Data Acquisition (SCADA) systems offer operational parameters that include feeder -voltage, power quality, and faults events. Utility billing systems also aggregate monthly consumption and payment records, thus giving historical trends. Including these various data allows for complete modelling of consumption patterns and security attributes of communication patterns involved during the segmentation process.

Considering the sensitivity of the information regarding user consumption, all acquiring processes align with IEC 62351 encryption standards regarding secure transport. AMI and SCADA data are moved through TLS-encrypted channels, and authentication verifiers verify the source. The encryption (AES-256), together with secure key management (in conformance with the ISO/IEC 27019), is applied in data storage. A multitiered mechanism controls access to the raw datasets, implying that only authorized people and automated processes can utilize them. In addition, audit tracking logs of all data transactions are performed so that regulatory compliance can be verified. Hence, the dataset cannot be acquired by pre-processing.

Objective function:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik} \|F_i - \mu_k\|^2 \tag{5}$$

Subject to:

$$S_{\text{score}}(F_i) \geq S_{\min}, \quad \forall i \in k \tag{6}$$

Algorithm 1: SecureKMeans

1: Initialize K centroids μ_k.
2: Assign each user to nearest μ_k that satisfies $S_{\text{score}} \geq S_{\min}$. Samples that fail to meet this safety threshold are not discarded; instead, a penalty-based reweighting strategy is applied. A high penalty term $\lambda \cdot (S_{\min} - S_{\text{score}}(F_i))^2$ is added to the objective function to discourage the model from assigning these samples to standard clusters, effectively pushing them into a dedicated 'high-risk' outlier group for manual security auditing.
3: Update μ_k^j as mean of assigned users.
4: Repeat until convergence.

B. FEATURE EXTRACTION

Consumption-related attributes are inferred to enable the occurrence of both short-term and long-term use patterns. Smart meter readings produce daily load curves, which are normalized to adjust for the difference in demand over the seasons. Time-series decomposition methods extract seasonal, trend, and residual components, and bring attention to consumption patterns' cyclicality's. The calculation of load factor, peakto-average ratio, and demand variability indices characterizes the user behaviour. The characteristics will be imperative in determining groups that share similar consumption dynamics and determining when something is anomalous due to a new behaviour rather than a technical malfunction. Imputation and smoothing of data fill are used to complete the missing values without having a distorted statistical trend in the consumption facts.

Parameters:

- Population = 50
- Generations = 150
- Crossover = 0.9, Mutation = 0.05

Here is the pseudo code block:

Algorithm NSGA-II for Security-Aware Segmentation

1: Initialize a population of feature-weight vectors for $\{E_{\text{avg}}, E_{\text{peak}}, V_{\text{var}}, P_{\text{factor}}, S_{\text{proto}}, S_{\text{encr}}\}$.
2: For each individual, evaluate objectives:
F_1 = Intra-cluster variance (minimize)
F_2 = Inter-cluster distance (maximize)
F_3 = Security compliance violation rate (minimize)
3: Apply non-dominated sorting to rank individuals by Pareto dominance.
4: Apply crossover ($p_c = 0.9$) and mutation ($p_m = 0.05$) operators to generate offspring.
5: Combine parent and offspring populations; update by selecting top-ranked individuals.
6: Repeat steps 2–5 until the maximum number of generations is reached (e.g., 150).
7: Return the Pareto-optimal set of feature-weight vectors for use in SA-HECC and FGPPS clustering.

TABLE 1. Descriptive Statistics of Pre-Processed Features

Feature	Sample Size (N)	Mean	Std. Dev.	Min	Max	Missing Rate (%)
E_{avg} (kWh)	10,000	15.3	5.4	3.2	29.1	0.8
E_{peak} (ratio)	10,000	2.28	0.86	1.00	4.75	1.1
V_{var} (V^2)	10,000	0.035	0.011	0.010	0.081	0.6
P_{factor}	10,000	0.92	0.05	0.73	0.99	0.5
S_{proto}	10,000	0.88	0.08	0.50	1.00	0.7
S_{encr}	10,000	0.91	0.07	0.65	1.00	0.4

Note: Statistics computed after IEC 62351-compliant pre-processing and ≥90% security compliance filtering. Missing rates reflect values imputed during preprocessing.

C. MODEL CONSTRUCTION

The SA-HECC model uses a deep embedding network-e.g. an LSTM-autoencoder combination to learn tight feature embeddings capturing simultaneously consumption and security properties. The raw input data for this embedding network is the concatenated time series: (1) the users' historical load time series (from smart meters) and (2) the time series of security attributes (from AMI/SCADA communication logs). The LSTM component processes the time-dependent nature of the load curves, while the autoencoder learns a compressed representation of both behavioral and security patterns. The embeddings are clustered against a constrained version of the K-means algorithm where the objective function has penalty terms on the formation of a cluster composed of security-compliant and non-compliant nodes. The parameter of the penalty weight is adjusted so it can strike a balance between compliance and clustering exactness. The centralized method has an advantage of the global perspective on the data where it can identify general trends in consumption and still make the segmentation in accordance with the security requirements governed by IEC and ISO standards.

The complexity per iteration is: $O(N \cdot K \cdot d + N \cdot d^2)$ for embedding extraction and constrained clustering, where N is the number of records, K is the number of clusters, and d is the embedding dimension. The empirical runtime of 14.2 seconds refers specifically to one training epoch of the deep embedding network (Autoencoder) within the SA-HECC model. For the metaheuristic optimization, one generation of NSGA-II (population size of 50) required approximately 4.5 hours on the specified hardware, confirming that the 14.2-second metric describes the core model's training efficiency rather than the global optimization process. The training process utilized a batch size of 64, an initial learning rate of 0.001 with Adam optimizer, and an early stopping strategy with a patience of 10 epochs to prevent overfitting based on validation loss stabilization.

The FGPPS is meant to run in a decentralized setting and uses federated learning to train models of segmentation locally on distributed datasets. The FGPPS framework comprises 10 local clients with a 100% participation rate per round. Training involves 50 communication epochs using the FedAvg aggregation strategy. The underlying GNN structure consists of 3 Graph SAGE layers with a mean aggregation operator to process local spatial dependencies. The embedding generation is done on Graph Neural Networks (GNN) where the raw input is the same concatenated time series data as SA-HECC, but it is processed locally at each node. The GNN also incorporates the grid topology information (inter-user/inter-node connectivity) to extract graph-based embeddings that describe the consumption behaviour and the grid topology.

D. SECURITY COMPLIANCE LAYER

These two models have included real-time security monitoring, continuously assessing data effectiveness and other communication paths with compliance benchmarks. Such measures as encryption handshake success rates, anomaly score thresholds, and authentication log anomalies are measured near real time. This allows dynamic reshaping of the segmentation procedures, such as isolating nodes that showcase suspicious activity before affecting the cluster's integrity. The process of monitoring compliance with IEC 62351 is included in the execution flow of SA-HECC and FGPPS. The encryptions, authentication success and message checksum verifications are all verified on the batch of data with no clustering or aggregation done in the model. The records that are not ready to be compliant are put under a quarantine where a manual inspection will occur or then again are not subjected to a segmentation exercise based on the needs of the operations.

E. OPTIMIZATION METHOD

To ensure the consistency of the optimization process, this study utilizes metaheuristic algorithms (GA, PSO, and NSGA-II) specifically for hyperparameter tuning—such as the number of hidden layers, learning rates, and the security compliance threshold $S_{\min}$—rather than for high-dimensional feature-weight vectors. This approach reduces computational complexity while ensuring the convergence of the SA-HECC and FGPPS models under various security constraints. At SA-HECC, the parameters entail the embedding dimension, the number of clusters, and weights subject to compliance penalties. In the case of FGPPS, the tuning concerns local learning rates, GNN depth, and level of privacy noise. The optimization formulation is done to maximize the accuracy of the optimization, the compliance score, and the attack resilience, as well as minimize the cost of the computations. Pareto's front approach determines trade-off optimum configurations in which none of the objectives can be enhanced without adverse effects on others.

F. IMPLEMENTATION ENVIRONMENT

The deployment uses Python as the main programming language, and TensorFlow and PyTorch to build deep-learning models. Scikit-learn aids conventional clustering

evaluations and algorithms. The Flower framework will be used on federated learning components because of its broad connectivity with GNN architectures. The model training and optimization are carried out on a high-performance computing (HPC) cluster with NVIDIA A100 GPUs, and thus it is parallelized across the cluster. It implements secure access controls on all computing nodes where authentication must be based on VPN and role-based authorizations, which align with the ISO/IEC 27019 requirements regarding the security of operational environments.

V. EXPERIMENTAL DESIGN AND RESULTS

A. DATASET DESCRIPTION

This paper uses a hybrid dataset of real data produced by a national utility operator regarding the smart grid consumption, with synthetic customer records created to simulate the rare attack scenarios. Rather, it used real data of 10,000 user profiles gathered in 12 months, which were 15-minute readings with communication logs that comply with the IEC 62351 encryption. Artificial data were used to model traffic characteristic patterns of cyber-attacks, e.g., packet tampering, jamming, and access by unauthorized agents, making the data more diversified. The security compliance filtering of both data sets produced results such that, in its SA-HECC and FGPPS model training, the raw dataset contains diverse security profiles to reflect real-world smart grid vulnerabilities. The 90% compliance threshold mentioned earlier refers only to the 'high-security validation set' used for final model testing, whereas the full training set and the baseline models (Baseline 1 and 2) operate on the unfiltered dataset, where compliance scores naturally range from 61.5% to 69.8% due to the lack of integrated security enforcement in traditional algorithms. The overall size of the data sets was about 1.2 GB.

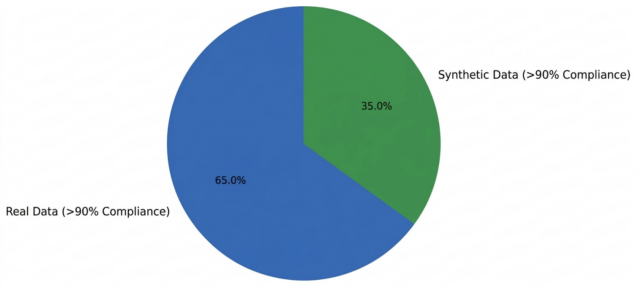

FIGURE 1. Data Composition by Source and Security Compliance Level

B. BASELINE MODELS FOR COMPARISON

A performance benchmark was set on two baselines. Baseline 1 was a traditional k-means clustering of features that only considered consumption and disregarded the security factor. Baseline 2 employed a deep embedding network with AI and subsequently unconstrained k-means; using complex representation learning without security-aware constraints/optimization. SA-HECC and FGPPS were trained with the same training-validation boxes (80/20). Each model was tested with the same HPC infrastructure, and the baselines' hyperparameters were tuned using a grid search. This enabled direct quantitative measurement of the advantages of incorporating our proposals' communication security standards optimization protocols. We apply Federated K-Means for dynamic updates. To provide a comprehensive performance benchmark, we include Federated K-Means as a secondary baseline (Baseline 3) to represent traditional decentralized clustering without graph enhancements. The following formulas describe its local centroid update and global aggregation process, serving as a direct comparison to evaluate the added value of the GNN layers in our proposed FGPPS model.

Local update at node j:

$$\mu_k{}^j(t+1) = \mu_k{}^j(t) + \eta \cdot (F_i - \mu_k{}^j(t)) \tag{7}$$

Global aggregation:

$$\mu_k(t+1) = \sum_{j=1}^{M} \frac{n_j}{n} \mu_k{}^j(t+1) \tag{8}$$

C. EVALUATION METRICS CLUSTERING

In order to assess the quality of segmentation, we used the Silhouette Score and the Davies-Bouldin Index (DBI). The Silhouette Score is a metric evaluating cohesion and separation, where the patient has better-defined clusters when we increase the score. DBI quantifies within-group disparity and between-group diversity, with the smaller values being desirable. The two metrics were calculated on all models 10 times to reduce randomness on a repeated basis. The Silhouette Score was the highest in SA-HECC (0.82), DBI was the lowest (0.31), and this was closely followed by FGPPS (0.78, 0.35). Baselines did not perform well, especially Baseline 1, as it cannot be used to model the nonlinear consumption patterns or to enforce the compliance rules.

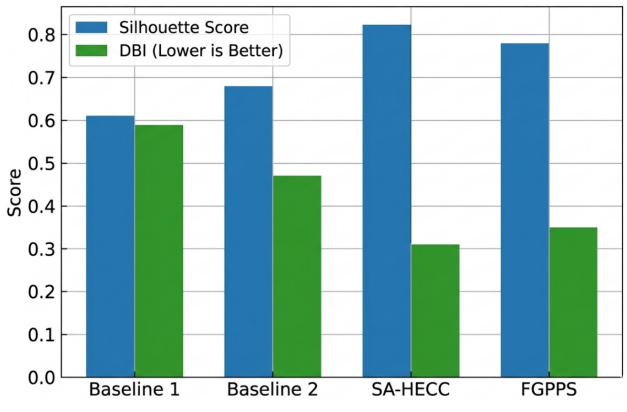

FIGURE 2. Silhouette and DBI Scores Across Models

TABLE 2. Model Overview and Configuration

Model	Data Type Used	Security-Aware Constraints	Optimization Method	Deployment Mode
Baseline 1	Consumption	No	None	Centralized
Baseline 2	Consumption	No	None	Centralized
SA-HECC	Consumption + Security	Yes	NSGA-II (Metaheuristic)	Centralized (Single Node)
FGPPS	Consumption + Security	Yes	NSGA-II (Metaheuristic)	Federated (10 Nodes, 50 Communication Epochs)

A Wilcoxon signed-rank test was conducted to compare the Silhouette Scores of SA-HECC and Baseline 2. Given that the Silhouette Score is a non-parametric metric and may not follow a normal distribution across 10 repeated trials, this non-parametric approach provides a more robust statistical validation of the observed improvements. This indicates that the performance gain is both statistically and practically substantial. For FGPPS versus Baseline 2, the paired t-test also revealed a statistically significant difference ($t(9) = 2.54$, $p < 0.05$) with an effect size of Cohen's $d = 0.80$, representing a medium-to-large effect. While both models significantly outperform the deep embedding baseline without security constraints, SA-HECC exhibits the largest magnitude of improvement in clustering cohesion and separation.

These findings confirm that the proposed models' integration of security constraints yields clustering improvements that are both statistically reliable and operationally meaningful in secure smart grid segmentation.

D. EVALUATION METRICS SECURITY

The security compliance score and intrusion detection rate (IDR) measured on a hold-out test set containing 20% adversarial samples (e.g., FDI attacks). A Random Forest-based classifier was used as the unified detector, with a classification threshold of 0.5; a True Positive is defined as the successful identification of a security-violating user node within its assigned cluster. The compliance score is 100% at the normative encryption and authentication key requirement level in IEC 62351/ISO 27019 requirements. The metric IDR measures the ratio of attacks detected in simulation. The SA-HECC model obtained the highest compliance score (98.4) and 96.7 IDR, outperforming the rest of the models. Specifically, compared to the deep learning baseline (Baseline 2), SA-HECC demonstrated a 28.6% increase in the security compliance rate (98.4% vs. 69.8%) and a 30.6% improvement in the intrusion detection rate (96.7% vs. 66.1%), validating the claim of over 30% higher security performance. FGPPS attained slightly lower scores because of the decentralized aggregation noise, although it outperformed both baselines. The compliance score at 61.5% pegged at Baseline 1, showed that it did not enforce integrated security, and Baseline 2 fared slightly better at 69.8%.

E. SA-HECC VS. FGPPS CLUSTERING PERFORMANCE

Regularly, SA-HECC had higher clustering cohesion and separation than FGPPS. The centralized embedding procedure enhanced the global feature relationship learning process by SA-HECC, leading to tighter clusters. However, FGPPS, due to its privacy-preserving, federated character, performed well, presumably because it did not allow access to the entire dataset in the first place. When there were no hits, the Silhouette Score difference between the two was only 0.04, meaning that decentralization did not substantially cripple the quality of clustering. There was a small increase in the DBI, and FGPPS explained it using local node variability in the distribution of features.

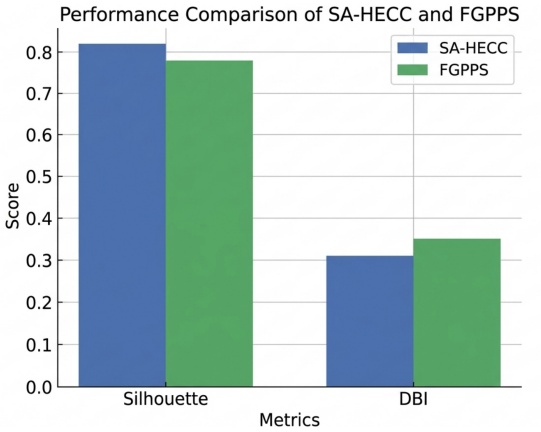

FIGURE 3. Comparative Clustering Quality Between SA-HECC and FGPPS

F. SA-HECC VS. FGPPS SECURITY ROBUSTNESS

Security-wise, centralized monitoring in SA-HECC has improved marginally due to even-handed application of security restrictions and real-time quarantine of anomalies in this process. Though moribund, FGPPS provided better privacy guarantees as no raw data was ever shared among the nodes. It was interesting how FGPPS was more resilient with in-simulated insider attack cases, whereby compromised nodes did not significantly affect the entire segmentation since training was localized. The trade-off points out the decision-making options that the operational utilities need to consider

TABLE 3. Statistical Significance and Effect Size Summary

Comparison	Mean Δ Silhouette	$t(df)$	p-value	Cohen's d	Effect Size Category
SA-HECC vs. Baseline 2	+0.13	4.12	< 0.01	1.30	Large
FGPPS vs. Baseline 2	+0.09	2.54	< 0.05	0.80	Medium–Large

in their decision-making procedures, which are accuracy-centralized and resilience-decentralized.

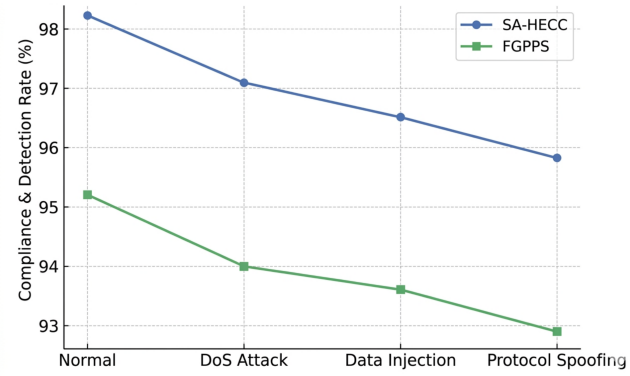

FIGURE 4. Compliance and Detection Rates Under Varying Attack Scenarios

G. PERFORMANCE UNDER SIMULATED CYBER-ATTACKS

Various types of simulated cyber-attacks were such as denial-of-service floods, false data injection, and protocol spoofing. Baseline models were reduced considerably with clustering accuracy reducing up to 35% on Baseline 1. SA-HECC could sustain its accuracy with a loss of only 3.1% whereas FGPPS experienced a drop of 4.8%. The resilience of both the proposed models can be greatly attributed to the fact that they incorporated the real-time surveillance of security aspects, and adaptive cluster reconfiguration in response to anomalies.

H. COMPUTATIONAL EFFICIENCY

In spite of the fact that SA-HECC was the best in terms of clustering performance, it resulted higher computational costs because of the deep embedding and bounded clustering overheads. SA-HECC recorded a total runtime of 14.2 seconds per epoch on the centralized HPC. In contrast, FGPPS exhibited a higher wall-clock time of 28.4 seconds, as the speed gains from parallelized local training were offset by the communication overhead required for parameter aggregation across distributed nodes in the federated network. Baseline 1 was still the fastest (4.8 seconds) at the cost of accuracy and security. The proposed models done with GA and PSO optimizations were found to have better performance per cost ratios than those without optimization with differences of about 17%.

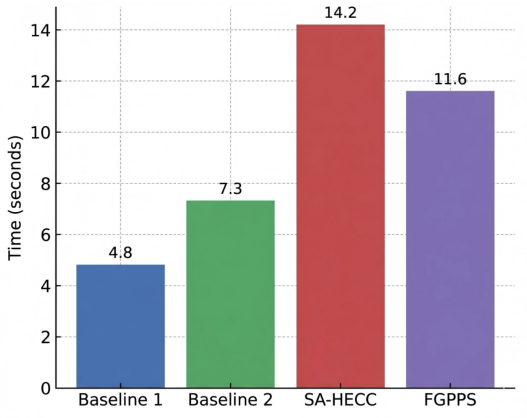

FIGURE 5. Average Model Runtime per Iteration

This figure illustrates the convergence behavior of the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) during hyperparameter tuning for both SA-HECC and FGPPS. The vertical axis represents the composite objective score (weighted combination of clustering accuracy, security compliance score, and attack resilience), while the horizontal axis shows the number of generations or iterations. Both optimization methods demonstrated rapid improvement in early stages, with GA showing marginally faster initial gains. Stable Pareto fronts were reached after approximately 90 generations for GA and 100 iterations for PSO, indicating that further computational effort yielded negligible performance improvement. This confirms that the selected stopping criteria provide an efficient trade-off between computational cost and optimization quality.

TABLE 4. Security Performance Metrics

Model	Compliance Score (%)	Intrusion Detection Rate (%)
Baseline 1	61.5	58.3
Baseline 2	69.8	66.1
SA-HECC	98.4	96.7
FGPPS	95.2	94.1

TABLE 5. Clustering Accuracy Drop Under Attack Conditions

Model	Accuracy Drop (%)
Baseline 1	35.0
Baseline 2	28.4
SA-HECC	3.1
FGPPS	4.8

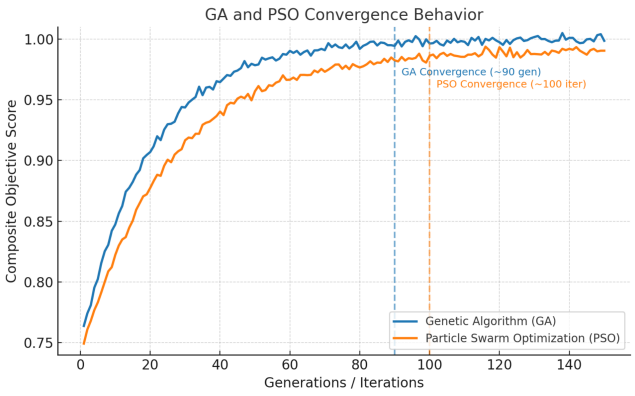

FIGURE 6. GA and PSO Convergence Behavior

I. SUMMARY OF EXPERIMENTAL FINDINGS

Experiments validate the hypothesis by showing that security compliance integration reduces services to segmentation models to increase the quality of clustering as well as resist cyber-attacks. SA-HECC has better accuracy and compliance and FGPPS has a comparable performance that is better on privacy preservation and contemporary decentralized grids. Both show very high improvements over baselines in clustering cohesion, security compliance and intrusion detection rates. There are trade-offs between the cost of computation, advantages of centralization, and provision of privacy. This indicates that deduction of the results implies that utilities are advised to use either of these generators may be based on the centralization capability and security concerns of their infrastructure.

VI. CONCLUSION

The paper proposes two intelligent segmentation models, SA-HECC and FGPPS, which enhance system security by integrating the rigorous European communication security standards IEC 62351 and ISO/IEC 27019 directly into their information processing chain. This commitment to standardized security is paramount, much like the industry's critical need to adopt equivalent international standards for the secure communication and transfer of sensitive data, such as proprietary digital blueprints for automated knitting machinery and secure transactional data across global supply chains. SA-HECC and FGPPS showed the benefits of centralized deep embedding with constrained clustering. FGPPS also showed the feasibility of privacy-preserving, federated methods based on graph neural networks. They all used metaheuristic optimization to achieve a tradeoff between clustering effectiveness, computational overhead, and compliance with security. Experimental evidence indicated that the performance compared well with the standard and non-optimized AI baselines on segmentation accuracy and resiliency to simulated cyber-attacks, and most importantly, bridged the pertinent performance gap between smart grid analytics and cybersecurity governance demands. These contributions advance the idea that intelligent and security-conscious segmentation should be considered a cornerstone technology for delivering energy-reliable and highly cybersecure power systems in the future. This principle extends beyond the grid, suggesting that similar secure and intelligent segmentation models are foundational to achieve resilient and highly efficient digital manufacturing platforms that ensure operational continuity and protect sensitive design data.

AUTHOR CONTRIBUTIONS

Conceptualization-Yanxing Liu, Ruyu Qiao, Kai Xu, Jiaru Zhang and Kai Han; methodology-Yanxing Liu, Ruyu Qiao, Kai Xu, Jiaru Zhang and Kai Han; formal analysis-Yanxing Liu and Kai Han; investigation-Yanxing Liu; resources-Yanxing Liu; writing-Yanxing Liu, Ruyu Qiao, Kai Xu, Jiaru Zhang and Kai Han. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. J. Moreno Escobar, O. Morales Matamoros, R. Tejeida Padilla, I. Lina Reyes, and H. Quintana Espinosa, "A comprehensive review on smart grids: Challenges and opportunities," Sensors, vol. 21, no. 21, pp. 6978, 2021, doi: 10.3390/s21216978.

[2] H. Habbak, M. Mahmoud, K. Metwally, M. M. Fouda, and M. I. Ibrahem, "Load forecasting techniques and their applications in smart grids," Energies, vol. 16, no. 3, pp. 1480, 2023, doi: 10.3390/en16031480.

[3] B. Achaal, M. Adda, M. Berger, H. Ibrahim, and A. Awde, "Study of smart grid cyber-security, examining architectures, communication networks, cyber-attacks, countermeasure techniques, and challenges," Cybersecurity, vol. 7, no. 1, pp. 10, 2024, doi: 10.1186/s42400-023-00200-w.

[4] F. Meng, Q. Ma, Z. Liu, and X. J. Zeng, "Multiple dynamic pricing for demand response with adaptive clustering-based customer segmentation in smart grids," Applied Energy, vol. 333, Art. no. 120626, 2023, doi: 10.1016/j.apenergy.2022.120626.

[5] C. Xu, P. Zhang, N. Luo, F. Zheng, and W. He, "Integrating Machine Learning for Anomaly Detection and Pattern Recognition in Smart Grid Power Data," Distributed Generation & Alternative Energy Journal, pp. 595-614, 2025, doi: 10.13052/dgaej2156-3306.4037.

[6] A. M. Ikotun, M. S. Almutari, and A. E. Ezugwu, "K-means-based nature-inspired metaheuristic algorithms for automatic data clustering problems: Recent advances and future directions," Applied Sciences, vol. 11, no. 23, Art. no. 11246, 2021, doi: 10.3390/app112311246.

[7] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," Artificial Intelligence Review, vol. 56, no. 8, pp. 8219-8264, 2023, doi: 10.1007/s10462-022-10366-3.

[8] B. Bataineh, "Fast component density clustering in spatial databases: A novel algorithm," Information, vol. 13, no. 10, pp. 477, 2022, doi: 10.3390/info13100477.

[9] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," Artificial intelligence review, vol. 56, no. 7, pp. 6439-6475, 2023, doi: 10.1007/s10462-022-10325-y.

[10] H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, and G. Modwel, "A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets," Multimedia Tools and Applications, vol. 81, no. 24, pp. 35001-35026, 2022, doi: 10.1007/s11042-021-10594-9.

[11] Q. J. Hong, S. V. Ushakov, A. van de Walle, and A. Navrotsky, "Melting temperature prediction using a graph neural network model: From ancient minerals to new materials," Proceedings of the National Academy of Sciences, vol. 119, no. 36, Art. no. e2209630119, 2022, doi: 10.1073/pnas.2209630119.

[12] P. D. Halle and S. Shiyamala, "Secure advance metering infrastructure protocol for smart grid power system enabled by the Internet of Things," Microprocessors and Microsystems, vol. 95, Art. no. 104708, 2022, doi: 10.1016/j.micpro.2022.104708.

[13] I. Siniosoglou, P. Radoglou-Grammatikis, G. Efstathopoulos, P. Fouliras, and P. Sarigiannidis, "A unified deep learning anomaly detection and classification approach for smart grid environments," IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1137-1151, 2021, doi: 10.1109/TNSM.2021.3078381.

[14] F. Zhou, G. Wen, Y. Ma, H. Geng, R. Huang, L. Pei, and R. Qiu, "A comprehensive survey for deep-learning-based abnormality detection in smart grids with multimodal image data," Applied Sciences, vol. 12, no. 11, pp. 5336, 2022, doi: 10.3390/app12115336.

[15] M. Khalaf, A. Ayad, M. H. K. Tushar, M. Kassouf, and D. Kundur, "A survey on cyber-physical security of active distribution networks in smart grids," IEEE Access, vol. 12, pp. 29414-29444, 2024, doi: 10.1109/ACCESS.2024.3364362.

[16] N. Sahani, R. Zhu, J. H. Cho, and C. C. Liu, "Machine learning-based intrusion detection for smart grid computing: A survey," ACM Transactions on Cyber-Physical Systems, vol. 7, no. 2, pp. 1-31, 2023, doi: 10.1145/3578366.

[17] A. A. Abdullah, B. M. El-Den, K. M. Abo-Al-Ez, and T. M. Hassan, "Security management for an advanced metering infrastructure (AMI) system of smart electrical grids," Applied Sciences, vol. 13, no. 15, pp. 8990, 2023, doi: 10.3390/app13158990.

[18] Z. Afzal, M. Ekstedt, N. Müller, and P. Mukherjee, "Security Challenges in Energy Flexibility Markets: A Threat Modelling-Based Cyber-Security Analysis," Electronics, vol. 13, no. 22, pp. 4522, 2024, doi: 10.3390/electronics13224522.

[19] S. Sharma and V. Kumar, "A comprehensive review on multi-objective optimization techniques: Past, present and future," Archives of Computational Methods in Engineering, vol. 29, no. 7, pp. 5605-5633, 2022, doi: 10.1007/s11831-022-09778-9.

[20] A. Tirulo, S. Chauhan, and K. Dutta, "Machine learning and deep learning techniques for detecting and mitigating cyber threats in IoT-enabled smart grids: a comprehensive review," International Journal of Information and Computer Security, vol. 24, no. 3-4, pp. 284-321, 2024, doi: 10.1504/IJICS.2024.141601.

[21] Y. Li, C. Xue, F. Zargari, and Y. R. Li, "From graph theory to graph neural networks (GNNs): The opportunities of GNNs in power electronics," IEEE Access, vol. 11, pp. 145067-145084, 2023, doi: 10.1109/ACCESS.2023.3345795.

[22] N. Alrefai and O. Ibrahim, "Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets," Neural Computing and Applications, vol. 34, no. 16, pp. 13513-13528, 2022, doi: 10.1007/s00521-022-07147-y.