Clustered Federated Learning for Early Stage Cyber Attacks Detection in Power Systems

May Myat Thwe, Peter Palensky, Alexandru Ștefanov

Department of Electrical Sustainable Energy
Delft University of Technology
Delft, The Netherlands
A.I.Stefanov@tudelft.nl

2025 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe) | 979-8-3315-2503-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISGTEurope64741.2025.11305511

***Abstract*— The increasing digitalization of Cyber-Physical Power Systems (CPPS) has enhanced power system operation and control but has also expanded the attack surface for cyber threats. Detection of early-stage attacks such as reconnaissance and Denial-of-Service (DoS) is critical to prevent power system-wide disruptions. Centralized Machine Learning (ML)-based techniques have been proposed for detecting cyber attacks. However, they struggle to ensure data privacy. Federated Learning (FL) can address this issue through collaborative model training without raw data sharing. Yet, FL's performance degrades under non-Independent and Identically Distributed (non-IID) data, a common scenario in real-world CPPS environments. In this paper, we propose a cluster-based FL method using Bidirectional Long Short-Term Memory (BiLSTM) for attack detection at the early stages of the cyber kill chain. It uses unsupervised clustering of client model updates for aggregation robustness and model generalization across heterogeneous clients. By grouping clients based on similarity in model updates, our method mitigates the adverse effects of data heterogeneity while preserving data privacy. The UNSW-NB15 dataset is used for distributed training under non-IID conditions and evaluation of the proposed method. Experimental results demonstrate that our cluster-based FL method achieves over 95% detection accuracy, proving its effectiveness in distributed cyber attacks detection in power systems.**

Keywords—federated learning, clustering, power systems, cyber attacks detection.

I. INTRODUCTION

Digitalization of Cyber-Physical Power Systems (CPPS) enhances power grid monitoring, control, and advanced analytics. While it offers efficiency in power system operation, it introduces new vulnerabilities and cyber security threats. Cyber attacks on power systems can potentially cause equipment damage, loss of load, instability, and power outages. Furthermore, sophisticated cyber attacks may cause power system-wide cascading failures leading to a blackout. Cyber attacks on Ukraine's power grid in 2015 and 2016, show the consequences of such threats [1]. They disrupted the operation of more than 30 substations and affected hundreds of thousands of customers. Such events highlight the urgent need to detect cyber threats in their earliest stages of the cyber kill chain. Particularly reconnaissance and Denial-of-Service (DoS) attacks often serve as kill chain stages for more severe advanced persistent threats [2]. Machine Learning (ML) advances have enabled intrusion detection and effectively identified complex threat patterns in power systems [3], [4]. ML models such as Support Vectors Machines (SVM), Decision Trees (DT), and Deep Learning (DL) networks were proposed to detect cyber threats in Supervisory Control and Data Acquisition (SCADA) systems and substations. However, traditional centralized ML training relies on aggregating data from multiple substations or regions and has limitations due to privacy and regulatory concerns. ML-based intrusion detection systems deployed in centralized architectures are facing significant challenges in large-scale power grids. The first challenge is the need for data sharing and privacy for effective attack detection. For instance, cyber threats originating at the distribution system level can escalate to transmission systems. In real-world scenarios, each substation operator typically has a limited number or type of attack samples. Since attacks on power systems can be distributed, no single substation can gather enough diverse attack data. Transmission System Operators (TSOs) and Distribution System Operators (DSOs) cannot share such attack examples (nor those normal behavior examples) with third parties because of national security concerns. Moreover, global regulations such as the General Data Protection Regulation (GDPR) and Cyber Resilience Act (CRA) limit the sharing and processing of sensitive data for critical digital infrastructures [5], [6].

Furthermore, digital substations exhibit non-Independent and non-Identically Distributed (Non-IID) data where different substations have different network traffic patterns, attack surfaces, and operational behaviors due to geographical, infrastructural, and environmental factors [7]. Even within a single organization managing multiple substations, a centrally trained machine learning may fail to generalize across different substations due to a lack of shared intelligence such as evolving attack patterns in different substations. The limitation of global attack information can influence the ML model's generalization ability and accuracy of detecting new or unseen threats. Federated Learning (FL) allows distributed and collaborative model training across different clients without transferring raw data. FL has shown potential in detecting network intrusions and identifying anomalous activities in industrial CPS [8]. Li et al. [9] applied federated learning to Distributed DoS (DDoS) detection. It combines fog computing and FL with training a Recurrent Neural Network (RNN) model in a distributed way to protect IoT devices from DDoS attacks. However, their work does not consider different data distributions. In this case, the global model obtained using the traditional federated averaging method may not be applicable to each client as it assumes that each client's data is independently and identically distributed, but this is difficult to achieve in real life. Lv et al. [10] proposed a Convolutional Neural Network (CNN)-based federated learning framework designed for DDoS attack detection and classification. They proposed a personalized federated learning framework that blends local and global CNN models to improve performance under non-IID conditions. Their approach achieves 0.9684 accuracy for binary and 0.8796 accuracy for multiclass classification.

However, it requires 1,000 communication rounds to reach convergence, introducing a significant communication overhead. Zhang et al. [11] proposed a federated RNN-based framework that incorporates K-Means clustering for hierarchical model aggregation and uses Synthetic Minority Oversampling Technique (SMOTE) to address local class imbalance. While their method achieves strong results with an average classification accuracy of 95%, it presents limitations. K-Means clustering assumes spherical data distributions and requires predefined cluster numbers. The hierarchical aggregation strategy requires approximately 4,800 communication rounds for model converge, which introduces communication overheads.

In this paper, we propose a cluster-based FL method i.e., Fedster, using Bidirectional Long Short-Term Memory (BiLSTM) for attack detection at the early stages of the cyber kill chain. It groups digital substations (clients) based on model update similarity and performs aggregation within each cluster. It mitigates the adverse effects of data heterogeneity while preserving data privacy. Our method enhances model accuracy with fewer communication rounds, for detecting reconnaissance and DoS attacks. The scientific contributions of this paper are summarized as follows:

1) We propose a cluster-based federated learning method using BiLSTM and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to address data heterogeneity in power system cyber attacks detection.
2) We simulate realistic non-IID conditions to evaluate the robustness of the approach.
3) We demonstrate the feasibility of the proposed method using the UNSW-NB15 dataset under IID and non-IID data distributions and compare its performance with FL methods such as FedAvg and FedProx.

The rest of this paper is organized as follows. In Section II, we provided background information on cyber attacks on power systems. Section III details the proposed method for federated learning attack detection. The experimental setup and results are provided in Section IV. Section V presents the conclusions and future work.

II. Cyber Attacks on Power Systems

While power systems are now more susceptible to cyber threats due to the increased digitalization, reconnaissance and DoS have emerged as critical attacks at the early stages of the cyber kill chain. These attack types are often part of advanced persistent threats and can be executed without directly compromising specific substation protocols.

A. Network Reconnaissance

Network reconnaissance involves gathering vital information about a communication network, including connected devices, network structure, protocols, applications, and services in an Information Technology (IT) – Operational Technology (OT) environment. It is is an important step in the cyber attack kill chain, identifying vulnerabilities that can be exploited by the attackers. Reconnaissance involves collecting network topology, communication protocols used, and running services, especially across integrated IT and OT environments. For instance, attackers use tools such as Network Mapper (Nmap) or tcpdump to perform ping sweeps and identify active Internet Protocol (IP) addresses for exploiting the Internet Control Message Protocol (ICMP). Transmission Control Protocol-Syncrhonize (TCP SYN) scans are also used to detect open ports and responsive services. In addition, attackers can perform host fingerprinting to determine operating systems and software versions, uncovering exploitable vulnerabilities. In the 2015 cyber attacks on the Ukrainian power grid, attackers exploited Microsoft Active Directory to infiltrate in the control centers, while in 2016, they identified critical OT communication protocols, i.e., International Electrotechnical Commission (IEC) 101, IEC 104, and IEC 61850, to open circuit breakers in substations [12].

B. Denial-of-Service Attacks

A DoS attack hinders the legitimate access of users or network devices to vital system resources such as network connections, computing power, and diverse application services. For example, they can disrupt the flow of critical measurements and controls by jamming SCADA communication channels or compromising field devices to prevent data reporting. Attackers may also target routing protocols or networks by flooding the communication channel with massive amounts of illegitimate traffic. Thus, legitimate users can experience significant disruptions in their communication, severely degrading overall network performance. It also pose significant risks to SCADA systems and threaten data availability [13]. Extensive research has been conducted on the susceptibility of SCADA systems to these types of attacks. How a DoS attack could exploit specific communication protocols used in substations, adversely affecting Remote Terminal Unit (RTU) processing and communication functions can be found in [14].

III. Federated Learning-based Attack Detection

Federated learning was introduced in 2016 by Google. It a distributed machine learning paradigm that trains machine learning models on separate datasets distributed across different devices or parties [15]. FL is an emerging area of research particularly due to the growth of big data, legal regulations, and global standards for data privacy protection [5], [6]. Unlike traditional centralized models that require data aggregation from multiple sources, FL enables decentralized model training across distributed devices while keeping the data localized. In power systems, FL offers advantages for cyber attack detection. System operators manage power system substations and generate vast amounts of operational data that are highly sensitive and subject to strict privacy regulations. In contrast to the centralized approach, this is a distributed model that facilitates global knowledge collected from all the distributed clients. A general or pre-trained model is initially distributed to clients, who then personalize the model with their local raw data. Clients perform ML tasks locally and send their parameters

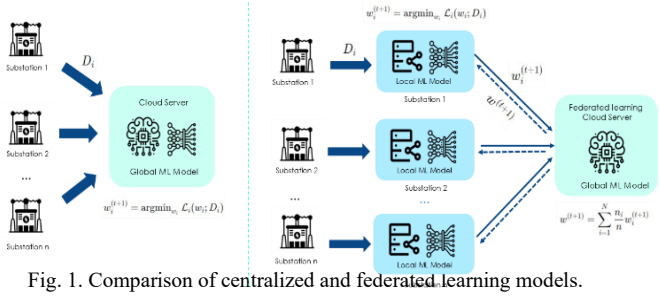

Fig. 1. Comparison of centralized and federated learning models.

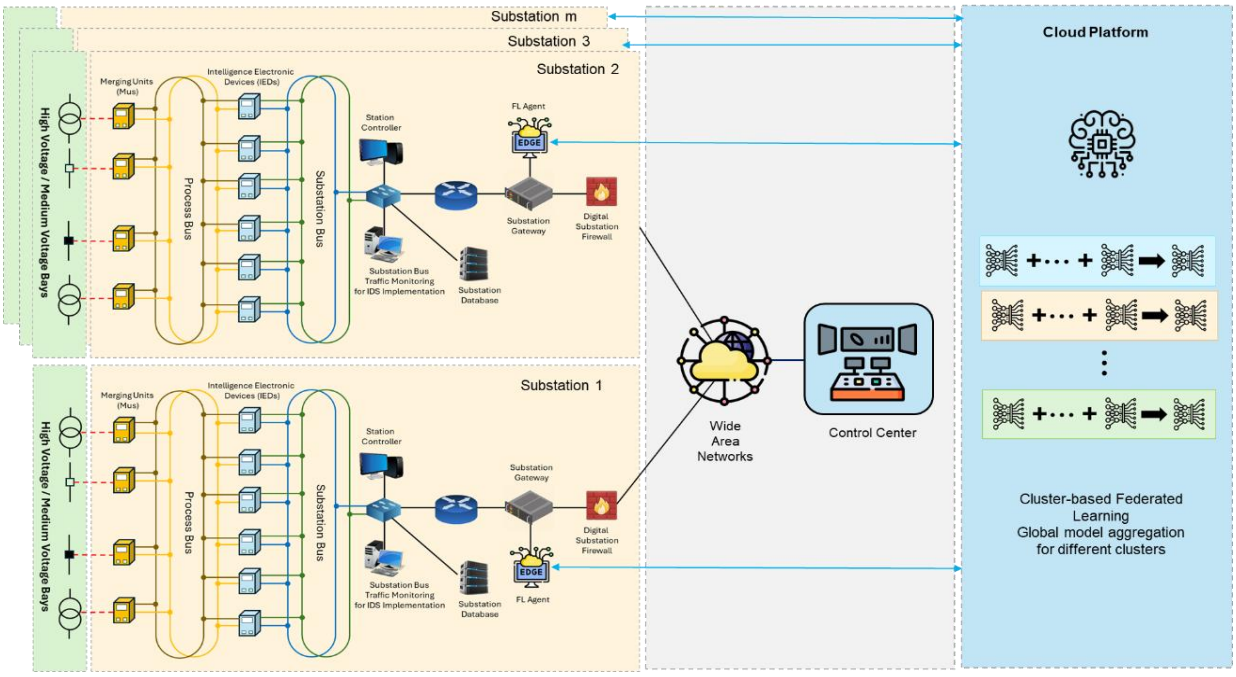

Fig. 2. System model for cluster-based federated learning in power systems.

to the server. The server aggregates all the updates received from the clients, performs ML tasks, and distributes the updated model to the clients. This process significantly reduces the need for extensive data transfer, thereby relieving potential data privacy concerns, while incorporating global attack intelligence. The comparison of centralized and federated learning is shown in Fig. 1.

A. System Models and Assumptions

The proposed method is designed based on the following assumptions, which reflect a realistic deployment scenario in a power system environment involving digital substations and a central coordination platform as illustrated in Fig. 2.

Substation Clients (FL Agents): Each substation in the power system has Merging Units (MUs) and Intelligent Electronic Devices (IEDs) that facilitate data acquisition and processing. A FL agent is deployed at the substation gateway, i.e., router, as it has comprehensive access to both inbound and outbound OT communication traffic. The FL agents responsible for monitoring the OT communication network activities and training a localized machine learning model for cyber attacks detection. In this scenario, substations operate independently and do not share raw data to ensure data confidentiality and regulatory compliance. Instead, they participate in FL-based collaborative learning, where only secure model updates, i.e., weights, are exchanged with the FL aggregation platform. This approach enables collaborative learning while protecting sensitive operational data. Moreover, substations have heterogeneous data distributions, e.g., operational conditions, OT network traffic patterns, and cyber threats. This non-IID nature of data presents challenges for traditional FL models, requiring a different aggregation strategy to reduce performance variations across digital substations.

Cloud Platform (FL Aggregation Server): It is responsible for computing and distributing the global intrusion detection model. Unlike traditional FL approaches that apply a single global aggregation, our approach implements a cluster-based FL strategy. The cloud platform does not store raw network logs; it only processes model gradients. Instead of treating all digital substations uniformly, the aggregation server clusters substations based on the similarity of their weight updates, to reduce bias and improve model convergence.

B. Data Preprocessing and Feature Engineering

Raw network features are first preprocessed to handle categorical encodings, i.e., one-hot and label encodings. We apply Min-Max normalization to scale all numerical features to a standard range [0, 1]. To address class imbalance, especially in attack data, we introduce a SMOTE-based oversampling technique that synthesizes minority class examples using adaptive k-nearest neighbor interpolation. This is included to prevent overfitting and enhance the model's ability for detection. In addition, to reduce feature redundancy and computational complexity, we applied Principal Component Analysis (PCA) with a retained variance threshold of 95%. PCA identifies the principal components, i.e., directions of maximum variance, in the data and projects the data onto a lower-dimensional subspace while retaining a significant portion of the original variance.

C. Clustered Federated Learning for Attack Detection

The proposed clustered federated learning method enhances traditional FL by addressing the non-IID data distribution challenge in power systems. Unlike standard FL methods such as Federated Averaging (FedAvg), which assume uniform data across clients, our method applies a DBSCAN-based clustering algorithm to group clients based on the similarity of their model updates. This enables localized aggregation improving model generalization, convergence, and early-stage intrusion detection performance. Each substation is modelled as an FL client that collects private traffic data and trains a local detection model without sharing raw data. The FL process consists of the following steps:

1. *Initialization:* The central FL server initializes the global BiLSTM model parameters.
2. *Client Selection:* In each communication round t, the server selects a random subset of clients to participate in training.

3. *Model Distribution:* The selected clients receive the current global model parameters from the server.
4. *Local Training:* Each substation (client) trains a local Bi-LSTM model using private network traffic data. LSTM networks are well-suited for analyzing sequential data such as network traffic flows, as they can capture temporal dependencies. A BiLSTM extends the standard LSTM by processing the input sequence in both forward and backward directions, capturing information from past and future time steps, which is crucial for detecting sophisticated cyber attacks. The local training process is defined as in (1).

$$w_i^{\{t+1\}} = w_i^t - \eta \nabla L(w_i^t, D_i) \quad (1)$$

where w_i^t represents the model parameters of client *i* at iteration *t*, η is the learning rate, ∇L is the gradient of the loss function, and D_i is the local dataset.

5. *Model Update Communication:* After each client computes the change in model parameters w_i, it securely transmits it to the server. The raw data remains on the client side to preserve privacy.
6. *Update Clustering:* Upon receiving updates, the server applies DBSCAN to group clients based on the similarity of their model updates. Unlike clustering techniques like K-means, DBSCAN does not require a predefined number of clusters, which is useful when the number of distinct attack patterns varies dynamically. In addition, DBSCAN inherently identifies outliers and excludes them from clusters. It is also capable of detecting clusters of arbitrary shape, which is important when model updates have diverse distribution patterns due to non-IID conditions. Instead of relying on centroid-based aggregation, DBSCAN clusters are based on the density of data points. By using a distance metric such as Euclidean distance, DBSCAN defines local neighborhoods and subsequently groups points that are in densely populated regions.
7. *Cluster-based Aggregation:* Instead of a single global aggregation, e.g., standard FedAvg, the server performs aggregation separately within each identified cluster. DBSCAN assigns cluster labels $C \in \{-1,0,1,\ldots,k\}$ on each round, where -1 represent the noise. For each cluster C_j (where $j \neq 1$), the cluster weight is calculated as in (2).

$$w_j^{cluster} = \frac{\sum_{i \in C_j} n_i . w_i}{\sum_{i \in C_j} n_i} \quad (2)$$

where n_i and w_i is the total number of training samples and the local weights of client i, and C_j is the set of clients belonging to cluster j. After obtaining cluster-level aggregation, the FL server aggregates global model updates by computing the mean of all weight updates within the cluster as given in (3).

$$w^{global} = \frac{\sum_{j \in C} N_j . w_j^{cluster}}{\sum_{j \in C} N_j} \quad (3)$$

where N_j is the total sample count in cluster *j*, and $w_j^{cluster}$ represents the aggregated cluster-specific model weights after round *t*+1.

8. *Model Distribution:* The server distributes back the relevant cluster-specific model to the corresponding clients. Clients belonging to the same cluster receive the same aggregated model, which is better adapted to their specific data distributions.
9. *Iteration:* Steps 2-8 are repeated for a predefined number of communication rounds.

IV. Experimental Results

A. Data Preprocessing

To evaluate the proposed method, we used the UNSW-NB15 dataset. It comprises multiple attack types, including our focus attack types: reconnaissance and DoS. It provides comprehensive representations of cyber attack behaviors alongside real-world network traffic, which is also relevant to early-stage cyber attacks in power systems. The dataset comprises 49 features, including flow-based, time-based, and traffic volume-related attributes. After preprocessing and feature selection, we split the dataset into 80% training and 20% testing sets. Non-IID data conditions were simulated using label skew, where clients receive data containing only a subset of attack types, and quantity skew, where clients randomly assign uneven data volumes, reflecting real-world operational traffic imbalance across substations.

B. Experimental Setup

The experimental setup includes both centralized and federated learning configurations. All models were implemented using PyTorch (v1.13.1) and Flower FL framework, executed on a system with an Intel® Xeon® W-2245 CPU (8 cores, 16 threads), 64 GB RAM, and Windows 10 OS. Each client uses a BiLSTM model trained on PCA-transformed input sequences. The architecture includes two bidirectional LSTM layers (64 hidden units), dropout (rate 0.3), and two fully connected layers, followed by a sigmoid-activated output for binary classification. The model is optimized using Adam and binary cross-entropy loss with a learning rate set to 0.001.

a) Centralized Setup: It serves as the performance upper bound, using the full preprocessed dataset with 50 training epochs (early stopping enabled) with a batch size of 32.

b) Federated Learning Setup: We distributed data across 100 clients, simulating substations. The model is trained in 10 communication rounds, with local training of five epochs per round. Each round involves 10 clients for training and 25 for evaluation. Both IID and non-IID scenarios are evaluated to assess robustness and data heterogeneity. Clients are grouped based on the similarity of their local model weight updates, defined by DBSCAN parameters: ϵ value is 0.05 and MinPts is 5. The cluster-based approach, i.e., Fedster, is compared with the other two baseline FL aggregation strategies, i.e., FedAvg and FedProx.

C. Attack Detection Results

To evaluate detection performance, we employ accuracy, precision, recall, and F1-score metrics and compare three federated learning strategies including the proposed strategies' performance under IID and non-IID scenarios against the centralized baseline. Table I presents the performance results of three federated learning strategies: FedAvg, FedProx, and Fedster. Fedster, our proposed

method, demonstrates marginal improvements over traditional FL methods, indicating its potential for enhanced intrusion detection, particularly in non-IID environments. Fedster outperforms FedAvg and FedProx across all metrics, achieving the highest accuracy of 0.9572 and an F1-score improvement of 0.004 (0.4%) relative to FedProx. The higher Recall (0.8875) of Fedster suggests that it minimizes missed attacks, i.e., False Negatives, while the balance between Precision and Recall in the Fedster reduces false positive rates and improves attack detection.

With regards to performance compared to the centralized approach, the centralized approach achieves the highest accuracy (0.9784), benefiting from comprehensive access to the entire dataset. The IID setting results in slightly lower accuracy (0.9700) as compared to the centralized one due to the structured partitioning of data across clients. In the non-IID setting, which reflects more accurately real-world heterogeneous substation environments, Fedster maintained strong performance with an accuracy of 0.9572, despite data variability. These results underscore that centralized models exhibit superior accuracy due to complete data access. Federated learning models, particularly Fedster still demonstrates good performance in distributed environments. This is crucial for substation-level cyber security, where data sharing is restricted due to privacy, regulatory, and operational constraints. The findings reinforce the practical feasibility of federated learning for large-scale, distributed cyber attack detection in power grids.

TABLE I. PERFORMANCE COMPARISON OF FL STRATEGIES.

FL Strategies	**Performance Metrics**			
	Accuracy	***Precision***	***Recall***	***F1***
FedAvg	0.9540	0.8749	0.8802	0.8771
FedProx	0.9568	0.8823	0.8862	0.8843
Fedster	0.9572	0.8831	0.8875	0.8852

V. CONCLUSIONS AND FUTURE WORK

This paper proposed a cluster-based federated learning method, i.e., Fedster, using BiLSTM for early-stage cyber attack detection in power systems, particularly reconnaissance and DoS attacks. To address data heterogeneity across substations, the proposed method applies DBSCAN-based clustering to group clients with similar model updates before aggregation. Evaluated on the public dataset under simulated non-IID conditions with label and quantity skew, Fedster demonstrated modest but consistent improvements in accuracy and F1-score over baseline FL methods such as FedAvg and FedProx. In particular, its higher recall suggests improved detection of actual attack instances, which is critical in substation-level cyber security. Nevertheless, the slight performance difference compared to baseline methods underscores the need for further refinement of both the model architecture and the aggregation strategy. One key threat to the validity of this approach can be using a BiLSTM-based architecture for local model training. It may not be ideal for deployment in resource-constrained environments such as substation gateways. Thus, future work will explore lightweight architectures or adjust the current model architecture to better provide real-time inference at the edge. In addition, while UNSW-NB15 is a widely used benchmark for network intrusion detection, it may not fully reflect the traffic characteristics of CPPS. To address this, we plan to generate a custom communication traffic dataset using a CPPS testbed that incorporates realistic transmission and distribution system interactions. Finally, to strengthen privacy guarantees in Fedster, future work will also focus on securing model updates by incorporating differential privacy to ensure that client data remains protected beyond the raw data level.

ACKNOWLEDGMENT

This work has received funding from the European Commission in Horizon Europe TwinEU project under Grant Agreement 101136119.

REFERENCES

[1] K. Zetter, "Inside the cunning unprecedented hack of Ukraines power grid", Wired, Mar. 2016.

[2] W. Zhe, C. Wei and L. Chunlin, "DoS attack detection model of smart grid based on machine learning method," in Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS), Shenyang, China, 2020, pp. 735–738, doi: 10.1109/ICPICS50287.2020.9202401.

[3] A. Abu Nassar and W. G. Morsi, "Detection of cyber-attacks and power disturbances in smart digital substations using continuous wavelet transform and convolution neural networks," Electr. Power Syst. Res., vol. 229, 2024.

[4] W. Danilczyk, Y. L. Sun and H. He, "Smart grid anomaly detection using a deep learning digital twin," in Proc. IEEE North Amer. Power Symp. (NAPS), Tempe, AZ, USA, 2021.

[5] European Data Protection Supervisor, "Federated learning," [Online]. Available: https://www.edps.europa.eu/press-publications/publications/techsonar/federated-learning_en

[6] European Cyber Resilience Act (CRA). [Online]. Available: https://www.european-cyber-resilience-act.com/

[7] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020.

[8] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for Internet of Things," IEEE Internet Things J., vol. 9, no. 11, pp. 8229–8249, Jun. 1, 2022, doi: 10.1109/JIOT.2022.3150363.

[9] J. Li, L. Lyu, X. Liu, X. Zhang and X. Lv, "FLEAM: A federated learning empowered architecture to mitigate DDoS in industrial IoT," IEEE Trans. Ind. Informat., 2021.

[10] D. Lv, Z. Chen, X. Liu, Y. Liang and H. Li, "DDoS attack detection based on CNN and federated learning," in Proc. Int. Conf. Adv. Cloud Big Data (CBD), 2022.

[11] J. Zhang, P. Yu, L. Qi, S. Liu, H. Zhang and J. Zhang, "FLDDoS: DDoS attack detection model based on federated learning," in Proc. IEEE Int. Conf. Trust, Security Privacy Comput. Commun. (TrustCom), Shenyang, China, 2021, pp. 635–642, doi: 10.1109/TrustCom53373.2021.000.

[12] J. Horalek, J. Matyska and V. Sobeslav, "Communication protocols in substation automation and IEC 61850 based proposal," in Proc. IEEE Int. Symp. Comput. Intell. Informat. (CINTI), 2013, pp. 321–326, doi: 10.1109/CINTI.2013.6705214.

[13] J. D. Markovic-Petrovic and M. D. Stojanovic, "Analysis of SCADA system vulnerabilities to DDoS attacks," in Proc. Int. Conf. Telecommun. Modern Satellite, Cable Broadcast. Services, Serbia, 2013, pp. 591–594.

[14] R. Kalluri, L. Mahendra, R. K. S. Kumar and G. L. G. Prasad, "Simulation and impact analysis of denial-of-service attacks on power SCADA," in Proc. Nat. Power Syst. Conf., India, 2016, pp. 1–5.

[15] M. Wen, Y. Yang, D. Wang, J. Yu, Y. Xiang and H. Jin, "FedDetect: A novel privacy-preserving federated learning framework for energy theft detection in smart grid," IEEE Internet Things J., vol. 9, no. 8, pp. 6069–6080, 2021.