A Novel Inherently Explainable Intrusion Detection System for Cyberattacks in Cyber-Physical Systems

Matthew Oinonen
Department of Electrical Engineering
Ontario Tech University (UOIT)
Oshawa, Ontario Canada
matthew.oinonen@ontariotechu.net

Yusef Mahmoud
Department of Electrical Engineering
Ontario Tech University (UOIT)
Oshawa, Ontario Canada
yusef.mahmoud@ontariotechu.net

Walid G. Morsi
Department of Electrical Engineering
Ontario Tech University (UOIT)
Oshawa, Ontario Canada
walidmorsi.ibrahim@ontariotechu.ca

***Abstract*—Smart Digital Substations (SDSs) are becoming vulnerable to cybersecurity threats due to the integration of Information and Communication Technologies (ICT). Attacks such as Denial-of-Service (DoS) on IEC 61850 Generic Object-Oriented Substation Event (GOOSE) messages can delay critical protection messages and compromise grid resiliency, hence Intrusion Detection Systems (IDSs) are necessary to detect such attacks. In this paper, a novel IDS based on XFrequentNet is developed, by adding an enhanced built-in visual explanation of the most important features for classification. Stockwell Transform is used to generate time-frequency spectrograms of the GOOSE message data that are then used by XFrequentNet to detect the attacks. In this work, a co-simulation is used, which includes an EXata network emulator that generates cyberattack traffic while the SDS is modeled in real-time using an OPAL-RT. The results show the proposed IDS achieving a high detection accuracy of 99.22% over all cases and a per-class accuracy of 99.78%. Visual explanations for the classifications are generated significantly faster than existing works, allowing for real-time explainability. The visual explanations revealed distinct high-energy time-frequency bands and recurring bursts in the Stockwell spectrograms of real DoS attacks. By embedding explainability directly within the model in real-time, the proposed IDS bridges the gap between accuracy and interpretability thus increasing the IDS reliability, transparency and trust.**

***Index Terms*—Cyber-physical Systems, Intrusion Detection Systems, Stockwell Transform, Substation Automation.**

I. Introduction

A. Problem Statement and Motivation

The integration of Information and Communication Technology (ICT) into the power systems has introduced new vulnerabilities to the cyber-physical systems (CPSs). Within the Smart Grid (SG) many modern substations employ IEC 61850 standard to enable high-speed communication between Intelligent Electronic Devices (IEDs) [1]. This improves the protection, control, and monitoring of substations, but also exposes critical equipment to cyberattacks. Specifically, DoS incidents targeting GOOSE traffic can degrade message delivery and impede the functioning of the system. Smart Digital Substations (SDSs) rely on IEC 61850 process-bus communication, particularly GOOSE messages, for protection and control. This dependence exposes the system to cyberattacks such as DoS, where delays or loss in network packets may compromise protection timing and grid resiliency. Intrusion Detection Systems (IDSs) using Machine Learning (ML) have been applied in this context with promising results [2]. However, these methods often act as black-box classifiers, limiting interpretability and operator trust. On the other hand, interpretable deep learning models such as FrequentNet [3] have shown how frequency-domain basis functions can replace backpropagation-trained filters to improve transparency. As cyberattacks on CPS grow in frequency and sophistication, there is an urgent need for detection frameworks that are accurate, real-time, and trusted by operators.

B. Previous Work

Table I summarizes the previous work on cyberattack detection in CPS. The work in [4] develops an unsupervised anomaly detection framework on a CPS testbed for transmission lines but lacks real-time operation capability. The work in [5] introduces a hybrid Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model for detecting False Data Injection (FDI) attacks on Photovoltaic system controllers but lacks explainability. The work in [6] develops an IDS based on a Deep Neural Network to detect cyberattacks on GOOSE messages, but only simulates a limited number of attacks and lacks explainability. The work in [7] develops a distributed IDS with semantic rules for DNP3 protocol in a Supervisory Control and Data Acquisition (SCADA) context that can operate in real-time but does not model the attacks realistically. In [8], real-time anomaly detection of spoofed GOOSE traffic within IEC 61850 substations is implemented. While [7] and [8] are rule-based and are technically explainable, these explanations are only possible in terms of their rules, and not all the features. While Explainable Artificial Intelligence (XAI) is used to explain the

979-8-3315-8138-1

decisions of black-box ML models, few works apply it to cybersecurity in the CPS. In [9] and [10] Shapely Additive exPlanations (SHAP) are used to explain decisions but is limited as SHAP is computationally expensive and cannot be used in real-time. The work in [11] and [12] also used SHAP, along with Local Interpretable Model-agnostic Explanations (LIME), to explain the decisions of a model detecting DoS attacks. In this case, both XAI methods require a significant amount of computations and do not allow for real-time explanations. The work in [13] applied Occlusion Sensitivity Maps (OSM) to explain the decisions of a CNN-based IDS but this XAI methodology is also computationally complex as it requires several forward passes of the CNN used. From Table I, it is notable that real cyberattacks and real-time capability lack explanations of their decisions. This real-time capability has been identified as open research path [14]. Several works that do investigate SDSs, lack realistic attacks, limiting the usefulness of their system in a real application. There is a need for an IDS that implements a computationally inexpensive, explainable methodology on data from realistic attacks to ensure that the designed systems are trustworthy and effective in real-time real-world SDS applications.

TABLE I
SUMMARY OF THE WORKS IN THE LITERATURE

Papers	SDS Context	Simulated Cyberattacks	Real Cyberattacks	Real-Time Capability	Feature-Based Explainability
[4]	✗	✗	✓	✗	✗
[5]	✗	✗	✓	✓	✗
[6]	✓	✓	✗	✗	✗
[7]	✓	✓	✗	✓	✗
[8]	✓	✗	✓	✓	✗
[9]	✗	✓	✗	✗	✓
[12]	✗	✓	✗	✗	✓
[10]	✗	✓	✗	✗	✓
[11]	✗	✓	✗	✗	✓
[13]	✓	✓	✗	✗	✓
This work	✓	✓	✓	✓	✓

C. Contributions

1) A novel deep-learning classifier named XFrequentNet is developed that implements inherent explainability by generating visual explanations of the important features for cyberattacks detection and classification.

2) A novel IDS for CPS based on the Stockwell transform and FrequentNet is developed to enhance the time-frequency representation of input signals, enabling more interpretable and discriminative feature extraction.

3) The developed IDS can provide real-time explianability thanks to the significant improvents to the computational performance of the IDS compared to existing deep learning and traditional approaches.

II. METHODOLOGY

A. Stockwell Transform and Novel Explainable FrequentNet

In this paper, the existing FrequentNet, which does not provide explanations to the decisions made, is modified by adding inherent explainability, thus becoming the explainable DL-method eXplainable FrequentNet (XFrequentNet). Fig. 1 depicts an overview of the proposed cyberattack classification approach, which includes a one-stage FrequentNet augmented with inherent explainability.

1) Stockwell Transform: Time-series samples of GOOSE message features gathered from a SDS under normal, power disturbance, and attack scenarios, are converted into time-frequency spectrograms using the Stockwell Transform (ST). The Stockwell Transform (ST) was introduced in [15] as a new technique for joint time-frequency analysis of nonstationary signals. For a given window function $\psi \in L^2(R)$, the ST expression for a signal $f(t) \in L^2(R)$ with respect to ψ is defined as (1), with $a \in R^+$ and $b \in R$.

$$[S_\psi f(t)](a,b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t)\, a\psi^*\big(a(t-b)\big) e^{-j\alpha t} dt \quad (1)$$

The ST can be viewed as a phase-corrected form of the Continuous Wavelet Transform (CWT). Unlike the CWT where oscillatory functions shift with the window, the ST keeps sinusoidal basis functions fixed in time while the Gaussian window dilates and translates across frequencies. This enables recovery of both amplitude and absolute phase spectra, ensures linearity in the presence of noise, and allows efficient computation through the Fast Fourier Transform [16]. An example ST spectrogram is given in Fig. 2 showing distinct high-energy streaks, outlined in red, correspond to transient bursts in GOOSE traffic. These localized patterns serve as discriminative cues for the proposed XFrequentNet to identify and explain cyberattacks in real-time.

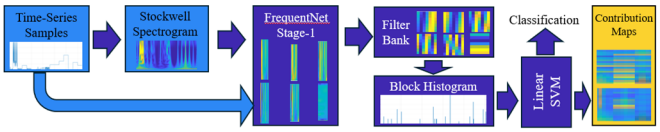

Fig. 1 Overview of the proposed IDS with explanations. The new contribution to FrequentNet, the development of contribution maps, is outlined in yellow.

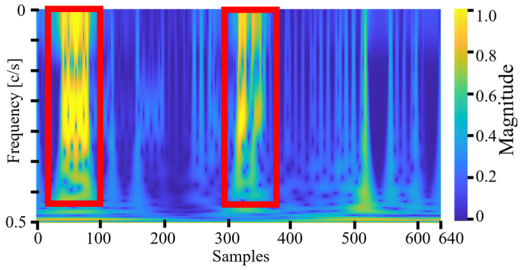

Fig. 2 An example ST spectrogram and the map of magnitudes. The y-axis is in terms of cycles per sample (c/s).

2) FrequentNet Stage-1: FrequentNet was proposed as a transparent alternative to conventional CNNs, where trainable filters are replaced with predefined orthogonal bases [3]. In its FourierNet implementation, the basis set consists of discrete Fourier functions. Given an input image (spectrogram from ST), candidate filters are drawn from the 1D discrete Fourier basis vectors $fb_1(w_k)$ and $fb_2(w_k)$, given by (2) and (3):

$$fb_1(w_k) = \frac{1}{\sqrt{n}}[1, cos(w_k), \dots, cos\big((n-1)w_k\big)]^T \quad (2)$$

$$fb_2(w_k) = \frac{1}{\sqrt{n}}[1, sin(w_k), \dots, sin\big((n-1)w_k\big)]^T \quad (3)$$

where, $w_k = 2\pi k/n$ denotes the discrete frequency, k is the frequency index, n is an index, and T is the transpose operator. Convolving the image with these selected filters produces the first-stage feature maps. A second stage repeats the procedure: patches are sampled from the first-stage maps, another set of

top-ranked Fourier bases is selected, and the corresponding convolutions yield a deeper set of frequency-based representations feature maps. At the output of these processes, the feature maps are binarized and aggregated into block-wise histograms, following the encoding scheme of Principal Component Analysis network (PCANet) [17]. PCAnet is a simple deep learning network for image classification. Each output map is binarized as (4):

$$T_i^{l_1} = \sum_{l_2=1}^{L_2} 2^{l_2-1} H(I_i^{l_1} * u_{l_2}) \quad (4)$$

where, $T_i^{l_1}$ is the binarized feature map obtained for the i^{th} image at the first stage, l_1. The total number of filters used in the second stage is denoted by L_2, which is the index of each filter within that stage. H denotes the binary function, which maps positive values to 1 and non-positive values to 0. $I_i^{l_1}$ refers to the input feature map or image corresponding to the i^{th} sample at stage l_1, while u_{l_2} represents the l_2^{th} basis filter selected from the Fourier basis set. This framework retains the hierarchical structure of CNNs while ensuring that every filter corresponds to a known Fourier frequency, thereby combining efficiency, interpretability, and performance [3] and [17]. These feature maps then serve as input to the binary hashing and block histogram encoding stage.

3) Block Histogram: these are generated by tiling the input feature maps with 9×9 blocks with a stride of 3 and forming a histogram of the packed codes within each block. A 9×9 block captures local time-frequency patterns with enough detail, while the stride of 3 adds overlap so no transient is missed. Each block's feature code forms a histogram, and combining all block histograms gives the final feature vector f. Before constructing the global feature vector, each pixel location within the feature maps must first be encoded into a compact integer code that represents the joint binary responses of all filters at that spatial position.

$$B(x,y) = \sum_{l=1}^{L_1} 2^{l-1} \cdot I(R_l(x,y) > 0) \quad (5)$$

Where, $B(x,y)$ is the integer-encoded value at spatial position (x,y) formed by combining the binary responses from all L_1 feature maps. L_1 denotes the number of filters or output maps in the first stage. $R_l(x,y)$ represents the response value at position (x,y) in the l^{th} feature map. I is the indicator function that returns 1 if the condition inside is true and 0 otherwise. The term 2^{n-1} assigns a unique binary weight to each feature map so that the combined result $B(x,y)$ forms an 8-bit code corresponding to 256 possible values.

4) Linear SVM with Error Correcting Output Codes: This 8-bit code in (5) is then fed to a Linear SVM for classifications into normal, disturbance or attacks. Since the output of the block histogram contains multiple classes, a one-vs-one Error-Correcting Output Codes (ECOC) scheme using a ridge-regularized linear learner is used with the Linear SVM. The ECOC decomposition gives a linear classifier S:

$$S = W f + b \quad (6)$$

where, W is a weight matrix of size C by D, C is the number of classes, D the number of features, f the feature vector, and b the bias vector. The class score, S_c, is given by (7):

$$S_c = \sum_d W_c(d) \cdot f(d) + b(c) \quad (7)$$

where, each row of W, W_c, corresponds to one class where c is the given class. The term in the class score $W_c(d) \cdot f(d)$ in (7) represents how much feature d supports class c, and by projecting these contributions back into the spectrogram created by ST of the input signal, maps that show where evidence for the prediction is concentrated can be generated.

5) Inherent Explanaibility Via Maps: Since both the filters and the classifier are linear and explicitly parameterized, contribution maps are developed that highlight where (blocks/bins) and how (filter combinations) evidence accumulates for each class. The design follows the FrequentNet principle of replacing learned convolutional kernels with analytic frequency bases to preserve interpretability [3]. Contribution maps quantify the average feature-weight interactions over all training samples of a given class, identifying the frequency-time regions that most consistently characterize that class. These maps convey the model's aggregate understanding of class-defining patterns.

6) Training Protocol: Training data was taken as 80% of the input cases as this yielded the best results during testing. The data were selected in a stratified manner.

7) FrequentNet Hyperparameters: the following hyperparameters were selected as they yielded the best results: an image size of 60×16, a patch size of 7×7, 8 $L1$ filters, block sizes of 9×9, a block stride of 3, 256 hash bins, a regularization term of $\lambda = 0.0003$, and a random seed of 42.

B. Performance Evaluation Metrics

The overall accuracy of the classifier (τ) is determined using (8). For each class the metrics are accuracy τ_c, precision p_c, recall r_c, and F1 score, as seen in equations (9)-(12), where N_o is the number of testcases in the dataset, and TP_c, TN_c, FP_c, and FN_c are the number of true positive, true negative, false positive, and false negative cases per class respectively. The macro values across all classes are also reported for each metric, and is the arithmetic mean of the values across all classes.

$$\tau = (\sum_{c=1:C} TP_c)/N_o \quad (8)$$

$$\tau_c = (TP_c + TN_c)/N_o) \quad (9)$$

$$p_c = TP_c/(TP_c + FP_c) \quad (10)$$

$$r_c = TP_c/(TP_c + FN_c) \quad (11)$$

$$F1_c = 2 \times TP_c/(2 \times TP_c + FP_c + FN_c) \quad (12)$$

To quantify the reliability of the observed per-class accuracy on a finite test set, the standard Z-test (13) as recommended in [18] is used, where Z=1.96 for the 95% confidence interval.

$$CI_c = \frac{2N_o\tau_c + Z^2 \pm Z\sqrt{Z^2 + 4N_o\tau_c - 4N_o\tau_c^2}}{2(N_o + Z^2)} \quad (13)$$

III. Evaluation

A. Cyber-Physical Test System Co-Simulation Platform

Fig. 3 depicts the cyber-physical test bed, which includes a real-time cyber-physical co-simulation platform. The system integrates EXata CPS 8.1.3.0 and OPAL-RT OP5033XG Real-

Time Digital Simulator (RTDS). A detailed depiction of the setup of the CPS co-simulation platform is shown in Fig. 4.

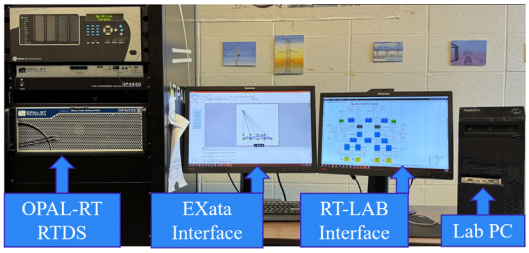

Fig. 3 The cyber-physical testbed used in this work, showing the OPAL-RT, and the PC that hosts the model, with the EXata and RT-LAB interfaces.

1) SDS Model in OPAL-RT: The SDS system in [19] was used and it consists of two transformers that step down 66 kV to 22 kV as in Fig. 4. This system was modified to work with OPAL-RT. Five IEDs are implemented within OPAL-RT: TRSF1 and TRSF2 for transformer overcurrent protection, FDR for overcurrent protection on the outgoing feeders, BFP for breaker-failure protection monitoring of the SDS, and CB CTRL, the control of the circuit breakers. The attacks in [19] are extended with new FDI, Replay, Message Suppression (MS), and DoS attack cases. FDI and Replay attacks are implemented by modifying the GOOSE data within the OPAL-RT model. MS attacks are implemented by disabling GOOSE message publishing from the OPAL-RT GOOSE drivers, and simulated DoS attacks by publishing GOOSE messages from an affected IED every timestep.

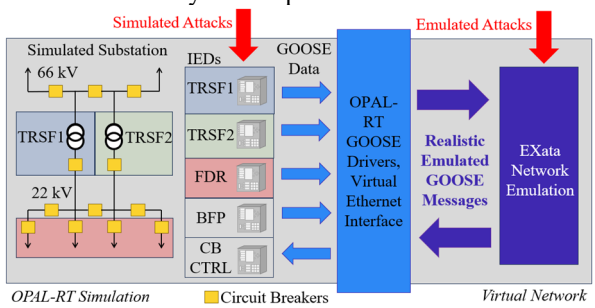

Fig. 4 Diagram of the test system and the flow of traffic on the network.

2) Real-Time Test System Set-up: the system was implemented as shown in Fig. 4 using a real-time cyber-physical co-simulation platform. EXata CPS emulates the communication network while OPAL-RT models the SDS in real-time with a fixed-timestep of 50 µs. The IEDs in the model publish and subscribe to GOOSE messages through virtual ethernet ports connected to the emulated network in EXata. This setup enables hardware-in-the-loop (HIL) experimentation where cyberattacks generated in EXata directly affect the behavior of protection and control functions in OPAL-RT. This configuration provides two key advantages, including realism and synchronization thus traffic-level effects of DoS attacks (e.g., packet delays, drops, jitter) are preserved.

3) Dataset Description: GOOSE messages were captured and dissected using Wireshark. MATLAB was then used to clean and normalize the packet dissections, after which feature extraction was used to collect the same features as in the Group 4 dataset from [20] as listed in Table II.

TABLE II
COMPOSITION OF TEST DATASET

Class	Number of Test Cases	Class	Number of Test Cases
Normal	895	Attack, MS	266
Disturbance	249	Attack, DoS-Sim	249
Attack, FDI	253	Attack, DoS-Real	266
Attack, Replay	256	**Total**	2434

B. Results

1) IDS Performance Results: To evaluate the performance of the system, a PC with an Intel i7-14700K processor, an Nvidia A4000 Graphics Processing Unit, and 16 GB of memory was used. The results have shown that XFrequentNet performed effectively on the EXata-based dataset, with an overall accuracy of 99.22%. Per-class metrics are reported in Table III, with high-per class accuracies in each class, and high confidence intervals. Three classes: FDI, MS, and DoS-Real, achieved perfect accuracy, precision, and recall. Normal and Disturbance achieved perfect recall, while DoS-Sim achieved a perfect precision score. End-to-end training and evaluation of the system was completed in 1.6181 seconds.

TABLE III
CLASS-WISE PERFORMANCE RESULTS OF THE SYSTEM

Class	Accuracy [%]	Precision [%]	Recall [%]	F1-Score [%]	CI+ [%]	CI- [%]
Normal	99.34	98.24	100.00	99.11	98.93	99.59
Disturbance	99.88	98.81	100.00	99.40	99.64	99.96
Attack, FDI	99.84	100.00	98.42	99.20	99.58	99.94
Attack, Replay	100.00	100.00	100.00	100.00	99.84	100.00
Attack, MS	100.00	100.00	100.00	100.00	99.84	100.00
Attack, DoS-Sim	99.38	100.00	93.98	96.89	98.99	99.63
Attack, DoS-Real	100.00	100.00	100.00	100.00	99.84	100.00
Macro-Averaged	99.78	99.58	98.91	99.23	99.52	99.87

2) Explanations Generated by XFrequentNet: Unlike black-box classifiers, XFrequentNet provides intrinsic explanations at the block and filter level. Maps for each case are given in Fig. 5, which show that each testcase is decided upon by unique features, as each feature contribution map is unique. For all maps in this work the same magnitude colour map as in Fig. 2 is used. Examples from each testcase are analyzed in terms of the feature maps generated, seen in Fig. 5, where each class exhibits distinct spatial patterns that are highlighted in red in each case. In Normal cases, Fig. 5 a), there is a unique strong feature in the mid-frequency range. For Disturbance, Fig. 5 b), and MS, Fig. 5 d), cases, features are concentrated in central blocks. FDI attack cases, Fig. 5 c) contain a strong feature in the mid sample range that extends across all frequencies. Replay testcases, Fig. 5 (e), produce strong vertical mid-sample activations. For DoS, Fig. 5 f), the strongest evidence appears in distinct mid-sample blocks.

IV. DISCUSSION

A. Performance Comparision with Literature Works

The accuracy of the proposed approach, 99.22%, is competitive among literature works that report accuracy metrics, which range from 94.48% [2] to 99.93% [5]. The proposed method is further compared to literature works in terms of performance in Table IV. The proposed approach has markedly lower inference times, which enables true real-time

performance justified by the significantly reduced inference time. The proposed system is also the only method with feature-based explainability.

TABLE IV
COMPUTATIONAL PERFORMANCE COMPARISON AMONG LITERATURE

Work	Inference Time [ms]	Feature-Based Explainability
[2]	2	✗
[4]	N/A	✗
[5]	N/A	✗
[6]	1–10	✗
[7]	N/A	✗
[8]	4	✗
This Work	**0.008**	✓

B. Performance Comparison Against XAI Methods

The proposed method is compared against the time it takes to compare the XAI methods used in literature [9] – [13] to a representative explainable IDS consisting of a pretrained CNN (AlexNet) trained on the ST spectrograms, refer to Table V. It is seen that the proposed method is faster than other methods by several orders of magnitude, allowing for true real-time explanations that the other XAI methods cannot achieve.

TABLE V
COMPUTATIONAL PERFORMANCE COMPARISON

XAI Methodology	Time to Generate an Explanation
SHAP	7.765 s
LIME	4.392 s
OSM	0.3468 s
This Work, XFrequentNet	**0.0003679s**

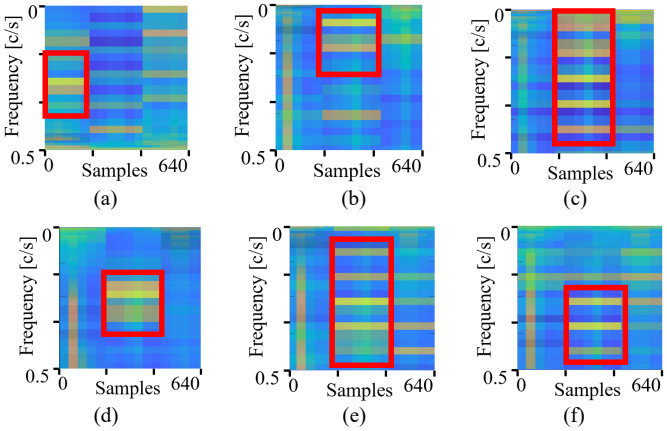

Fig. 5 Feature maps for (a) Normal, b) Disturbance, c) FDI, d) MS, e) Replay and f) DoS cases.

V. CONCLUSION

In this work a novel cyberattack detection methodology for CPS is proposed, named XFrequentNet, that adds real-time inherent explanations. This system is more trustworthy as it is inherently explainable and is highly computationally efficient compared to other DL models and so is suited for real-time detection as is required in SDSs. The proposed system was tested on a novel dataset on a co-simulation platform using OPAL-RT and EXata CPS and achieved a high detection accuracy of 99.22% and macro-averaged per-class accuracy of 99.78%. Using the inherent explainability, distinct features of the attacks and the disturbances were observed, which help in explaining the decisions made by the proposed IDS.

REFERENCES

[1] J. C. Lozano, K. Koneru, N. Oritiz and A. A. Cardenas, "Digital Substations and IEC 61850: A Primer," *IEEE Communications Magazine*, vol. 61, no. 6, pp. 28-34, Jun. 2023.

[2] T. S. Ustun, S. M. S. Hussain, A. Ulutas, A. Onen, M. M. Roomi, and D. Mashima, "Machine Learning-Based Intrusion Detection for Achieving Cybersecurity in Smart Grids Using IEC 61850 GOOSE Messages," *Symmetry*, vol. 13, no. 5, Art. no. 826, May 2021.

[3] Y. Li, K. Song, Y. Sun, and L. Zhu, "FrequentNet: A Novel Interpretable Deep Learning Model for Image Classification," *arXiv.org*, 2021.

[4] P. Gong, H. Yang, H. Wu, H. Lu, Y. Liu, Z. Qi, W. Wang, D. Wu, and X. Gao, "Co-Simulation Platform with Hardware-in-the-Loop Using RTDS and EXata for Smart Grid," *Electronics*, vol. 12, no. 17, Art. no. 3710, Sep. 2023.

[5] Y. B. Tadese, et al., "Hybrid CNN-LSTM Detection of Cyber-Attacks on MPPT Controllers in PV Systems," in *2025 IEEE Texas Power and Energy Conference (TPEC)*, College Station, TX, USA, 2025, pp. 1-6.

[6] L. E. da Silva and D. V. Coury, "A new methodology for real-time detection of attacks in IEC 61850-based systems," *Electric Power Systems Research*, vol. 143, pp. 825-833, Jan. 2017.

[7] S. N. Mohan, G. Ravikumar, and M. Govindarasu, "Distributed Intrusion Detection System using Semantic-based Rules for SCADA in Smart Grid," in *2020 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, Chicago, IL, USA, 2020, pp. 1-5.

[8] H. Nhung-Nguyen, M. Girdhar, Y.-H. Kim, and J. Hong, "Machine-Learning-Based Anomaly Detection for GOOSE in Digital Substations," *Energies*, vol. 17, no. 15, Art. no. 3745, 2022.

[9] I. Elgarhy, M. M. Badr, M. Mahmoud, J. Ni, M. Alsabaan, and T. Alshawi, "Investigation of the Robustness of XAI-Based Federated Learning Against Adversarial Attacks for Smart Grid False Data Detection," *IEEE Internet Things J.*, vol. 12, no. 15, pp. 32179-32192, Aug. 2025.

[10] C. Benrebbouh, H. Mansouri, S. Cherbal, S. Djahel, and D. Arrar, "An explainable CNN-based intrusion detection system for enhanced smart grid security," in *2024 Int. Conf. on Inf. and Commun. Technol. for Disaster Manag. (ICT-DM)*, Setif, Algeria, 2024, pp. 1-7.

[11] R. Upadhyay, P. Patel, L. Pathak, K. Shah, R. Gupta, S. Tanwar, and G. Bhardwaj, "XAI-based DoS attack detection framework for reliable energy distribution in smart grid systems," in *2025 3rd Int. Conf. on Intell. Data Commun. Technol. and Internet of Things (IDCIoT)*, Bengaluru, India, 2025, pp. 340-345.

[12] C. I. Nwakanma, L. A. C. Ahakonye, T. Jun, J. M. Lee, and D. -S. Kim, "Explainable SCADA-edge network intrusion detection system: Tree-LIME approach," in *2023 IEEE Int. Conf. on Commun., Control, and Comput. Technol. for Smart Grids (SmartGridComm)*, Glasgow, United Kingdom, 2023, pp. 1-7.

[13] A. A. Nassar, M. Oinonen, and W. G. Morsi, "An Integrated Trustworthy Detection and Classification of Cyber-Physical Attacks in the Presence of Disturbances Using Morphological Image Processing and Explainable AI," *IEEE Trans. Ind. Cyber-Phys. Syst.*, vol. 3, pp. 442-453, 2025.

[14] T. R. Gadekallu, P. K. R. Maddikunta, P. Boopathy, N. Deepa, R. Chengoden, N. Victor, W. Wang, W. Wang, Y. Zhu, and K. Dev, "XAI for Industry 5.0 – Concepts, opportunities, challenges, and future directions," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 2706-2729, 2025.

[15] R. G. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: the S transform," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 998-1001, April 1996.

[16] D. Wei and Y. Zhang, "Fractional Stockwell transform: Theory and applications," *Digital Signal Process.*, vol. 115, Art. no. 103090, Aug. 2021.

[17] T. -H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017-5032, Dec. 2015.

[18] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Pearson Addison-Wesley, 2005.

[19] X. Wang, C. Fidge, G. Nourbakhsh, E. Foo, Z. Jadidi, and C. Li, "Anomaly Detection for Insider Attacks from Untrusted Intelligent Electronic Devices in Substation Automation Systems," *IEEE Access*, vol. 10, pp. 6629-6649, Jan. 2022.

[20] The MITRE Corporation, "ICS Matrix," 2024. [Online]. Available: https://attack.mitre.org/matrices/ics/