Transfer Learning-Based Classification of Cyber Attacks Against Power Grids

Joshua Foster*, Abdulrahman Takiddin†, Muhammad Ismail*, Shady S. Refaat‡
*Department of Computer Science, Tennessee Technological University, Cookeville, Tennessee, USA
†Department of Electrical & Computer Engineering, Florida State University, Tallahassee, Florida, USA
‡Department of Engineering and Technology, University of Hertfordshire, Hertford, UK
{jtfoster42, mismail}@tntech.edu, atakiddin@fsu.edu, herts.ac.uk

***Abstract*—Power grids serve as the backbone of critical infrastructures, enabling the efficient control and distribution of electricity. Subsequently, the number of cyber attacks against power grids continues to grow, especially during times of international conflict and uncertainty. Thus, we must continue to further the research that secures and ensures the stability of power grids. Related research has made progress to this end through the application of artificial intelligence-based systems, but such works suffer from the following limitations: (a) they strongly focus on attack detection, neglecting attack classification, (b) they produce complex systems requiring large amounts of computation without concern for reducing complexity, and (c) they lack concern for how the systems can be adapted when new attacks arise. These limitations motivate our work, where we develop efficient and high-performing classification systems that are adaptable to new attacks. Specifically, we develop diverse benign and attack datasets consisting of cyber and physical layer data using our power system testbed. Additionally, we adopt SHapley Additive exPlanations to reduce the total number of features required to accurately classify attacks by 83% while maintaining a superior accuracy of 97%. Lastly, we use transfer learning to enhance fine-tuning and adapt our classification system to classify new cyber attacks in power grid environments.**

***Index Terms*—Attack classification, machine learning, power grids, SHapley Additive exPlanations (SHAP), transfer learning.**

I. Introduction

Power grids serve as the backbone of critical infrastructures where lots of communications take place to ensure customer demand is met efficiently and reliably through real-time monitoring, distributed control, and secure data exchange between cyber and physical components. Power grids are continuously adapted to become more efficient to meet consumer needs [1]. To improve efficiency, power grids have been enhanced with advanced digital technologies such as programmable logic controllers (PLCs), human machine interfaces (HMIs), advanced supervisory control and data acquisition (SCADA) systems, and smart meters [2]. However, the addition of such devices has made power grids increasingly more vulnerable to cyber attacks [3]. Thus, it is vital we continue to further the research in this domain to ensure that power grids are resilient to cyber attacks. Related research is commonly motivated by the events of the 2015 Ukraine power grid attack, where Russia used sophisticated cyber attacks to render the Ukraine power grid incapacitated, leaving over 200,000 customers without electricity [4]. Thus, enhancing the security of power grids is essential. Today's power grids present cyber-physical systems that produce vast amounts of power measurements and network traffic that can be used to evaluate the status of power grids [5]. Related research has proven the superior abilities of artificial intelligence (AI) to aid in securing power grids. A common example from related literature is AI-based intrusion detection systems (IDSs) that learn the normal behaviors of power grids to detect when abnormal activities occur [6].

Acknowledgement of National Science Foundation for funding via NSF Grant Award Number 2043324.

A. Related Works

When it comes to enhancing the resilience of smart grids against cyber attacks, existing works focus on either attack detection or classification. Attack detection has been extensively studied [1], [5], [7], which offers limited insight into the nature, source, or cause of the threat, unlike attack classification. We group the attack classification studies based on the consideration of cyber-physical features, use of SHapley Additive exPlanations (SHAP), and adoption of transfer learning as follows.

Naeem et al. [8], Ganjkhani et al. [9], and Aligholian et al. [10] each studied classification systems for power systems but failed to consider the cyber-physical nature of power systems. Instead, they focused only on the physical power measurements in open-source datasets. The work done by Aligholian et al. and Presekal et al. [11] showed the positive effects of using a graph-based classification system to classify events in power systems, but such a study solely utilizes physical power measurements. The work done by Ahmed et al. in [12] is one of the first to consider both cyber and physical features in a graph-based classification system in power systems. However, instead of using a single graph model to classify attacks, the study utilized two individual models to make individual classifications based on the cyber data and then the physical data, which impacts the complexity and efficiency of the model. Additionally, the work did not classify specific cyber attacks but instead classified various events as normal, cyber, physical, or cyber-physical. These works fail to consider the benefits of using a singular graph-based model that is aware of both cyber and physical features at the same time to efficiently classify cyber attacks.

When looking at the use of SHAP in works that classify cyber attacks, few have been applied to power systems. Naeem et al. and Trivedi et al. [13] applied SHAP to better understand and improve their attack classification systems. However, both of these works relied solely on physical power measurements in their systems. Outside of the domain of power systems, Wei et al. [14], Assadhan et al. [15], and Kalutharage et al. [16] all used SHAP on their classification systems that were trained on open-source distributed denial-of-service (DDoS) attack datasets. These works used datasets consisting of only cyber features for IoT systems that do not belong to power systems. The use of SHAP on systems trained on only cyber features or only physical features fails to provide a holistic analysis of the features present in power systems.

Lastly, we consider works that apply transfer learning to adapt classification systems or IDSs to classify or detect new system behaviors. Li et al. [17], Xia et al. [18], and Quraishi et al. [19] all assessed how transfer learning could be used to adapt systems to detect either faults and trips or cyber attacks. However, each of these works only used physical power measurements in their analysis.

Related literature in this domain offers enhanced abilities to secure our power grids, but they suffer from the following limitations: (a) most focus on cyber attack detection instead of classifying cyber attacks in power grids, (b) their analysis does not consider the complexity of the systems and how to make them more efficient, and (c) they do not adapt their systems to classify new attacks as they arise in our power grids. Attack detection provides operators with information that tells them that an anomaly has occurred. On the other hand, classification systems tell operators which cyber attack occurred, which can be used to make informed remediation decisions. This is a valuable characteristic of classification systems, as each cyber attack may have different response strategies required to return the grid to standard working conditions. Furthermore, power grids produce vast amounts of data that AI systems must be trained on. With such large amounts of data, producing more efficient classification systems that require less data to achieve good performance is vital to improve computational and time efficiency. Additionally, these systems must be adapted when new cyber attacks are launched against our power grids. The process of developing new classification systems when each new attack is launched could be computationally taxing and require large amounts of time. We must consider alternative methods for adapting our trained classification systems to improve the process of classifying new cyber attacks against our power grids without starting from scratch.

B. Contributions

The aforementioned limitations motivate the work done in this paper to develop more efficient classification systems that can be easily adapted to classify new cyber attacks in power grid environments. Through this work, we offer the following contributions to further the domain of AI-enhanced security in power grids:

- We use a realistic cyber-physical power system testbed to create benign and malicious datasets consisting of cyber and physical features.
- We develop spatio-temporally aware and unaware classification systems to test the classification abilities of various models for power grid-specific datasets.
- We use SHAP to analyze the importance of each cyber and physical feature in sample classification to produce more efficient classification systems.
- We use transfer learning to adapt our classification system to classify new cyber attacks in power grid environments.

The rest of this paper is organized as follows. Section II defines the testbed used and datasets created to facilitate this work. Section III discusses the methodologies used to efficiently develop our classification systems and adapt them to new attacks. Section IV provides our results for the development and adaptation of the classification systems. Section V discusses our conclusions and directions for future work.

II. Testbed and Data Collection

To facilitate the work done in this paper, we used the Tennessee Technological University cyber-physical power system testbed. The power system testbed consists of two main components: the physical and cyber layers. The former uses a commonly used hardware-in-the-loop simulator, Opal-RT (model OP4610XG) [20], to simulate the physical power system. The latter consists of a Docker container network that simulates the virtual connections between substations and the devices within those substations.

A. Physical Layer

The physical layer of the power system testbed runs the IEEE 14-bus test system to simulate the realistic behaviors of a power grid for our experiments. In the simulation, we use a 6-month load profile for all experiments to mimic consumer consumption patterns over 6 months. The load profile was scaled to 20 minutes to collect data during the experiments. Lastly, the OPAL-RT is configured with a Modbus TCP server that stores each PLC's physical power measurements. Therefore, the physical PLCs exist on the Opal-RT.

B. Cyber Layer

The cyber layer of the testbed uses a Docker container network to represent other commonly found devices in a power grid environment. In total, the Docker network consists of 10 substations, each consisting of at least one of the following devices, which enable us to collect network traffic in addition to the physical power measurements that represent the power grid:

- **PLC** - These Docker containers exist to collect realistic network traffic for the PLCs that exist on the OpalRT.
- **Relay** - Polls power measurements from their respective PLCs and forwards them to ElasticSearch for storage.
- **Router** - Connects all of the substations and devices within the substations together so communication is possible.

TABLE I
CYBER AND PHYSICAL LAYER FEATURES

Cyber	Physical
Source MAC Address	Voltage (V1)
Destination MAC Address	Current (I1)
Source IP Address	Theta
Destination IP Address	Active Power (P)
Packet Size (Bytes)	Reactive Power (Q)
Packet Protocol	1_P-14_P
Source TCP Port	1_Q-14_Q
Destination TCP Port	BreakerStatus_1-BreakerStatus_14
Source UDP Port	BreakerStatus_GEN
Destination UDP Port	BreakerStatus_LOAD

C. Attack Types

To create our malicious datasets for our classification models, we conduct four attacks against the power system testbed. The following attacks are chosen due to their realistic use and impact on our power grids:

- **False Data Injection (FDI) -** The attacker intercepts communications with the victim PLC, flips the breaker for that PLC, and then replays valid power measurements.
- **Ransomware -** A trusted user unknowingly downloads and runs ransomware on the victim PLC. The ransomware exfiltrates data and locks down that PLC.
- **Brute Force -** The attacker brute forces the admin user password of the victim PLC and then flips the breaker for that PLC.
- **Reverse Shell -** The attacker exploits a vulnerable web server running on the victim PLC which provides the attacker with a reverse shell to flip the breaker of that PLC.

D. Data Collection

In the physical layer, the Docker container Relays use ModbusTCP to poll the physical power measurements from each PLC and send those along to ElasticSearch (our data aggregate) for storage and later use. For the cyber layer, we use *tcpdump* on each of the substations' local networks to gather the network traffic for all devices in the testbed. Once the data is collected, we use a data imputation script to join the physical power measurements with the network traffic associated with the devices belonging to the physical power measurements. The result of this process is a dataset where each sample has the cyber and physical features shown in Table I. We collect 55 features in total, where 10 belong to the cyber layer and 45 belong to the physical layer. In Table I, *1_P-14_P* represents the active power and *1_Q-14_Q* represents reactive power for PLCs 1 through 14. The same logic applies for *BreakerStatus_1-BreakerStatus_14* where these features represent the breaker status of PLCs 1 through 14. For these attributes, the values are only present where there are connections between PLCs.

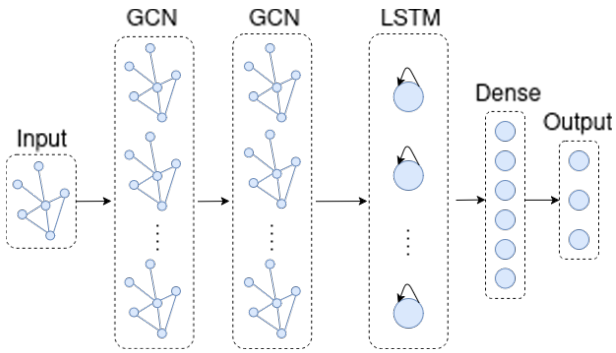

Fig. 1. Illustration of the GCN-LSTM model architecture.

III. PROPOSED CLASSIFICATION METHODOLOGIES

In this section, we introduce our classification system along with benchmarks, followed by the used SHAP analysis to enhance the classification efficiency and the used transfer learning to adapt our classification system to classify new cyber attacks in power grid environments.

A. Classification Models

We adopt a spatio-temporally aware graph-based model and three traditional benchmark deep learning models to show the superior performance of the adopted graph model. Each model is trained on benign, FDI, and ransomware samples as a supervised deep learning model. The remaining attacks are reserved for our later analysis of the adaptability of the classification systems.

1) Model Architecture: Our graph-based model is a graph-convolutional neural network with a long short-term memory (LSTM) layer (GCN-LSTM). Our GCN-LSTM model specializes in learning the spatial and temporal relationships in the datasets. Fig. 1 illustrates the GCN-LSTM model architecture. For a comparative analysis, we also build feed-forward neural network (FNN), convolutional neural network (CNN), and recurrent neural network (RNN)-based benchmark models. These benchmark models give us a robust comparison of the spatio-temporally aware and unaware architectures. In addition to using multiple architectures, we also train the models on different subsets of features to analyze the importance of the cyber and physical features when performing attack classifications. Thus, we train the aforementioned models on cyber-only, physical-only, and cyber-physical features

2) Model Hyperparameters: We adopt a randomized grid search to find the suitable hyperparameters for each model trained on each feature subset. Specifically, we use a sliding window mechanism on our datasets to aid the models in learning the temporal patterns. For the sliding windows, we tune the window size and stride, which results in a window size and stride of 50 and 10, respectively.

B. SHAP Analysis

To improve the efficiency of the classification systems, we use the SHAP analysis to understand the impact of each

feature on the classification of each sample. SHAP works by calculating local and global explanations that define the importance of each feature [21]. Local explanations give us insights as to which features had the greatest impact based on the sample class. Global explanations tell us which features have the overall highest impact, regardless of sample class. In this work, we focus on the best-performing classification systems from our initial analysis of the classification systems. To calculate the SHAP values, we use

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{1}$$

to produce unique explanations where $g(z')$ is a linear version of the model's prediction, ϕ_0 is the prediction when no features are used, ϕ_j shows how much the j^{th} feature changes the prediction, and z'_j indicates whether the j^{th} feature appears in the explanation. The unique explanations produced quantify each feature importance as a numerical value. Using SHAP, we gather which features are the most important for each classification system to make its classification. The results of our SHAP analysis allows us to improve our best-performing classification system by making it more efficient.

C. Transfer Learning

Transfer learning is a method commonly used to adapt models that have been trained to perform a specific task to then perform a different, yet similar, task [22]. In this work, we implement a transfer learning-based approach to show an efficient method to adapt a classification system to classify new attacks, without starting over from scratch. As mentioned in Section III-A, we originally train our classification systems on only benign, FDI, and ransomware samples. The other two attack datasets, brute force and reverse shell, are used in this portion of our analysis to see if we can adapt the models to classify new attacks. For our analysis of adaptability, we focus on the optimal classification system, namely, the GCN-LSTM model trained on the best features from our SHAP analysis.

D. Experimental Setup

In our experiments, we use two methods of transfer learning. For both methods, we take the saved state of the trained GCN-LSTM model from our SHAP analysis and freeze all of the weights for each layer. As shown in Fig. 1, for Method 1, we unfreeze the dense and output layer weights so that they are able to be retrained to learn the patterns of the new attack datasets. In this method, both the GCN and LSTM layer weights are frozen and only utilize the information learned from the initial model training. This process saves time from having to retrain the entire model from scratch, while retaining the information already learned. In Method 2, we unfreeze the LSTM layer weights in addition to the dense and output layer weights. This method provides us with a more granular fine-tuning, which improves performance at the expense of potential time and computation used to adapt those weights. This method is a viable option as the GCN layers are the most complex layers that require the greatest computation to fine-tune. Keeping these weights frozen will save computational resources while retaining the learned spatial relationships between benign and attack datasets. In this work, we test each method in three ways. In the first way, we introduce only the brute force samples into the training dataset with the benign, FDI, and ransomware samples. In the second way, we introduce only the reverse shell samples into the training dataset with the benign, FDI, and ransomware samples. In the third way, we introduce both brute force and reverse shell samples into the training dataset with the benign, FDI, and ransomware samples. The result of our different methods and ways of testing is six unique GCN-LSTM models.

IV. Experimental Results

In this section, we first introduce the performance metrics. Then, we discuss the results for the developed classification systems. We then define the feature importance for our best classification systems and how we develop a more efficient classification system. The last section proves the abilities of transfer learning to adapt our classification systems for new attacks.

A. Performance Metrics

To evaluate the performance of each system, we calculate the accuracy $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$, precision $= \frac{\text{TP}}{\text{TP+FP}}$, recall $= \frac{\text{TP}}{\text{TP+FN}}$, and F1-score $= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision+recall}}$, where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative samples, respectively.

B. Classification Results

As stated in Section III, we train each of our four models three times using: cyber features collected from the power system testbed, physical features collected from the power system testbed, and both the cyber and physical features at the same time. The results of all the models developed are shown in Table II. Through our experiments, we draw a few conclusions based on the feature subsets and model architecture.

1) Feature Subsets: When looking at the feature subsets, we see that the models trained on the cyber-only datasets outperform the physical-only datasets by about $2-3\%$ across the different models. Hence, we conclude that all of the tested architectures lead to increased classification capabilities when looking at the network traffic in comparison to the physical power measurements. However, we also see that all architectures trained on the cyber-physical datasets outperform their cyber-only and physical-only counterparts by $4-5\%$. Thus, we conclude that the inclusion of both cyber and physical features leads to increased performance due to the greater amount of information gained from both network traffic and the physical power measurements at the same time.

2) Model Architecture: According to Table II, our GCN-LSTM graph-based models outperform the benchmark model architectures when classifying cyber attacks by up to 6%. The superior performance of the GCN-LSTM graph-based models

TABLE II
MODEL PERFORMANCE ON DIFFERENT DATASETS (%)

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Physical	FNN	87.3	87.7	87.8	87.7
	RNN	88.3	89	88.5	88.4
	CNN	90	90.2	90	89
	GCN-LSTM	**91.9**	**92.5**	**91.9**	**92**
Cyber	FNN	90.3	90.5	90.8	90.6
	RNN	91	90.9	90.9	90.9
	CNN	92.5	92.5	92.7	92.6
	GCN-LSTM	**93.5**	**93.7**	**93.5**	**93.5**
Cyber-Physical	FNN	91.7	92.1	92.3	92.2
	RNN	94.6	94.7	94.7	94.7
	CNN	95.2	95.2	95.2	95.2
	GCN-LSTM	**97.2**	**97.4**	**97.2**	**97.2**

is due to their innate ability to infer complex patterns in graph-based systems such as power grids. Overall, we conclude that the best performing model is the GCN-LSTM model when trained on both cyber and physical features due to learning the complex spatial and temporal patterns, which are best suited to classify cyber attacks in power grids. Additionally, we conclude that the inclusion of network traffic and the physical power measurements lead to increased classification abilities for diverse deep learning model architectures.

C. SHAP Results

As stated in Section III, we perform our SHAP analysis on the models where we focus on all of the model architectures trained on the cyber-physical features to produce a robust SHAP analysis to understand feature importance. To simplify the results of our analysis, we focus on the top ten features for each architecture to understand the model behavior. Through our SHAP analysis, we see consistent patterns among the FNN, CNN, and GCN-LSTM models, where they have a strong dependence on *tcp.srcport* when classifying all three classes.

1) Benchmark Models: Unlike the RNN model, the FNN and CNN models have a balanced dependence on the cyber and physical features at the same time, relying on five cyber and five physical features in the top ten features. A balanced dependence on the two types of data sources shows how the attacks are shown in both the network traffic and the power measurements. However, a dependence on data from two different sources adds additional complexity when it comes to collecting the samples in real-time, which adds to the overall complexity of using these models. Alternatively, for the RNN model, the most important feature is *BreakerStatus_7* as it exhibits behaviors different than the other architectures. In addition to the unique dependence on *BreakerStatus_7*, the RNN model also exhibits unique behaviors across all of its top ten features. Unlike the other models, the top features with the RNN model are all physical layer features belonging to the physical power measurements.

2) Graph Models: In addition to the results shown from the benchmark models, we see some more unique behaviors from the GCN-LSTM model. When looking at the results for the

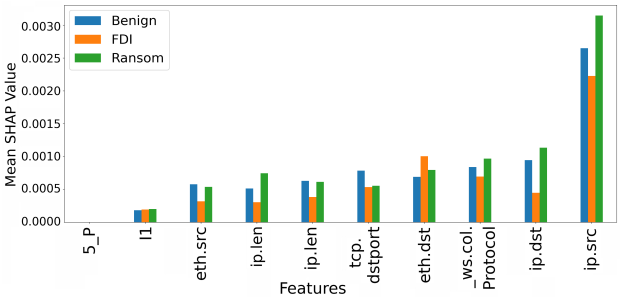

Fig. 2. SHAP results of the GCN-LSTM model showing the best features.

TABLE III
GCN-LSTM PERFORMANCE ON BEST FEATURES (%)

Features Used	Accuracy	Precision	Recall	F1-Score
All 55 Features	97.2	97.4	97.2	97.2
Best 9 Features	**97.3**	**97.4**	**97.2**	**97.2**

GCN-LSTM model in Fig. 2, we can see a stronger reliance on the cyber features in comparison to the physical features. In the top ten features, eight of those features belong to the cyber layer network traffic. A stronger reliance on one of the two data sources can reduce complexity by only needing to collect those specific features. Additionally, we only see the top nine features for the GCN-LSTM model have SHAP values (shown on the y-axis of Fig. 2). This means that only those features are used to make the classifications. The use of limited features means we could make our classification systems more efficient. The top nine features in order of importance are *tcp.srcport*, *ip.src*, *ip.dst*, *_ws.col.Protocol*, *eth.dst*, *tcp.dstport*, *ip.len*, *eth.src*, and *I1 (Current)*. We then retrain the GCN-LSTM model using only these nine features to reduce computational complexity. Specifically, Table III shows the positive results of the GCN-LSTM model trained on only the best features. We see the new GCN-LSTM model maintains its high performance, with all performance metrics matching the performance of the original GCN-LSTM model with a 0.1% increase in accuracy. Through the use of our SHAP analysis, we are able to reduce the number of features needed by roughly 83% while maintaining superior performance.

D. Transfer Learning Results

In addition to improving the efficiency of our classification systems, we must also adapt them when new attacks arise through transfer learning. For this portion of our analysis, we focus on the more efficient, high-performing GCN-LSTM model trained on the best nine features. The results of the new model are presented in Table IV, highlighting the abilities of the two transfer learning methods (introduced in Section III-D. Using Method 1, we see that the models are able to adapt to learn the new attacks and achieve performances around 90% when only one attack is introduced to the training process. However, we see higher performance deterioration when both attacks are introduced into the training process at the same time due to the lack of fine-tunable weights to learn the

TABLE IV
GCN-LSTM WITH TRANSFER LEARNING PERFORMANCE (%)

Method	Attack Added	Accuracy	Precision	Recall	F1-Score
Method 1	Brute Force	89.5	90	89.5	89
	Reverse Shell	91.2	92.1	91.7	91.6
	Both	82.4	82.7	82.5	81.4
Method 2	Brute Force	96.9	96.9	96.9	96.8
	Reverse Shell	98.4	98.5	98.4	98.4
	Both	97.8	97.8	97.8	97.8

patterns in both of the new attacks. Using Method 2, we can see improved performance when each attack is introduced in comparison to Method 1. When the individual attacks are added, we see performance metrics between $96-98\%$. When both attacks are introduced into the training process, we see the performance metrics are around 97%. We see greater performance in Method 2 than Method 1 due to the addition of the LSTM layer weights in the training process to learn the distinct temporal patterns represented in the new attacks. Using Method 2, the performance of the GCN-LSTM model surpasses the performance of the original GCN-LSTM systems trained on only three classes. These results prove the viability of transfer learning to adapt classification systems to classify new attacks in power grid environments.

V. CONCLUSIONS

In this paper, we developed five new cyber-physical datasets demonstrating benign and malicious system behaviors. We developed graph-based and non-graph-based classification systems for power grids. Additionally, we performed a SHAP analysis to understand the importance of the cyber and physical features present in power grids, which are then used to build efficient classification systems. Lastly, we used transfer learning to enhance the adaptability of classification systems when new attacks arise. In this work, we concluded the following. Graph-based deep learning models provide superior performance (by $4-5\%$) when classifying cyber attacks in power grids due to their ability to learn the spatio-temporal patterns in power grids. The best-performing GCN-LSTM classification system can be made more efficient by reducing the feature dimensions by 83% using the results of our SHAP analysis. Transfer learning is a suitable method to adapt GCN-LSTM classification systems to classify new attacks while maintaining performances around $96-98\%$. Specifically, freezing the weights of the GCN layers while unfreezing the LSTM, dense, and output layer weights leads to the best performance. Future work in this domain could focus on larger test systems, such as the IEEE 30-bus or IEEE 118-bus test systems with more diverse benign scenarios, such as faults or trips, to prove the scalability of this work.

REFERENCES

[1] R. Atat, A. Takiddin, M. Ismail, and E. Serpedin, "Graphon neural networks-based detection of false data injection attacks in dynamic spatio-temporal power systems," *IEEE Open Access Journal of Power and Energy*, vol. 12, pp. 24–35, Jan. 2025.

[2] N. K. Barsha and N. Hubballi, "Detecting cyber attacks in smart-grid networks with probability distribution comparison," in *IEEE Consumer Communications Networking Conference (CCNC)*, pp. 648–649, Las Vegas, NV, USA, 06–09 Jan. 2024.

[3] W. Liao, A. Takiddin, M. Tariq, S. Chen, L. Ge, and Z. Yang, "Sample adaptive transfer for electricity theft detection with distribution shifts," *IEEE Transactions on Power Systems*, vol. 39, pp. 7012–7024, Nov. 2024.

[4] CISA, "Cyber-attack against ukrainian critical infrastructure," 2021.

[5] A. Takiddin, M. Ismail, R. Atat, and E. Serpedin, "Spatio-temporal graph-based generation and detection of adversarial false data injection evasion attacks in smart grids," *IEEE Transactions on Artificial Intelligence*, vol. 5, pp. 6601–6616, Dec. 2024.

[6] J. Sweeten, A. Elshazly, A. Takiddin, M. Ismail, S. S. Refaat, and R. Atat, "Cyber-physical fusion for gnn-based attack detection in smart power grids," *IEEE Open Access Journal of Power and Energy*, vol. 12, pp. 515–528, July 2025.

[7] A. Takiddin, M. Ismail, R. Atat, K. R. Davis, and E. Serpedin, "Robust graph autoencoder-based detection of false data injection attacks against data poisoning in smart grids," *IEEE Transactions on Artificial Intelligence*, vol. 5, pp. 1287–1301, Mar. 2024.

[8] H. Naeem, F. Ullah, and G. Srivastava, "Classification of intrusion cyber-attacks in smart power grids using deep ensemble learning with metaheuristic-based optimization," *Expert Systems*, vol. 42, p. e13556, Feb. 2025.

[9] M. Ganjkhani, M. Gilanifar, J. Giraldo, and M. Parvania, "Integrated cyber and physical anomaly location and classification in power distribution systems," *IEEE Transactions on Industrial Informatics*, vol. 17, pp. 7040–7049, Oct. 2021.

[10] A. Aligholian and H. Mohsenian-Rad, "Graphpmu: Event clustering via graph representation learning using locationally-scarce distribution-level fundamental and harmonic pmu measurements," *IEEE Transactions on Smart Grid*, vol. 14, pp. 2960–2972, July 2023.

[11] A. Presekal, A. Ştefanov, V. S. Rajkumar, and P. Palensky, "Attack graph model for cyber-physical power systems using hybrid deep learning," *IEEE Transactions on Smart Grid*, vol. 14, pp. 4007–4020, Sept. 2023.

[12] A. Ahmed, S. Basumallik, A. Gholami, S. K. Sadanandan, M. H. N. Namaki, A. K. Srivastava, and Y. Wu, "Spatio-temporal deep graph network for event detection, localization, and classification in cyber-physical electric distribution system," *IEEE Transactions on Industrial Informatics*, vol. 20, pp. 2397–2407, Feb. 2024.

[13] R. Trivedi, S. Patra, and S. Khadem, "Data-centric explainable artificial intelligence techniques for cyber-attack detection in microgrid networks," *Energy Reports*, vol. 13, pp. 217–229, June 2025.

[14] Y. Wei, J. Jang-Jaccard, A. Singh, F. Sabrina, and S. Camtepe, "Classification and explanation of distributed denial-of-service (DDoS) attack detection using machine learning and shapley additive explanation (SHAP) methods," June 2023.

[15] B. Assadhan, A. Bashaiwth, and H. Binsalleeh, "Enhancing explanation of lstm-based ddos attack classification using shap with pattern dependency," *IEEE Access*, vol. 12, pp. 90707–90725, July 2024.

[16] C. S. Kalutharage, X. Liu, C. Chrysoulas, N. Pitropakis, and P. Papadopoulos, "Explainable ai-based ddos attack identification method for iot networks," *Computers*, vol. 12, Feb. 2023.

[17] H. Li, Z. Ma, and Y. Weng, "A transfer learning framework for power system event identification," *IEEE Transactions on Power Systems*, vol. 37, pp. 4424–4435, Nov. 2022.

[18] Y. Xia, Y. Xu, S. Mondal, and A. K. Gupta, "A transfer learning-based method for cyber-attack tolerance in distributed control of microgrids," *IEEE Transactions on Smart Grid*, vol. 15, pp. 1258–1270, March 2024.

[19] A. Quraishi, M. A. Rusho, A. Prasad, I. Keshta, R. Rivera, and M. W. Bhatt, "Employing deep neural networks for real-time anomaly detection and mitigation in iot-based smart grid cybersecurity systems," in *Int. Conf. on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pp. 1–6, Hyderabad, India, 26–27 April 2024.

[20] OPAL-RT Technologies, "OP4610XG Simulator." https://www.opal-rt.com/simulator-platform-op4610xg/. Accessed: Jul. 2025.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 4766–4777, Long Beach, CA, USA, 04–09 Dec. 2017.

[22] W. Liao, R. Zhu, A. Takiddin, M. Tariq, G. Ruan, X. Cui, and Z. Yang, "Transfer learning-driven electricity theft detection in small sample cases," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–13, Oct. 2024.