X-HDNN: Explainable Hybrid DNN for Industrial Internet of Things Backdoor Attack Detection

Love Allen Chijioke Ahakonye, Cosmas Ifeanyi Nwakanma †, Jae Min Lee, Dong-Seong Kim
IT Convergence Engineering, † ICT Convergence Research Center,
Kumoh National Institute of Technology Gumi, South Korea
loveahakonye, cosmas.ifeanyi, ljmpaul, dskim@kumoh.ac.kr

***Abstract*—This study proposes a hybrid deep neural network (HDNN) framework, called X-HDNN, for detecting backdoor attacks in Industrial Internet of Things (IIoT) data. The X-HDNN combines LeakyReLU and focal loss functions to reduce false positives and losses. A comparative analysis of the performance of the primary deep neural network and the proposed X-HDNN had remarkable improvement in f-score value from 57% to 78% and loss of 0.0044 to 0.0014. It also incorporates the SHAP explainability technique to provide interpretable and reliable detection. Evaluating the X-HDNN model using backdoor impact ratio, feature importance score, and decision confidence helps understand the model's outcomes and the significance of each feature. The findings enhance trust in the model and facilitate better decision-making based on the provided explanations.**

***Index Terms*—Backdoor, Decision, and Confidence Impact Ratios, Deep Neural Networks, Focal Loss, LeakyReLU, SHAP, XAI,**

I. Introduction

Deep learning techniques have been extensively utilized in Industrial Internet of Things (IIoT) intrusion detection systems (IDS) [1], necessitating enhanced model security. However, IIoT applications continue to face threats from unauthorized access and attacks that exploit vulnerabilities in authentication mechanisms, potentially compromising critical infrastructures [1], [2]. Backdoor attacks pose a significant risk among these threats, as they disrupt the operational reliability, safety, and efficiency of critical infrastructures [2].

To mitigate the impact of unauthorized access on IDS model training caused by backdoor attacks, this study proposes an explainable hybrid deep neural network (X-HDNN) that addresses high false positives and loss in imbalanced data. The X-HDNN incorporates the Leaky rectified linear unit (LeakyReLU) activation function and focal loss technique to effectively manage false positives, reduce loss, and maintain classification accuracy. Additionally, an explainable artificial intelligence (XAI) technique, precisely the SHapley Additive exPlanations (SHAP), is employed to provide comprehensive interpretations of attack classifications, enhancing trustworthiness and model transparency [3], [4].

The X-HDNN framework leverages LeakyReLU to overcome the "dying ReLU" problem and improve gradient flow during training. The focal loss handles imbalanced data by assigning different weights to challenging samples, downweighing the contribution of well-classified instances. Combining LeakyReLU and focal loss reduces loss, mitigates false positives, and improves classification performance. The SHAP explainability approach further enhances attack classifications' interpretability and visual explainability, addressing the need for comprehensibility and trust in black box IDS algorithms [4], [5].

Specifically, this study focuses on the following:

1) A hybrid deep neural IDS framework incorporating LeakyReLU and focal loss to reduce false positives and loss; and model explainability to enhance the interpretability of model decisions for attack classification.
2) A substantial experimentation showing the model's performance with traditional activation and cross-entropy loss functions and the proposed hybrid of LeakyReLU activation and focal loss functions.
3) Presentation of the SHAP explainability technique assessment regarding impact decisions and comprehensibility proficiency for model transparency.

The study arrangement is thus: Following the introduction in Section I, Section II reviews existing works on backdoor attack detection utilizing deep neural networks. Section III discusses the proposed hybridized deep neural network model with XAI. Section IV highlights the experimentation environment and results. Section V concludes the study.

II. Related Works

Several studies have proposed deep neural networks to defend against backdoor attacks using pruning, tuning defense schemes, and analyzing internal activation values [6]–[8]. Another approach involves identifying prompts and utilizing entropy allocation to enhance defense mechanisms [9]. These studies collectively contribute to developing robust defenses against backdoor attacks.

A study by [10] proposed a conditional generative model that effectively identifies and prevents backdoor attacks by learning the likely distribution of possible prompts. In a related study, [11] extended this technique by demonstrating differential privacy's efficiency in detecting outliers of backdoor attack detection. Aimed at rendering backdoor attacks ineffective, [12] implemented a data augmentation policy and preprocessing input samples to invalidate prompts during the inference phase; [13] highlighted that the complexity of current defense methods increases as class labels are added. They suggest a more effective strategy that utilizes the K-arm optimization method to simplify the defense approach and

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

handle models with numerous classes. Although research contributions enhance the prevention and detection of backdoor attacks in various scenarios, model loss, and high false positives remain critical issues, particularly in the heterogeneous IIoT sensor data.

A novel approach named DeepGuard was introduced for privacy-preserving backdoor detection and identification in a multi-participant computation scenario within an outsourced cloud environment [14]. While the efficacy and efficiency of the backdoor attack detection and identification algorithm were demonstrated, there remains room for enhancing the model's handling of loss, reducing false positives, and improving the interpretability of its decision-making process.

III. Methodology

This section briefly overviews the X-HDNN model for backdoor attack detection and a systematic process of SHAP explainability interpretation leveraging the backdoor impact ratio, feature impact score, and the decision confidence score to evaluate the model decisions for backdoor attack classification. Fig. 1 is the pipeline of the system model process flow.

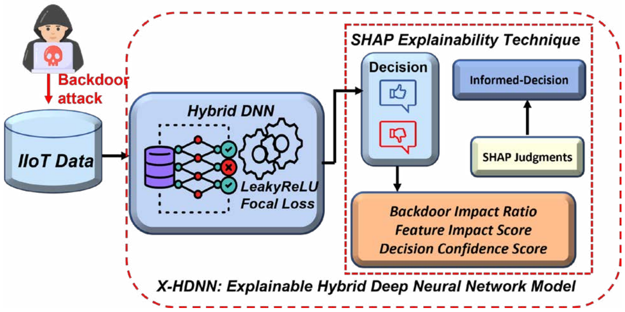

Fig. 1. The pipeline process flow of the proposed X-HDNN

A. Characterization of Backdoor Attacks

Backdoor attacks involve intentionally inserting hidden vulnerabilities or malicious functionality into a system, enabling unauthorized access or control [2]. These attacks manipulate data or insert hidden patterns to bypass security measures and compromise system integrity, resulting in high false positives and increased model error loss. Preventing and mitigating backdoor attacks is crucial for maintaining the security and integrity of network infrastructures.

Commonly used sophisticated strategies such as BadNets and trojan attacks enable backdoor attacks [10]. The BadNets approach injects backdoor attacks by introducing randomly chosen clean data triggers and modifying their labels [15]. On the other hand, trojan attacks allow intrusion into pre-trained deep neural network models without compromising clean datasets, achieved through reverse engineering and the generation of trojan triggers and training data [16]. These studies shed light on diverse backdoor attack strategies and their implications for machine learning model security, including false model performance in terms of high false positives or losses.

B. Hybrid Deep Neural Network Architecture for Backdoor Attack Detection

The proposed X-HDNN structure consists of multiple dense layers with LeakyReLU activation functions and dropout regularization to prevent overfitting. The network architecture includes three dense layers with 64, 32, and 16 neurons, each followed by a LeakyReLU activation function. Dropout layers with a rate of 0.5 are inserted after the first two dense layers. The output layer employs the softmax activation function for multi-class classification. During training, the model optimizes error loss and false positives using the focal loss function combined with the softmax function.

Given x as the output of a layer, Where W_1, W_2, W_3 are the weight matrices, b_1, b_2, b_3 are the bias vectors, α is the slope parameter for LeakyReLU activation, p is the dropout rate, Softmax is the softmax activation function and y_{true}, y_{pred} represents the focal loss function. The model's performance is optimized by adjusting these parameters to match the true class labels with the predicted probabilities. The hybrid deep neural network framework is implemented as follows in Equations 1, 2, 3, 4, 5 and 6 below:

$$x = W_1 \cdot \text{inputs} + b_1 x = \text{LeakyReLU}(x, \alpha), \tag{1}$$

$$x = \text{Dropout}(x, p), \tag{2}$$

$$x = W_2 \cdot x + b_2 x = \text{LeakyReLU}(x, \alpha), \tag{3}$$

$$x = \text{Dropout}(x, p), \tag{4}$$

$$x = W_3 \cdot x + b_3 x = \text{LeakyReLU}(x, \alpha), \tag{5}$$

$$\text{outputs} = \text{Softmax}(W_4 \cdot x + b_4). \tag{6}$$

The Python Keras, TensorFlow, and PyTorch libraries define the LeakyReLU, dropout, and softmax functions and the weight matrices W_1, W_2, W_3, and W_4 and bias vectors b_1, b_2, b_3, and b_4 specific to the X-HDNN architecture are initialized during the training process. Algorithm 1 summarizes the process flow of the X-HDNN for backdoor attack detection.

C. SHAP Model Explainability

Explainability is a crucial aspect of artificial intelligence (AI) in cybersecurity, encompassing various predictive models across domains. The utilization of SHAP values allows for a deeper understanding of the impact of each input feature on the model's decision-making process. In detecting backdoor attacks, the SHAP technique proves valuable in identifying the significant features influenced by backdoor triggers. [17] gave detailed explanations on the calculations and application of SHAP, emphasizing its importance in XAI for cybersecurity. The SHAP explainability interpretation is assessed based on the metrics below:

Algorithm 1: Explainable Hybrid Deep Neural Network

```
Input: $df$, where $df$ = Set of classified instances
Output: X-HDNN
Require: $df \neq \emptyset, num\_attributes > 0$
procedure BUILD HDNN WITH SHAP($a$)
    Data Preprocess
    Create the model architecture
    Compile Model
    Define early stopping criteria
    Model Test, validate and using early stopping
    Evaluate Model Prediction
    Generate SHAP Interpretation for Model Predictions
    Evaluate generated SHAP interpretations using
  decision and confidence impact ratios
end procedure
```

- Backdoor Impact Ratio (BIR) measures the proportion of backdoor instances correctly classified as backdoor attacks. It is calculated as the ratio of true positive backdoor instances to the total number of backdoor instances as shown in equation 7.

$$\text{BIR} = \frac{\text{Total Backdoor Instances}}{\text{True Positive (Backdoor)}}. \tag{7}$$

- Feature Importance Score (FIS) quantifies the contribution of each feature in the decision-making process of backdoor attack detection. It is calculated by analyzing the SHAP values for the affected features and summing their magnitudes in equation 8 below:

$$\text{FIS} = \sum_{i=1}^{n} \left| \text{SHAP Value}_{\text{feature}_i} \right|. \tag{8}$$

- Decision Confidence Score (DCS) measures the overall confidence of the model in making correct decisions regarding backdoor attack detection. It is the average confidence score of correctly classified backdoor instances as in equation9.

$$\text{DCS} = \frac{\sum_{i=1}^{k} \text{Confidence Score}_{\text{backdoor instance}_i}}{k}. \tag{9}$$

The total backdoor instances represent the total number of backdoor occurrences, and the true positive (Backdoor) is the number of correctly classified backdoor occurrences in equation 7. The SHAP Value in equation 8 is the specific feature for i ranging from 1 to n, and the total number of features considered for a backdoor, n, corresponds to the range of features contributing to the decision-making process. The confidence score in equation 9 is the value assigned to a backdoor instance, where k is the total number of correctly classified backdoor instances, denoting the count of instances where the model correctly identifies an instance as a backdoor attack.

Assessing the SHAP explainability of the X-HDNN for backdoor attack detection offers security experts the opportunity to subjectively evaluate their level of trust in the model and assess its decisions. It shows the confidence and reliability of the model's outputs, enabling informed judgments and decisions in backdoor attack detection and mitigation.

IV. Experimentation and Result Discussion

A. Dataset Description and Experimentation Environment

This study utilizes the WUSTL-IIoT-2021 dataset, a well-established dataset in cybersecurity research focused on Industrial Internet of Things networks [18], [19][1]. It spans approximately 53 hours of data samples, totaling 1,194,464 observations, 41 features, 87,016 attacks, and 1,107,448 normal samples. Amongst the 7.28% diverse attack types, the backdoor constitutes 0.25%, which is the focus of this study. The dataset was split into training (60%), testing (25%), and validation (15%) sets. The *StandardScaler* technique was applied to obfuscate the relationship between the original feature values and the classification, preventing direct inference of sensitive data. The choice of the dataset is due to its relevance to IIoT network cyberattacks and its ability to mimic authentic industrial systems. The experimentation was with Python on Google Colaboratory, running on a system with an Intel(R) Core(TM) i5-8500 CPU @ 3.00GHz and 8GB RAM, operating on Windows 11.

B. Proposed Hybrid Deep Neural Network Model Performance and Evaluation Metrics

The X-HDNN evaluated the backdoor attack classification using accuracy, f-score, precision, and recall. The priority of the f-score in this study is due to its ability to balance precision and recall, making it suitable for imbalanced scenarios. Unlike accuracy, which can be misleading in imbalanced datasets, the f-score considers precision and recall, comprehensively assessing the model's performance. The f-score is calculated as in equation 10 below:

$$\text{f-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

The f-score correctly identified the few backdoor instances while minimizing false positives and negatives, ensuring reliable and accurate detection. The X-HDNN performance in terms of f-score in Fig. 2 demonstrates the significant prediction of the few backdoor samples in highly imbalanced WUSTL-IIoT-2021 dataset with 78% f-score, 93% accuracy, 77% recall, 78% precision. Furthermore, the precision, recall, and f-score performance indices highlight the ambiguity associated with depending on accuracy alone, particularly in heterogeneous IIoT networks with highly imbalanced scenarios. Fig. 3 is the confusion matrix classification report of the few backdoor attack instances against normal instances showing only four (4) misclassified instances out of 52 total

[1] https://ieee-dataport.org/documents/wustl-iiot-2021

backdoor samples. Similarly, Figs. 4 and 5 illustrate the X-HDNN learning process path performance with a minimal loss of 0.014 upon test and validation.

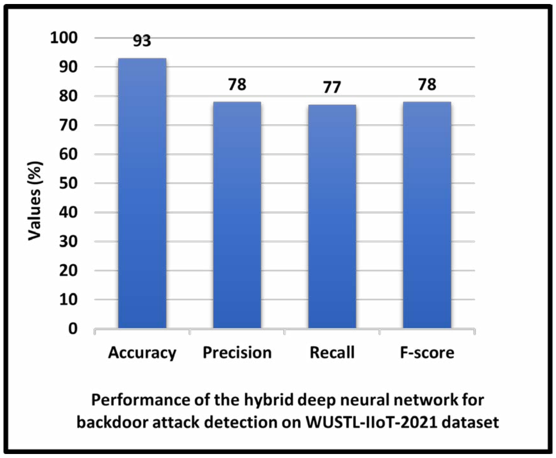

Fig. 2. Graph showing the X-HDNN performance detecting backdoor attacks in the WUSTL-IIoT-2021 dataset.

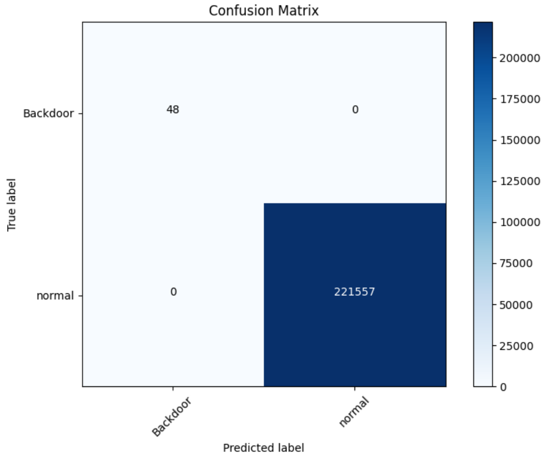

Fig. 3. Confusion metrics illustrating the X-HDNN capacity to significantly detect 48 out of 52 instances of backdoor attack samples

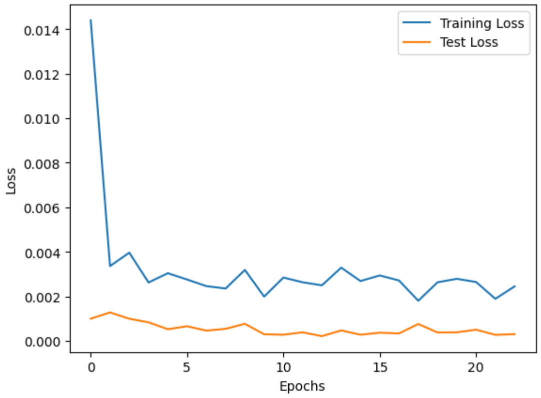

Fig. 4. Learning process path graph of loss versus epoch demonstrating the X-HDNN performance with minimal error loss during the testing phase.

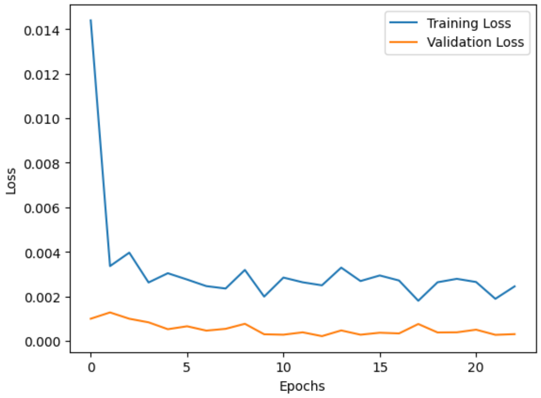

Fig. 5. Learning process path graph of loss versus epoch demonstrating the X-HDNN performance with minimal error loss during the validation phase.

TABLE I

EXPLAINABILITY ASSESSMENT OF THE SHAP INTERPRETATION OF THE X-DNN CLASSIFICATION BASED ON THE SHAP VALUE (0.0078) AND CONFIDENCE SCORES (0.9922) AND 78% F-SCORE (52 BACKDOOR INSTANCES AND 48 INSTANCES OF TRUE POSITIVE PREDICTIONS)

Basic Deep Neural Network		**Hybrid Deep Neural Network**
Accuracy (%)	93	**93**
Recall (%)	40	**77**
Precision (%)	95	**78**
Loss	0.044	**0.014**
F-score	**57**	**78**
SHAP Explainability Interpretation		
Backdoor impact ratio		**1.3**
Feature importance score		**0.0078**
Decision confidence score		**0.9922**

C. Explainable Hybrid Deep Neural Network Model Performance

The assessment of backdoor impact ratio, feature importance, and decision confidence score by the SHAP explainability approach allows a subjective evaluation of the X-HDNN classification of some level of trust and assesses its decisions. This analysis provides a comprehensive understanding of the confidence and reliability of the model's outputs, enabling informed judgments and decisions in backdoor attack detection.

The metrics presented in Table I provide valuable insights into the X-HDNN detection of backdoor instances. The comparative analysis of the performance of the primary deep neural network and the proposed X-HDNN shows an improved f-score of 78% from 57%, demonstrating the hybridized LeakyReLU's efficiency and focal loss. A backdoor impact ratio of 1.3 indicates a higher prevalence of backdoor instances, while a feature importance score of 0.0078 suggests low importance with a slight influence. The high decision confidence score of 0.9922 reflects the model's strong confidence in classifying backdoor instances. These metrics contribute to understanding the backdoor instances' prevalence, importance, and associated features, as depicted in Fig 6.

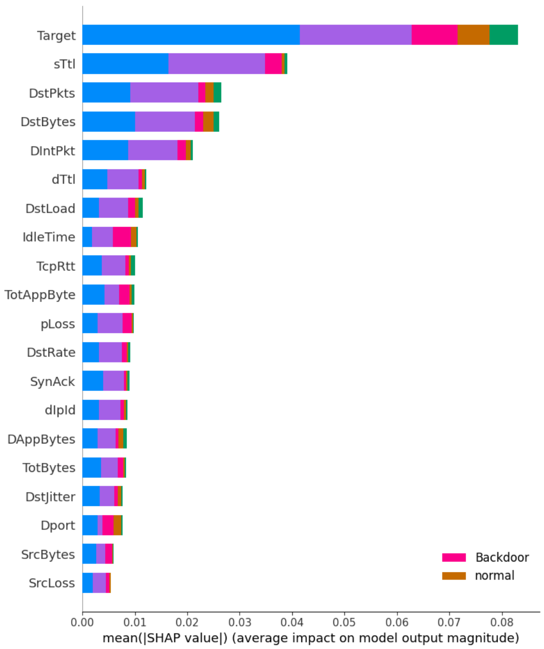

Fig. 6. Graph showing the average impact of the data features significant for backdoor attack prediction

V. Conclusion

This study presented the hybridization of the LeakyReLU activation function and focal loss to mitigate high false positives and loss. It analyzed the outcome with the SHAP interpretation's effects using decision confidence, feature importance scores, and backdoor impact ratio. A comparative analysis of the experimentation of the primary deep neural network and the proposed X-HDNN showed significant improvement in reducing false positives and loss. Interpretation of the SHAP explainability provides different perspectives to assess the robustness, reliability, and interpretability of the X-HDNN predictions. The experimentation result facilitates more comprehensibility of the classifier's decisions to understand the factors influencing the model's decisions and provide insights into the importance and impact of each feature. It aids in building trust in the model and enables better decision-making based on the explanations provided by the SHAP explainability technique. In the future, we intend to expand the scope of the study by exploring other explainability techniques' interpretability.

Acknowledgment

This research was supported by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003), and by MSIT under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP, and the project for Industry-University-Research Institute platform cooperation R&D funded Korea Ministry of SMEs and Startups in 2022.(S3311338)

References

[1] L. A. C. Ahakonye, C. I. Nwakanma, J. M. Lee, and D.-S. Kim, "Agnostic CH-DT Technique for SCADA Network High-Dimensional Data-Aware Intrusion Detection System," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10 344–10 356, 2023.

[2] B. Hou, J. Gao, X. Guo, T. Baker, Y. Zhang, Y. Wen, and Z. Liu, "Mitigating the Backdoor Attack by Federated Filters for Industrial IoT Applications," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 5, pp. 3562–3571, 2022.

[3] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "Trust xai: Model-agnostic explanations for ai with a case study on iiot security," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 2967–2978, 2023.

[4] C. I. Nwakanma, L. A. C. Ahakonye, J. N. Njoku, J. C. Odirichukwu, S. A. Okolie, C. Uzondu, C. C. Ndubuisi Nweke, and D.-S. Kim, "Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review," *Applied Sciences*, vol. 13, no. 3, p. 1252, 2023.

[5] L. Zou, H. L. Goh, C. J. Y. Liew, J. L. Quah, G. T. Gu, J. J. Chew, M. P. Kumar, C. G. L. Ang, and A. W. A. Ta, "Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 2, pp. 242–254, 2023.

[6] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against Backdooring Attacks on Deep Neural Networks," in *International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 273–294.

[7] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "Abs: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1265–1282.

[8] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 707–723.

[9] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A Defence Against Trojan Attacks on Deep Neural Networks," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 113–125.

[10] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks," in *IJCAI*, vol. 2, no. 5, 2019, p. 8.

[11] M. Du, R. Jia, and D. Song, "Robust Anomaly Detection and Backdoor Attack Detection via Differential Privacy," *arXiv preprint arXiv:1911.07116*, 2019.

[12] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, "Deepsweep: An Evaluation Framework for Mitigating DNN Backdoor Attacks using Data Augmentation," in *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2021, pp. 363–377.

[13] G. Shen, Y. Liu, G. Tao, S. An, Q. Xu, S. Cheng, S. Ma, and X. Zhang, "Backdoor Scanning for Deep Neural Networks through K-arm Optimization," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9525–9536.

[14] C. Chen, L. Wei, L. Zhang, Y. Peng, J. Ning *et al.*, "DeepGuard: Backdoor Attack Detection and Identification Schemes in Privacy-Preserving Deep Neural Networks," *Security and Communication Networks*, vol. 2022, 2022.

[15] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating Backdooring Attacks on Deep Neural Networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[16] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning Attack on Neural Networks," in *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018. [Online]. Available: http:dx.doi.org/10.14722/ndss.2018.23291

[17] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in Cybersecurity: A Survey," *IEEE Access*, vol. 10, pp. 93 575–93 600, 2022.

[18] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6822–6834, 2019.

[19] *WUSTL-IIOT-2021 Dataset for IIoT Cybersecurity Research*, 2021. [Online]. Available: http://www.cse.wustl.edu/ jain/iiot2/index.html