Self-Verifying Anomaly Detection using Explainable AI for Cybersecurity of DER Networks

Damilola Popoola, Souradeep Bhattacharya, and Manimaran Govindarasu
Department of Electrical and Computer Engineering
Iowa State University, Ames, IA 50011
Email: popson@iastate.edu, sbhatta@iastate.edu, gmani@iastate.edu

***Abstract*—The rapid growth of Distributed Energy Resources (DERs) has significantly expanded the cyber-attack surface of modern power grids. Furthermore, increasing sophistication in attack techniques demands anomaly detection systems (ADS) that are accurate, interpretable, and reliable to support DER cybersecurity. While ML-based ADS provide strong detection capabilities, their "black-box" nature reduces operator trust and limits Security Operation Center's (SOC) ability to effectively interpret alerts and respond, highlighting the need for explainable Artificial Intelligence (XAI) to ensure transparency and operational confidence. This paper presents an XAI-based anomaly detection framework tailored for DER networks (ExCYDER). The proposed framework uses a self-verifying mechanism that validates ADS alerts to ensure trustworthy decision-making. ExCYDER combines LightGBM with SHAP to check whether each model decision aligns with its feature-attribution evidence, allowing the system to confirm that its internal reasoning is consistent and reliable. Experiments on a realistic DNP3 dataset achieved over 98% detection accuracy, an average rule–SHAP consistency of 44.6%, a SHAP latency of 14.5 ms per alert, and a confidence deviation within $\pm$5%, demonstrating stable verification behavior with minimal computational overhead. The framework distinguished between coherent and inconsistent alerts without compromising detection accuracy, demonstrating that integrated verification within XAI-based ADS enhances interpretability, auditability, and operational robustness for DER-focused SOCs.**

***Index Terms*—Distributed Energy Resources, Explainable Artificial Intelligence, Machine Learning, Anomaly Detection, Security Operations Centers**

I. Introduction

Distributed Energy Resources (DERs) in modern power grids have transformed the energy landscape, enabling greater integration of renewable energy sources, enhanced grid flexibility, and improved energy efficiency [1]. However, this transformation has simultaneously introduced unprecedented cybersecurity challenges that threaten the stability, reliability, and security of critical energy infrastructure [2]. As DER systems become increasingly interconnected, they present an expanded attack surface that malicious actors can exploit to compromise grid operations [3]. Traditional security approaches that depend on fixed, perimeter-based defenses are increasingly ineffective against modern, adaptive cyber threats. Although Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL)-based anomaly detection systems (ADS) provide strong real-time detection capabilities [4], their deployment in critical infrastructure is hindered by the inherent "black-box" behavior of many models [5]. This opacity creates a trust deficit, as Security Operations Center (SOC) analysts require clear, justifiable explanations for security decisions.

Recognizing this gap, the field of Explainable Artificial Intelligence (XAI) has emerged to address the interpretability of AI models [6], but mere explainability is not sufficient. True reliability and interpretability require not only an explanation but also verification of that explanation. Considering the critical nature of DER networks, this paper presents an XAI-based ADS framework for enhancing the cybersecurity of DER networks (ExCYDER). The primary contributions of this work are as follows: (1) development of ensemble learning-based ADS with an explainability mechanism for interpretable and reliable cybersecurity analytics tailored for DER SOC environments; (2) development of an explainability-driven verification process that validates model decisions through rule–feature consistency analysis, enhancing the transparency, confidence, and auditability of SOC-level anomaly investigation; and (3) real-time testbed-based evaluation of ExCYDER using realistic DER datasets to demonstrate its detection accuracy, verification stability, and interpretability performance.

The remainder of this paper is organized as follows. Section II provides background on DER communication and cybersecurity challenges. Section III reviews related work in anomaly detection and explainable AI. Section IV presents the proposed ExCYDER methodology, including the edge-cloud workflow, explainability mechanisms, and verification logic. Section V describes the experimental setup, performance metrics, and evaluation results. Finally, Section VI concludes the paper and outlines directions for future work.

II. Background on DER Cybersecurity

A. DER Communication Architecture

The DER communication structure in this work adopts the three-tier DER architecture described in our earlier framework, CySIDER [7]. It comprises: (1) *DER clients and devices* such as solar PVs, battery systems, and EVs that regulate local generation and storage; (2) an *edge intelligence layer* that hosts anomaly detection and secure telemetry exchange through Edge Intelligent Devices (EIDs); and (3) a *utility and aggregator layer* that coordinates distributed assets and manages grid interaction through the DER Management System (DERMS). Standardized communication protocols (e.g.,

979-8-3315-8138-1

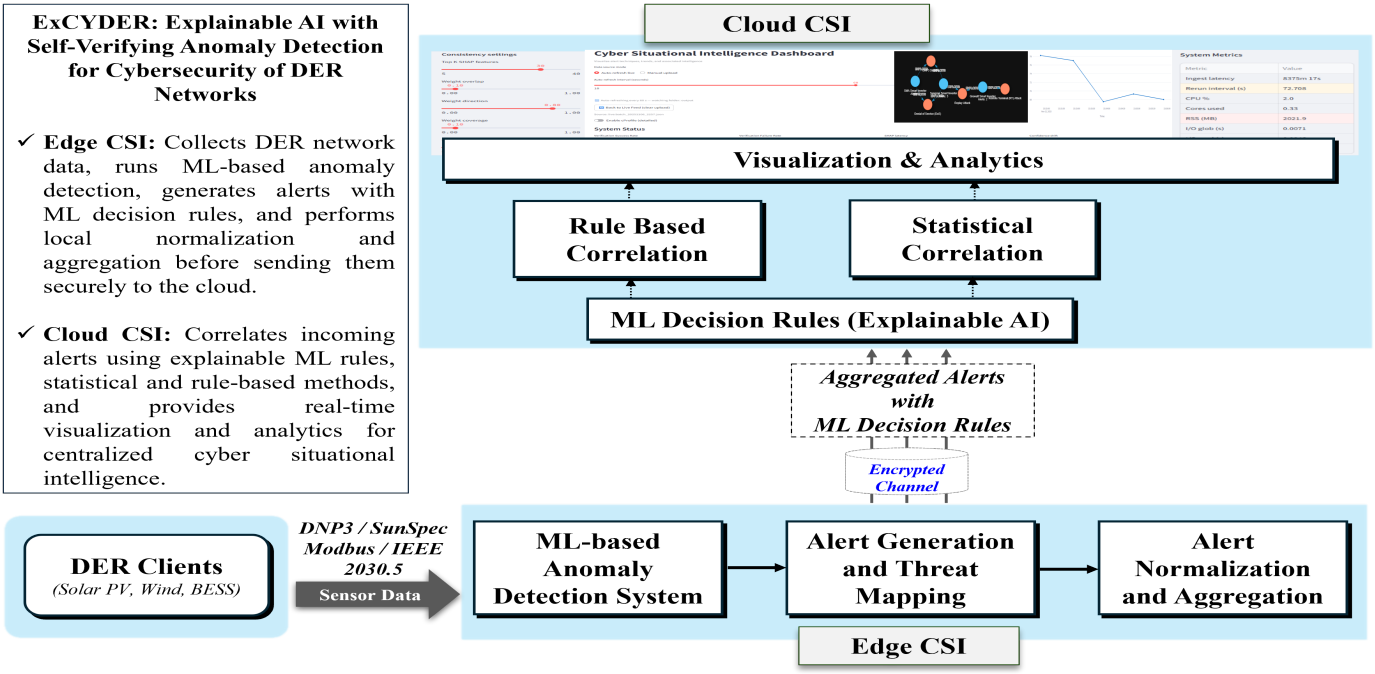

Fig. 1. ExCYDER workflow from edge-based anomaly detection and alert aggregation to cloud-level verification, correlation, and visualization

DNP3, Modbus, IEC 61850, IEEE 2030.5) enable interoperability across heterogeneous devices but are limited by weak or absent authentication and encryption mechanisms.

B. Cybersecurity Challenges

As detailed in [7], the growing digital interconnection of DER assets has significantly expanded the grid's attack surface. Protocols such as DNP3 and SunSpec Modbus were not originally designed with built-in cryptographic safeguards, exposing field devices, including smart inverters, EIDs, and vendor gateways to threats such as replay, spoofing, denial-of-service, and data manipulation. These vulnerabilities, combined with fragmented alert streams and opaque ML models, hinder timely detection and verification within SOCs. Consequently, modern DER cybersecurity requires explainable, verifiable, and SOC-integrated anomaly detection frameworks such as ExCYDER to ensure reasoning transparency, incident traceability, and operational resilience in distributed power networks.

C. Need for Explainability in ML based Anomaly Detection

DER networks increasingly require transparent anomaly detection models rather than traditional "black box" approaches. Although DL and ensemble methods achieve high accuracy, their opacity limits operator trust and slows incident response. XAI improves confidence by revealing the features behind each prediction and supporting debugging and auditability [5], [6]. However, post hoc explanations can be inconsistent or misaligned with model behavior, reducing their reliability in high stakes settings. This motivates the need for approaches that not only explain model outputs but also verify the consistency and reliability of those explanations.

D. Related Work

Traditional anomaly detection in power systems has relied on statistical models and threshold-based alarms, which often fail to cope with the complexity and high-dimensional data streams of modern DER networks. Recent research has shifted toward ML and DL techniques, which offer superior capability in identifying subtle attack patterns. For instance, ensemble methods and neural networks have been explored for detecting data injection and denial-of-service attacks in smart grid environments [8], [9]. While these models achieve high detection accuracy, their "black-box" nature poses a critical challenge for SOC analysts, who require clear, justifiable evidence to triage and respond to alerts in real time.

The field of XAI has emerged to bridge the gap between high model performance and interpretability. In the context of cybersecurity, XAI enhances transparency, fosters trust, and facilitates collaboration between human analysts and AI tools [10], [11]. XAI methods such as LIME and SHAP have been applied to energy systems to improve model transparency [12], [13]. These techniques assist analysts in understanding model predictions and debugging anomalous behavior. Recent studies also highlight the importance of interpreting power grid anomalies from a human centered perspective [14].

However, these XAI applications provide only post hoc explanations and do not verify the internal consistency of model reasoning. This limits their usefulness in SOC environments where rapid and reliable interpretation is required.

III. Proposed Methodology

To address the limitations identified in existing explainable anomaly detection approaches, this work proposes the ExCYDER framework, designed to support DER-specific SOC

operations as shown in Fig 1. ExCYDER integrates ensemble-based anomaly detection with a self-verifying explainability mechanism that internally validates each model decision through consistency analysis. The framework operates across two hierarchical layers: (1) an *edge intelligence layer* that performs real-time anomaly detection and alert generation near DER devices, and (2) a *cloud intelligence layer* that aggregates, verifies, and visualizes alerts to support analyst interpretation and situational intelligence.

A. ML Based Anomaly Detection

At the edge intelligence layer, ExCYDER adopts the ML-Based ADS framework developed in [15]. The adopted ML ADS pipeline performs systematic data preprocessing, including label encoding of categorical variables, feature selection to remove redundant or low variance attributes, and normalization using Min–Max scaling, followed by dataset balancing through the Synthetic Minority Oversampling Technique (SMOTE). The balanced dataset is then divided into 70% training and 30% testing subsets for model evaluation.

B. Model Explainability using SHAP and Decision Rules

ExCYDER derives its interpretability from the combined use of model decision rules and SHapley Additive exPlanations (SHAP). The decision rules extracted from the LightGBM model capture the logical path leading to each prediction, while SHAP quantifies the contribution of each feature to that decision. Together, they represent the model's decision logic and supporting evidence. The alignment between these two components forms the foundation of ExCYDER's verification process, enabling the system to evaluate how consistently the model's explanations reflect its internal decision logic.

C. Cloud-Level Verification

At the cloud layer, ExCYDER verifies each aggregated alert by cross-checking the model's rule-based decision path with corresponding SHAP feature attributions. This determines whether the model's explanation (evidence) aligns with its internal decision logic (claim), producing a consistency score for each alert before it is presented on the SOC dashboard.

The verification process employs a multi-term consistency metric inspired by recent studies on explanation fidelity and faithfulness [16]. The metric integrates three components: *Overlap* (O), *Direction* (D), and *Coverage* (C), which jointly measure the agreement between decision rules and SHAP attributions. Overlap captures shared features between the decision path and top-K SHAP features, Direction checks whether SHAP contributions follow the same causal trend as rule inequalities, and Coverage quantifies the SHAP importance explained by shared features. These terms form a unified consistency score optimized for SOC-level interpretability.

As illustrated in Fig. 2, the system performs rule extraction, SHAP attribution, consistency computation, and confidence blending. Alerts meeting the verification threshold θ are labeled *verified*, while others are flagged as *unverified* for additional analyst review.

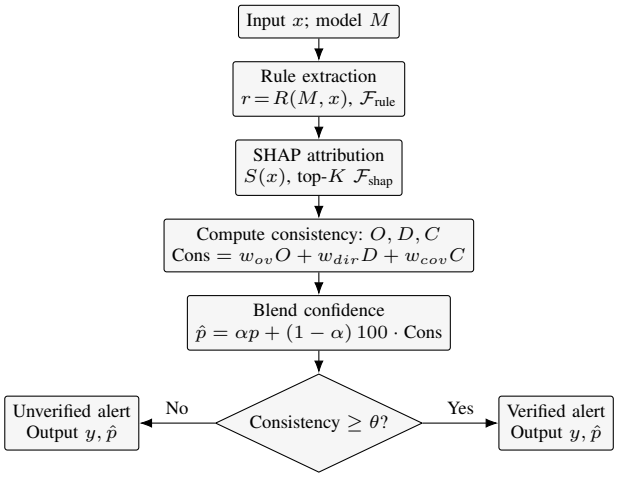

Fig. 2. ExCYDER workflow for explainability-driven alert verification

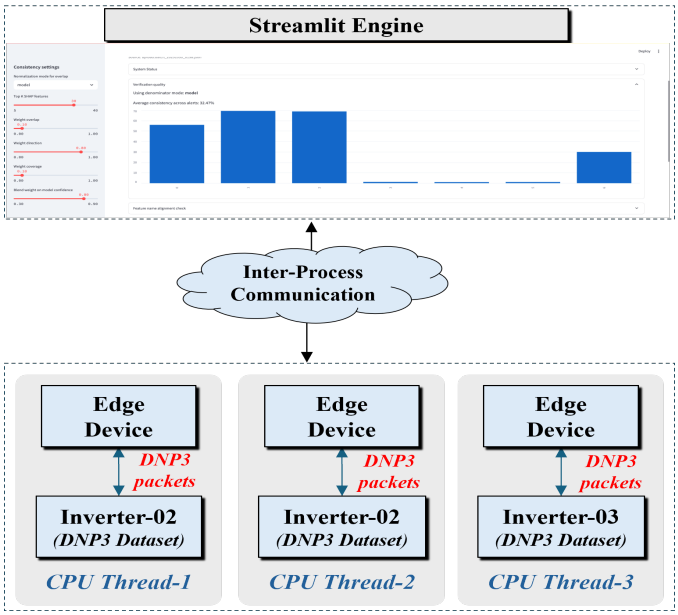

Fig. 3. Experimental setup showing communication between the DER devices, edge devices and the cloud analytics layer

IV. Experimental Evaluation

A. Experimental Setup

Fig. 3 illustrates the experimental setup, which emulates a distributed SOC-oriented cybersecurity environment for DER networks. Three virtual DER devices were implemented in Python 3.13 on an Apple M4 system (16 GB RAM, macOS 15.5), each replaying labeled DNP3 traffic in real time to ExCYDER's edge anomaly detection module. The edge layer performs local inference and aggregates alerts into JSON batches transmitted to the cloud every 60 seconds [17]. The cloud processes each batch on the same interval for verification and visualization, providing near-real-time situational awareness without excessive update frequency. Verified alerts are displayed on the SOC dashboard with their classifications, adjusted confidence values, and consistency scores to support analyst decision-making.

TABLE I
PERFORMANCE EVALUATION METRICS

Metric	Description / Formula
(a) Verification and Explainability Metrics	
Verification Success rate (%)	Ratio of alerts whose consistency exceeds threshold θ; reflects reasoning alignment.
Verification Failure Rate (%)	Fraction of alerts that did not pass the verification threshold at $\theta = 50\%$; represents unverified or low-consistency alerts.
Average consistency (%)	Mean rule–SHAP consistency across all alerts; measures interpretability and model coherence.
Confidence deviation (%)	Average difference between baseline and verified confidence; quantifies verification stability.
Average SHAP time (ms)	Mean computation time for SHAP feature attribution; represents explainability overhead.
(b) System Performance Metrics	
Ingest latency (s)	Delay between alert batch creation and visualization; end-to-end data transfer efficiency.
Rerun interval (s)	Time between dashboard refresh cycles; ensures alignment with the 60 s verification schedule.
CPU utilization (%)	Processor load per refresh; values above 70% indicate heavy computation overhead.
RSS memory (MB)	Resident memory footprint of the Streamlit process; tracks memory usage.
I/O glob (s)	Time spent listing all JSON batch files in the directory; measures disk or network latency.
I/O read (s)	Time to open and read the latest JSON file.
JSON parse (s)	Time to deserialize JSON content into Python objects; reflects data volume and parsing cost.

B. DER Dataset

This work uses the Iowa State University DER DNP3 dataset [18], which provides labeled DNP3 communication records collected from a DER network environment. The dataset defines six traffic classes that capture both normal operation and representative cyberattacks within DER communication: Benign Traffic, Denial of Service, Remote Terminal Attack, Scanning Attack, DNP3 Stealthy Attack, and Replay Attack. These labels form a reliable basis for supervised training and evaluation of the anomaly detection system by distinguishing routine behavior from malicious activity.

C. Results and Discussion

The evaluation focused on validating ExCYDER's anomaly detection performance, explanation fidelity, and verification consistency under a SOC oriented simulation environment. A total of 250 DER communication samples were simulated over a 7 minute interval, generating 121 malicious events that included DNP3 Stealthy, Denial of Service (DoS), and Replay attack scenarios, three prominent inverter layer vulnerabilities documented in [19]. These events were aggregated into seven JSON batches using the 60 second edge batching interval. Each batch was processed by ExCYDER's cloud level verification engine, where rule based decision paths and SHAP feature attributions were cross validated before being visualized on the SOC dashboard for analyst interpretation.

Machine Learning Detection Performance: As established in our earlier work [7], the ML based ADS demonstrates high detection reliability on DER communication traffic. That evaluation assessed LightGBM, XGBoost, and Random Forest using standard classification metrics, including Precision (correct identification of attacks among all predicted attacks), Recall (correct detection of attacks among all actual attacks), F1 score (the harmonic balance of Precision and Recall), and the False Negative Rate (missed attacks). LightGBM achieved the strongest performance with more than 99 percent Precision, Recall, and F1 score, and the lowest False Negative Rate of 0.51 percent, indicating minimal missed attacks and clear separation between benign and malicious traffic. XGBoost and Random Forest followed with higher False Negative

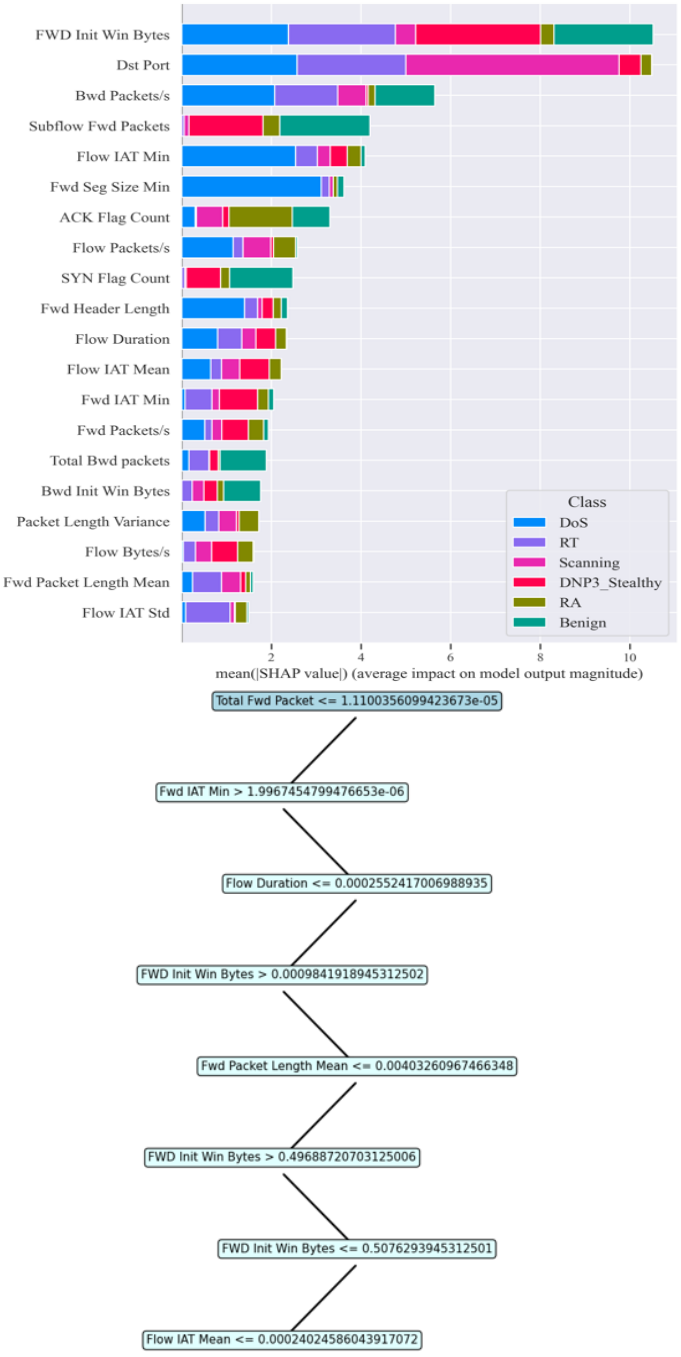

Fig. 4. (Top) Global SHAP summary plot for the LightGBM showing feature importance across alert classes. (Bottom) Decision rule path for DoS illustrating shared classification logic

Rate values of 0.88 percent and 5.10 percent, respectively. Based on these validated results, LightGBM is adopted as the operational anomaly detector in ExCYDER.

Explainability Analysis: Fig. 4 presents the global SHAP summary of feature contributions across all alert classes and decision-path for DOS attack. Attributes such as *FWD Init Win Bytes*, *Dst Port*, and *Bwd Packets/s* were the most influential features shaping model decisions, reflecting their high correlation with protocol-level activity. The decision-path illustrates how specific thresholds in these features lead to the classification of a DoS event. These insights confirm that the model captures meaningful traffic behaviors rather than relying on spurious correlations, thereby improving interpretability.

Verification Outcomes: Consistency scores varied across alerts due to differences in the model's decision paths and feature attributions, but the overall average remained 44.6%, indicating moderate alignment between rule driven reasoning and SHAP attributions. Overall, the verification module achieved a 54.5% success rate, with alerts below the 50% threshold filtered as unverified (45.5%). The confidence deviation stayed within $\pm 5\%$, confirming that verification preserved classifier reliability. The average SHAP computation time was 14.5 ms per alert, demonstrating the feasibility of explainability in near real time SOC contexts.

Operational Stability: During continuous runtime, system metrics confirmed the dashboard's responsiveness and computational efficiency. The average ingest latency remained below 10 s, and the rerun interval synchronized with the 60-second aggregation cycle. CPU utilization remained under 70%, and resident memory (RSS) stayed below 900 MB, confirming the lightweight nature of ExCYDER's verification loop. These metrics validate the framework's suitability for continuous SOC operation without performance degradation.

V. Conclusion

The results demonstrate that ExCYDER effectively integrates high-accuracy ML detection with interpretable and verifiable reasoning for DER-focused SOC environments. By combining rule-based reasoning, SHAP feature attribution, and consistency-driven verification, the framework transforms post-hoc explanations into actionable and audit-ready intelligence. This internal validation mechanism also reduces redundant or low-confidence alerts, supports transparent triage, and enhances analyst trust in automated decision support. Future work include: (1) real-time deployment of ExCYDER within operational DER networks to evaluate performance, latency, and scalability under production conditions and, (2) development of interactive interfaces that integrate operator expertise with AI-driven insights, improving human–AI collaboration for SOC decision-making and threat intelligence.

Acknowledgement

This work is funded in part by the US DOE Cybersecurity, Energy Security, and Emergency Response (CESER) Award Number DE-CR0000049.

References

[1] A. H. Javed, P. H. Nguyen, J. Morren, and J. G. H. Slootweg, "Review of operational challenges and solutions for DER integration with distribution networks," in *Proc. 56th Int. Universities Power Eng. Conf. (UPEC)*, Middlesbrough, UK, Aug. 2021, pp. 1–6.

[2] National Renewable Energy Laboratory, "The distributed energy resource cybersecurity framework," National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-82577, 2022. [Online]. Available: https://www.nrel.gov/docs/fy22osti/82577.pdf

[3] U.S. Department of Energy, "Cybersecurity considerations for distributed energy resources on the U.S. electric grid," U.S. Department of Energy, Washington, DC, USA, Tech. Rep., Oct. 2022.

[4] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: An overview from machine learning perspective," *J. Big Data*, vol. 7, no. 1, pp. 1–29, Dec. 2020.

[5] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[6] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed, "Explainable artificial intelligence approaches: A survey," *IEEE Access*, vol. 9, pp. 101 974–102 001, 2021.

[7] D. Popoola, S. Bhattacharya, and M. Govindarasu, "Cysider: Cybersecurity situational intelligence framework for DER networks," in *2025 Resilience Week (RWS)*, National Harbor, MD, USA, 2025, (Accepted).

[8] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Dec. 2019.

[9] W. Villegas-Ch, A. Jaramillo-Alcázar, A. M. Navarro, and A. Mera-Navarrete, "Integrating explainable artificial intelligence in anomaly detection for threat management in e-commerce platforms," *IEEE Access*, vol. 13, pp. 29 830–29 846, 2025.

[10] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *IEEE Access*, vol. 10, pp. 93 310–93 333, 2022.

[11] G. Rjoub, J. Bentahar, O. A. Wahab, R. Mizouni, and A. Mourad, "A survey on explainable artificial intelligence for cybersecurity," *IEEE Transactions on Network and Service Management*, vol. 20, pp. 3241–3260, 2023.

[12] S. Paul *et al.*, "Demystifying cyberattacks: Potential for securing energy systems with explainable AI," in *Proc. 2024 Workshop Comput., Netw. Commun. (CNC)*, Las Vegas, NV, USA, 2024, pp. 430–435.

[13] I. G. Al-Ashmali *et al.*, "Enhancing network security: Robust anomaly detection with ensemble learning and explainable AI," *Comput. Secur.*, vol. 125, p. 103118, Feb. 2023.

[14] D. Wang, T. C. Chen, and K. Mahapatra, "Interpretability, explainability and trust metrics in anomaly detection method for power grid sensors," in *2025 IEEE PES Grid Edge Technologies Conference and Exposition (Grid Edge)*, San Diego, CA, USA, April 2025.

[15] S. Bhattacharya, N. Saqib, and M. Govindarasu, "Ml-based anomaly detection system for IEC 61850 communication in substations," in *2024 IEEE Power & Energy Society General Meeting (PESGM)*, Seattle, WA, USA, 2024.

[16] U. Bhatt, A. Weller, and J. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4550–4561.

[17] S. Bhattacharya and M. Govindarasu, "Real-time cybersecurity situational awareness framework for agriculture machinery-based iot networks," in *2025 34th International Conference on Computer Communications and Networks (ICCCN)*, 2025, pp. 1–9.

[18] M. AbdElKhalek, "DER_ML_ADS: Machine Learning-Based Anomaly Detection for DER Networks," 2022, accessed: 2025-07-09. [Online]. Available: https://github.com/Moataz-AbdElKhalek/DER_ML_ADS/tree/main/dataset/DNP3

[19] F. R. Labs, "Sundown: Exploiting design flaws in industrial and solar power communication protocols," Forescout Technologies, Inc., San Jose, CA, USA, Tech. Rep., March 2025, technical Vulnerability Research Report. [Online]. Available: https://www.forescout.com/resources/sun-down-research-report/