International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Explainable and Adversarially Robust AI for Cyber Defense in Critical Infrastructure Systems

[1]Malik Huzaifa, [2]Hamza Afzal

Submitted:08/04/2025 **Revised**: 16/05/2025 **Accepted**: 29/06/2025

Abstract: Critical infrastructure systems (CIS) including water delivery facilities and electrical connections, transport, healthcare, and communication networks are being increasingly targeted by highly advanced forms of cyber-attack because they are digital and interconnected. Ransomware, phishing, DDoS, and APTs are among the advanced attacks that traditional cybersecurity mechanisms are inadequate. Artificial Intelligence (AI) has become a game-changer for cyber defense, offering cutting-edge capabilities in intrusion detection, anomaly detection, and predictive analysis. In the realm of cyber defense, AI is a revolutionary tool that brings advanced features such as intrusion detection, anomaly detection, and predictive analytics. But in most AI models, the black-box nature is not conducive to transparency and trust, especially in critical decision-making scenarios. Explainable Artificial Intelligence (XAI) offers solutions to this, as it makes the insights obtained by AI models interpretable through tools like SHAP, LIME, and rule-based models. Meanwhile, adversarial machine learning reveals the weaknesses in AI systems that could impact performance or be overlooked when attackers input data. In this paper, the synergy of explainable AI and adversarially robust AI in the context of securing critical infrastructure systems. It covers important techniques, defense mechanisms and problems, and shows the compromises in robustness and interpretation in high-risk Cybersecurity environments.

Keywords: Critical Infrastructure Security, Explainable Artificial Intelligence, Adversarial Machine Learning, Cyber Defense, Machine Learning, Deep Learning, Robust AI Models

I. INTRODUCTION

A critical infrastructure system (CIS) is essential to modern society and includes generators and power distribution systems, water treatment and supply systems, transportation systems, health services, communications systems and other critical systems [1]. Operational efficiency, scalability and automation have been markedly improved with the emergence development of virtual physical technologies, the Industrial Internet of Things (IIoT) and management platforms based on the cloud that increasingly digitize and connect these systems [2][3]. However, this change has led to an enhanced cyberattack surface and elevated the vulnerability of these infrastructures to sophisticated attacks [4]. Advanced Persistent Threats (APTs), ransomware, phishing, DDoS attacks, and zero-day flaws may all seriously harm systems, result in financial losses, interfere with fundamental services, and even jeopardize public security [5][6]. The current methods of cybersecurity, which are rule-based, signature-based and rely on manual monitoring, are no longer sufficient to defend and counteract today's intelligent and dynamic attack strategies.

In response to these constraints, the application of AI has become a game-changer in critical infrastructure cyber defense [7]. By examining web traffic, system records, and user behaviour, machine learning (ML) and deep learning (DL) techniques may be used to conduct sophisticated large-scale data security evaluation. Unlike traditional systems, these AI-based solutions provide real-time IDS/AIDS, malware classification and predictive threat intelligence, which enables more agile and responsive reactions. Although the models are quite powerful, the process of making decisions of the majority of the sophisticated models is difficult for an ordinary analyst to comprehend and is sometimes referred to as a "black-box." Gaining trust, taking risks into consideration, and adhering to rules are challenging, particularly in high-risk scenarios where mistakes might have detrimental effects on activities [8].

One way to tackle this major constraint is through XAI, which provides easily understandable insights into AI decisions. Models that use techniques like SHAP, LIME, decision trees, and rule-based systems can give security analysts an explanation for how a model is making a decision because the event is malicious or benign, which helps them to build trust and validate models and make better decisions more efficiently [9]. Meanwhile, adversarial machine learning has demonstrated some serious weaknesses in AI systems, where attackers can craft deceptive inputs or training data to pretend AI systems into missing the mark or failing to detect them. The adversarial attacks have emphasized the need to create a resilient AI that can function efficiently under a challenging environment. Even though some advancements have been made in both the areas of XAI and adversarial robustness, their implementation in

[1]*System Administrator FNR Solutions INC Baltimore Maryland* hmalik.student@wust.edu
[2]*Software engineer American Technology group LLC Baltimore Maryland* hafzal.student@wust.edu

cybersecurity is still limited. In this paper, the authors examine how explainable AI and adversarially robust AI complement one another and discuss various techniques, challenges, and future research avenues for building robust, explainable, and secure AI systems for cyber defense.

A. Structure of the Paper

The remainder of this essay is organized as follows: Major cybersecurity risks and AI's role in cyber defense for essential infrastructure systems are covered in Section II. The authors examine explainable AI methods and their potential applications in cybersecurity in Section III. Section IV discusses rivalry-based robust AI attacks and defense tactics. Section V includes a review of the literature and a comparative analysis of relevant studies. Lastly, Section VI provides an overview of the study and makes recommendations for subsequent research avenues to improve robust, transparent, and safe AI systems.

II. Artificial Intelligence for Cyber Defense in Critical Infrastructure Systems

Critical Infrastructures (CIs) are fundamental systems required to support key infrastructure sectors, including power, transport, water, telecommunications, healthcare, etc. In the era of increasing reliance on computers and other technologies, Cybersecurity of Critical Infrastructures is now essential to ensure national security and public safety. The security measures in CI have evolved rapidly over the years as computing and communications technologies have advanced.

A. Cybersecurity Threats in Critical Infrastructure

As vital infrastructure networks including energy, healthcare, finance, transportation, and water control become more digitalized and linked together, they are more vulnerable to cyberattacks [10]. The following are the identified cybersecurity threats as follows:

- **Malware and Ransomware Attacks:** When a perpetrator infects a device, vaults data, and then demands payment to decrypt it, this is known as a malicious program. These attacks can disrupt services and have a massive financial and operational impact.
- **Distributed Denial of Service (DDoS) Attacks:** DDoS attacks are attacks that flood the networks or servers with too much traffic, resulting in a denial of critical services, communication or operation.
- **Phishing and Social Engineering:** Fraudulent emails, messages or websites are used to deceive employees into disclosing their login information or into clicking links in them that lead to malicious software to be installed, allowing unauthorized access to the system.
- **ICS and SCADA Vulnerabilities:** ICS and SCADA are frequently vulnerable to cyber attacks because use outdated security mechanisms, weak authentication and software vulnerabilities.
- **Advanced Persistent Threats (APTs):** APTs are very complex and long-term cyberattacks that take place over an extended period to gain access to information or disrupt infrastructure in stealth mode.
- **Insider Threats:** Insiders (employees or contractors) can inadvertently or intentionally make it easier for others to get to security systems, either by neglecting security measures or by failing to practice proper security.
- **IoT Security Risks:** Internet of Things (IoT) in smart infrastructure is yet another area that is not considered as secure. It may serve as a gateway for an illegal entrance or cyberattack.
- **Supply Chain Attacks:** Attacking and compromising the third-party vendors or software providers.
- **Cloud and Data Breach Threats:** Data relating to sensitive infrastructure could be accessed by unauthorized individuals as well as subjected to cyber attacks when configuration in the cloud is not appropriate.

These dangers highlight how crucial it is to improve vital infrastructure safety for security and risk control.

B. Role of Artificial Intelligence in Cyber Defense

AI is an important element of the current cybersecurity system. Through its use, security becomes intelligent, adaptable, and automatic in order to counter the evolving cybersecurity threat. Signature-based detection and human monitoring are two methods used in conventional cyber security systems. However, these methods cannot detect some of the advanced attacks such as ransomware, phishing, zero-day attack, and APT. But the use of AI solve all these problems, because it can analyze vast quantities of data related to security and detect attack trends and take necessary steps to counter them. It can be employed to detect any threat or anomaly that may be part of an offensive strategy. ML and DL algorithms are deployed to analyze network traffic, system log files and user behavior for anomalies that indicate attacks.

The ability of AI to recognize unfamiliar threats and zero-day attacks in contrast to traditional systems is possible because AI is capable of detecting deviations in behavior as indicators of an attack [11]. AI technology could also be applied in cyber threat intelligence and prediction analysis, where the data for cybersecurity would be gathered from different sources, including firewalls, intrusion detection, security, and endpoints. The information gathered in cyber

threat intelligence could be used to determine the possible weaknesses of an organization and its probable attack vectors, thus providing proactive protection against attacks instead of reactive. Another equally important aspect that needs consideration is Malware Detection and Intrusion Prevention application. Software execution behavior and networking could be analyzed using AI systems to detect any malicious actions like ransomware and fileless malware [12].

Further, IDS and IPS, based on Artificial Intelligence, can respond to any kind of incidents in automatic mode through isolating the system, blacklisting the malicious IP address and notifying the users in real-time. Artificial Intelligence is being applied in the area of User and Entity Behavior Analytics that helps to monitor the activities of the users and detect any suspicious behavior. Likewise, Natural Language Processing (NLP) can help identify phishing and fraud by analyzing email content, URLs, and communication patterns. AI has several weaknesses even in the domain of cybersecurity; some of which include poor data quality, increased number of false alarms, and adversarial attacks on AI systems that confuse machine learning algorithms. Ethics with regard to privacy and explainability are also important.

C. *Machine Learning and Deep Learning for Threat Detection*

The concepts of ML and DL play key roles within the framework of modern cyber threat detection systems due to the multidimensional nature of security data available and the high amount of information that is generated. This can be done based on past and present data. The adaptive, knowledge-driven safety measures provided by ML and DL models may detect both known and zero-day threats, while signature-based models are static and heuristic-based, only identifying known assaults.

1) Supervised Learning for Threat Classification

When tagged data becomes available, supervised learning techniques have been widely used in cyber risk monitoring [13]. Methods that classify web traffic or system records as either safe or malicious include SVM, Random Forest, and gradient optimization. Differential patterns may be learned from past attack data and applied to classifying malware and detection of breaches.

2) Unsupervised Learning for Anomaly Detection

Unsupervised learning techniques are used when labeled attack data is scarce or absent. Clustering and other methods like density-based models detect anomalies from normal. In the field of cybersecurity, these methods can be especially effective when identifying an unknown or zero-day attack, since based on a model of normal network activity and any deviation from it is suspect.

3) Deep Learning for Complex Pattern Recognition

CNNs and RNNs is common in extracting complex spatial and temporal patterns from cybersecurity data [14]. CNNs are proven to be capable in analyzing malware images and extract features from network traffic, whereas RNNs and LSTM networks are appropriate for analyzing sequential data like system logs and time-series network traffic.

4) Ensemble Learning Methods

Numerous ML algorithms are implemented to increase detection accuracy and robustness. Bagging, boosting, and stacking are methods for improving performance, based on reducing both variance and bias. Ensemble models are frequently used in cyber threat detection to decrease the errors made by the system in complex and noisy data sets.

5) Real-Time Threat Detection Systems

Over time, the use of ML and DL in real-time cybersecurity systems has grown as a means of continuously monitoring the system to discover and detect intrusions. Methods like online learning and streams data analytics might be useful in identifying various types of hackers.

In summary, the use of ML and DL methods is highly beneficial for enhancing cyber threat detection, providing efficient solutions that are automated and scalable. Nevertheless, there are several issues that remain to be resolved, including data imbalance, adversarial attacks, and interpretability limitations.

III. Explainable AI for Cybersecurity and Threat Detection

The use of ML and DL algorithms is increasing in cyber threat management, and have been found to be useful in detecting malicious activities, intrusion on networks, and malware. There are, however, many models that are advanced but are difficult to explain decision-making process, called black-box models [15]. XAI techniques have gained popularity in recent years as a means of enhancing safety systems' oversight, openness, and confidence. These methods help analysts understand why a danger is identified and make more informed conclusions.

- **Local Interpretable Model-Agnostic Explanations (LIME):** LIME provides local explanations for single predictions made by complex machine learning models by fitting simpler models in the vicinity of the prediction [16]. It assists with incident investigation in cyber threat detection when evaluating why a given network event or suspicious activity is deemed malicious.
- **SHapley Additive exPlanations (SHAP):** SHAP is a method based on principles of game theory, which quantifies feature contributions to prediction results. It offers local and global interpretability, allowing analysts to grasp significant behaviors like unusual

system activity, suspicious traffic patterns, or unauthorized access attempts.

- **Decision Tree-Based Explanations:** Decision Trees are representation of classification rule that ease the process of determining classification in a transparent manner. In the field of cybersecurity, then assist cybersecurity analysts in understanding the reasoning behind the detection of a cyber attack, which facilitates the understanding and validation of predictions.
- **Rule-Based Explainability Systems:** These techniques are based on the IF–THEN rules to make meaningful links understandable in cyber threat detection. If these events occur in a pattern over time, or if the traffic pattern detect differs from expected, these events could be useful in determining whether a brute-force attack is currently taking place, providing additional context for these alerts.

In conclusion, clarification strategies help with incident resolution, reduce error rates, and improve honesty and confidence of AI-based cyber threat detection technologies.

A. XAI Methods in Intrusion Detection and Critical Infrastructure Security

Explainable Artificial Intelligence (XAI) is a crucial component for improving intrusion detection systems (IDS) and safeguarding critical facilities environments where openness about decision-making is just as vital as detection efficacy [17][18]. Critical infrastructures like industrial control systems, smart grids, and transportation networks, require not only high detection performance, but also justifications of the automated responses taken to ensure that the operator is able to trust the application and meet regulatory requirements.

1) Model Transparency for Security Assurance

XAI techniques are applied to reveal model inner workings in intrusion detection systems where predictive models are deployed in real-time monitoring systems. It is especially critical in critical infrastructures where quick verification of automated alerts is essential to prevent disruption in operations. Transparent models aid security analysts in the identification of an attack versus a benign anomaly due to system fluctuations.

2) Post-Hoc Explanation Frameworks

The most commonly used post-hoc XAI methods are those that are applied to interpret the already-trained intrusion detection models without changing its architecture. These frameworks offer explanations following the prediction generation, which is applicable to deep learning-based IDS in industrial environments that are either costly or impractical for retraining.

3) Behavioral Pattern Interpretation

The use of XAI techniques helps to interpret behavioral deviations in network and system activity. In intrusion detection, the abnormal sequence of user actions, device communications or process executions are converted to interpretable patterns that give the analysts a better understanding of the changing attack strategies like lateral movement or privilege escalation.

4) Temporal Explainability in Sequential Attacks

Many cyber-attacks against critical infrastructure are multi-stage attacks that are delivered over a period of time. Temporal explainability methods are used to understand how patterns in logs and network flows are treated in intrusion detection models. This, in turn, helps understand the attack progression in IPCS (Industrial and Cyber-Physical Systems).

5) Explainability in Industrial Control Systems (ICS)

XAI methods in ICS focus on explaining the anomalies in the sensor data, the behavior of actuators, and the process control signals. With these explanations, it is possible to identify cyber-physical attacks like false data injection or control manipulation, while keeping the system stable [19].

6) Risk-Based Explanation Models

Some XAI approaches in critical infrastructure security do not just make predictions, but give interpretations of the level of risk. The idea is to convert model results into an operational risk rating that could then be used for making decisions regarding response prioritization in view of system criticality and impact.

Concluding, XAI methods for intrusion detection and critical infrastructure security are not only about interpreting AI model outputs but are about creating transparency, providing real-time decision-making assistance, and validating AI-based security solutions from a human perspective. All equip users with competencies to be robust, responsible, and secure in AI operations.

B. Limitations and Challenges of Explainable AI

XAI is yet to be fully adopted in cybersecurity and critical systems due to various limitations. These obstacles are due to the complexity of the models, the need for more computing power, vulnerability to security attacks, and the absence of a standard model.

1) Accuracy–Explainability Trade-off

Simply making a model more interpretable typically comes at the cost of its accuracy, and models that are very accurate deep learning models are not easily interpretable. This trade-off restricts the optimal use of this as an intrusion detection system.

2) High Computational Cost

The high processing costs of several XAI techniques (such SHAP and LIME) make them unsuitable for IDS systems in real-time settings and big networks.

3) Lack of Standard Evaluation Metrics

There is no agreed standard for quality of explanation. There are a number of metrics, such as fidelity or interpretability, which are highly subjective and do not allow for an easy comparison between XAI methods.

4) Adversarial Exploitation Risk

Adversaries are able to reverse-engineer explanations in order to figure out how the algorithm functions and provide it with data for it to overcome security measures, making the system vulnerable.

In summary, these limitations underscore the significance of more research into creation of an effective and secure explanation system in field of cybersecurity generally, and in real-time scenarios specifically.

IV. Adversarially Robust AI Against Cyber Threats

AML is a branch of ML that is concerned with the weaknesses of AI systems when manipulated by an adversary to trick the ML models. Adversarial assaults against AI systems, such as security detection, malware examination, and detecting phishing, are made in the field of defence by altering input or training data provided to the system. The main types of attacks are evasion attacks, which involve providing manipulated inputs that elude detection when deployed and poisoning attacks, which involve providing corrupted data during the training phase to diminish the reliability of the model. Furthermore, model extraction and inference attacks try to generate models or to disclose private information in the training data. Relying on the attacker's level of expertise, the assaults can be executed in white-box, black-box, or gray-box environments. Knowing which attacks to expect is a critical element to understanding how to build a strong AI system that can withstand a changing cyber threat landscape and remain effective for cybersecurity.

A. Vulnerabilities of AI-Based Security Systems

AI-based security systems provide advanced features in threat management and predictions, but also vulnerable to various challenges such as adversarial attacks, data poisoning, exploitation of models, and changing threats.

- **Adversarial Attacks:** Adversarial inputs may impact AI systems through output manipulation or through evading detection mechanisms in cases involving malware and intrusion detections.
- **Data Poisoning Attacks:** These attacks involve providing training sets with corrupted or mislabeled data so that models learn misleading patterns, decreasing the detection accuracy.
- **Model Evasion Risks:** After being installed, these cybercriminals can tamper with files, links, and actions on the network that AI detection software finds difficult to detect.
- **Lack of Explainability:** Many AI systems are "black box" systems that are not easy to understand and analyze, and there are no security issues in them.
- **Bias and Data Dependency:** An AI system for security is only as effective as training data used, biased or inaccurate data can cause false positives or negatives.

In fact, adversaries may repeatedly request deployed systems to create their own AI models and learn the key principles of detection.

B. Adversarial Defense and Robustness Techniques

The importance of improving the resilience and robustness of AI in light of cyberattacks and manipulation cannot be overstated. This is due to the possibility of adversarial attack and data poisoning which can make an AI security model vulnerable, making it essential to have robust security features to ensure detection accuracy [20]. This is based on increasing model robustness, detecting malicious attacks, and responding to a changing threat landscape. In cybersecurity use cases, highly capable AI models are used for effective malware detection and prevention, intrusion protection, and anomaly detection. Also reduce negative implications on system performance.

- **Adversarial Training:** Models are trained using both normal and adversarial examples, thus resulting in models that are more resilient to adversarial examples while also being more capable at handling the distorted data.
- **Input Data Preprocessing:** Some of the techniques that could be used to avoid the occurrence of adversarial perturbations to the input data before feeding it into an AI system include normalization, feature squeezing, and noise filtering.
- **Ensemble Learning:** Combining several machine learning models is done to achieve better results and minimize susceptibility to attack.
- **Explainable AI (XAI):** Explanation methods improve explain ability, helping security analysts understand behavior of model and detect any potential threats.

In general, these robustness methods greatly improve the capabilities and dependability of AI-powered cybersecurity solutions against adversarial attacks.

C. Robust AI Models and Case Studies in Critical Infrastructure

Robust AI models are developed in such a way that resist cyberattacks, manipulation by adversaries, noisy input data, and unfamiliar working conditions without compromising their reliability and accuracy. Robust AI models are essential in fields such as cybersecurity and critical infrastructure because it ensure smooth and continuous operation and identify any anomalies and cyberattacks. Some common Robust AI models are:

1) Convolutional Neural Networks (CNNs)

Because CNNs can find and recognize complicated patterns in big, complex datasets—such as antivirus detection, network traffic analysis, and intrusion classification—they are widely employed in the cybersecurity industry.

2) Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

Models based on LSTM can identify threats, anomalies, and predict the future evolution of cyberattacks in network traffic that follows a time series.

3) Autoencoders (AE)

Additionally, detection systems for anomalies frequently employ autoencoder models to identify suspicious activity and unusual behavior in vital infrastructure networks.

4) Random Forest (RF)

In the case of phishing, malware, and intrusion detection, Random Forest models enhance cybersecurity strength by reducing overfitting and increasing classification accuracy.

5) Extreme Gradient Boosting (XGBoost)

A model of severe gradient boosting XGBoost is a very useful tool for cyber security because of its capacity to handle enormous volumes of data and spot patterns in attack routines, particularly when dealing with issues like risk prediction or threat detection.

6) Generative Adversarial Networks (GANs)

GAN-based methods were used for simulating adversarial attacks and derived synthetic cyber threat data for the purpose of enhancing the robustness of the model with adversarial training.

D. Integration of XAI and Adversarial Robustness AI

In such high-stake situations as critical infrastructure, it is essential to incorporate both adversarial robustness and XAI to develop trustworthy and safe AI systems. While goal of XAI is to uncover reasoning behind an AI model, adversarial robustness focuses on strengthening the models against corrupt input data that can deceive ML algorithms [21]. Together, the above-mentioned methods help create secure and comprehendible cybersecurity systems, mostly for humans in the context of their operations, in critical environments, such as power, water, and communication infrastructures. It is vital to integrate these two approaches to ensure that cybersecurity operations are performed in real-life environments as this would allow not only to identify possible attacks but also understand the logic of the AI alert (Table II).

TABLE I. KEY INTEGRATION APPROACHES

Approach	Description	Benefit
Adversarial Training with Interpretable Models	Uses robust training methods with inherently interpretable models, such as DT or rule-based systems	Improves both robustness and transparency
Post-hoc Explanation Methods	Applies techniques like SHAP or LIME to explain outputs of robust deep learning models	Enhances interpretability of complex models
Attention-based Hybrid Models	Combines deep learning with attention mechanisms for feature importance visualization	Balances accuracy and explainability
Ensemble-based Hybrid Systems	Integrates multiple models for robustness and explanation consistency	Improves reliability in threat detection

Despite all of this, there still are some challenges to be faced. This is where the trade-off between sturdiness and understanding occurs: very complicated deep learning models are less interpretable but more resilient to adversarial assault. However, simpler models are easier to understand, but they could not be as safe in the event of a more complex attack. The other challenge is that explanations are fragile in the face of conditions that are hostile: If inputs are perturbed slightly, predictions and explanations change significantly. In addition, the expenses of XAI and adversarial robustness involve some computation overhead, which is not desirable for critical infrastructure real-time systems. Moreover, there is no standardized evaluation metrics to evaluate explainability, robustness, and accuracy comprehensively, which is not easy to compare the outcomes of different studies.

A digital twin is considered as a live view of a real-world system that monitors the state

of its entities. Deeply, it is an environment that consists of a virtual and a physical machine.

Each machine (model) is represented as a simulation, a mirror, or a twin of the other.

So, the digital twin can list the life cycle of the physical entity which can be a human, an

object, or a process [68]. Each digital twin is connected to its counterpart by a unique key,

therefore a relationship between two entities can be established [48].

A digital twin is a partition of a Cyber-Physical System (CPS), which is a set of physical

systems connected to virtual cyberspace through the network [11, 49]. The communication

between a physical entity and its digital twin can be represented directly by physical con-

nections or indirectly via a cloud system. Also, it can be a seamless connection and con-

tinuous data exchange [26, 88]

V. LITERATURE OF REVIEW

The critical infrastructure field, as covered in the reviewed literature, has not many studies that have combined interpretability, robustness, and real-time deployment, which are all key concepts of explainable AI in cybersecurity.

Mohale and Obagbuwa (2025) carries out a thorough analysis of application of XAI in IDS to improve knowledge and transparency in the cybersecurity industry. This review's objective is to identify the most widely used XAI approaches, evaluate their effectiveness in IDS, and weigh their benefits and drawbacks using a thorough analysis of recent research. The most often used XAI models, according to the results, are rule-based and tree-based; nevertheless, maintaining interpretability while maintaining high detection accuracy is not always simple. Moreover, the review reveals the lack of standardization and scalability, and calls for hybrid approaches and real-time explainability [22]

Petihakis et al. (2024) explain the development of AIAS and examine the emerging risks of hostile AI. To strengthen the security of AI operations against these assaults, AIAS is a comprehensive security solution based on artificial intelligence. AIAS aims to transform cybersecurity and make AI more resistant to adversarial assaults by analyzing data and creating new techniques. It also intends to make the environment safer for the use of AI technology in vital applications. In order to encourage the use of AI security solutions, paper outlines components of the AIAS platform, discuss about how it works and suggests future lines of inquiry [23]

Nkoro et al. (2024) uses the 2023 Edge-IIoT set and 2023 CICIoT, two modern marine cybersecurity datasets, to demonstrate a zero-trust NIDS architecture for identifying modern maritime cyberattacks. In a multi-class experiment, the zero-trust NIDS model has an optimal MCC score of 97.33%. The XAI approach uses both quantitative and visual XAI tools, including the LIME algorithms and SHAP, to increase explain ability and interpretability. The study's conclusions demonstrate that by improving the reliability and comprehensibility of the black-box NIDS models used for maritime digital defense, it is feasible to enhance the overall cybersecurity posture of marine organizations [24]

Nguyen et al. (2023) present Mont image AI Platform (MAIP), a cutting-edge GUI-based DL platform that can identify and categorize illegal activity and provide an explanation for model's decision. The resulting DL model's predictions are interpreted using well-known XAI techniques. Additionally, use adversarial assaults to evaluate the model's robustness and accountability using several measurable indicators. They all conduct in-depth tests on both private and public network traffic. The model provides good performance and resilience, as shown by the experimental findings, and its results closely match the domain knowledge [25]

Almuqren et al. (2023) developed the XAI, XIDS Method for Secure Cyber-Physical Systems, or XAIID-SCPS. The identification and classification of intrusions into CPS platform is main objective of proposed XAIID-SCPS technique. The XAIID-SCPS method selects features using a HEGSO algorithm. The parameters of the Improved IENN model for intrusion detection were optimized using the EFFO technique. Furthermore, the XAIID-SCPS technique uses the XAI methodology LIME to enhance the black-box method's comprehensibility and explainability for accurate intrusion classification [26].

Neupane et al. (2022) explores the current status of XAI for IDS, its challenges, and how these challenges connect to development of an X-IDS. In particular, thoroughly discuss black and white box methods and demonstrate how they vary in terms of performance and ability to provide explanations. Furthermore, offer a generic design that considers human-in-the-loop and may be used as a template for developing an X-IDS. Three crucial perspectives are presented in the research recommendations: the necessity of defining explainability for IDS, the necessity of developing explanations customized for different stakeholders, and the necessity of developing metrics to assess explanations [27].

Table II summarizes prior studies on explainable and adversarial robust AI for cyber defense, highlighting their research focus, key findings, limitations, and future directions to identify existing research gaps.

TABLE II. COMPARATIVE ANALYSIS OF RECENT STUDIES FOR EXPLAINABLE AND ADVERSARIALLY ROBUST AI FOR CYBER DEFENSE

Authors	Focus	Findings	Limitations	Future Work
Mohale and Obagbuwa (2025)	Systematic review of XAI integration in IDS for transparency and interpretability in cybersecurity	Rule-based and tree-based XAI techniques provide better interpretability; hybrid models and real-time explainability are important for IDS	Limited standardization of XAI methods in IDS; scalability challenges; trade-off between explainability and detection accuracy	Develop standardized XAI frameworks, scalable real-time explainable IDS, and hybrid models balancing accuracy and interpretability
Petihakis et al. (2024)	Development of AIAS for securing AI systems against adversarial attacks	AIAS enhances resilience of AI systems against adversarial AI threats through a comprehensive security platform	Limited focus on explainability of decisions; implementation in diverse real-world critical infrastructure environments remains insufficient	Improve AI robustness mechanisms, integrate explainability with adversarial defense, and validate in critical infrastructure domains
Nkoro et al. (2024)	Zero-trust NIDS with XAI for marine cyberattack detection using Edge-IIoTset and CICIoT datasets	SHAP and LIME improve interpretability of NIDS; achieved high MCC score (97.33%) in marine cybersecurity	Framework mainly tested in marine cyber environments; limited evaluation against adversarial manipulation of IDS models	Extend framework to broader critical infrastructure systems and integrate adversarial robustness with explainable IDS
Nguyen et al. (2023)	Montimage AI Platform (MAIP) for malicious traffic detection using DL, XAI, and adversarial testing	High performance in malicious traffic detection; XAI methods improve interpretability; robustness validated through adversarial attacks	Limited applicability to large-scale, real-time critical infrastructure systems; explainability-performance trade-off not fully explored	Improve real-time deployment, optimize scalable explainability, and strengthen resilience against sophisticated adversarial attacks
Almuqren et al. (2023)	Explainable AI-enabled intrusion detection for secure Cyber-Physical Systems (XAIID-SCPS)	LIME improves understanding of black-box IDS; optimized feature selection and neural network improve intrusion detection performance	Focuses primarily on explainability without explicit adversarial robustness evaluation; tested on limited CPS contexts	Incorporate adversarial defense strategies into XAI-based CPS intrusion detection and evaluate across diverse infrastructure systems
Neupane et al. (2022)	Survey of XAI approaches for IDS and explainable IDS (X-IDS) design	Identifies trade-offs between black-box and white-box approaches; proposes human-in-the-loop architecture	Lack of standardized explainability metrics; absence of stakeholder-specific explanation models; limited practical deployment guidelines	Define explainability standards for IDS, create tailored explanations for stakeholders, and establish evaluation metrics for explainability

VI. CONCLUSION AND FUTURE WORK

The current critical infrastructure systems need to keep pace with the rapid digital transformation, cyber-physical systems, and complex cyber threats to make cybersecurity more complex. AI has significantly improved defense systems' capacity to recognize dangers, spot irregularities, and forecast security threats. Two major concerns with combining AI technology in high-stakes security situations

are the openness of AI decision-making processes and susceptibility of AI-driven systems to adversarial attack. XAI approaches have been created to improve trust, accountability, and commercial decision-making by giving people comprehensible interpretations of models' behaviors in order to solve the understanding barrier. Meanwhile, adversarial machine learning identifies key vulnerabilities that attackers could exploit through evasion, poisoning or extraction attacks against AI models. Combination of explainability and robustness to adversaries is another serious problem. The progress of developing safe and reliable AI systems for the security of ICTs for critical infrastructures. For future development, it would be useful to develop common approaches to improve interpretability, robustness and scalability for real-time applications. Standardized metrics for explainability and adversarial resistance are also required in a consistent manner. Furthermore, lightweight and adaptive AI models are needed for resource-constrained industrial settings, providing ongoing protection against emerging cyber threats while ensuring transparency and reliability.

REFERENCES

[1] E. Areghan and O. S. Ndibe, "Explainable AI for Autonomous Threat Detection in Critical Infrastructure Systems," *J. Comput. Anal. Appl.*, vol. 33, no. 8, pp. 6841–6857, 2024.

[2] S. Chatterjee, "A Data Governance Framework for Big Data Pipelines: Integrating Privacy, Security, and Quality in Multitenant Cloud Environments," *Tech. Int. J. Eng. Res.*, vol. 10, no. 5, 2023, doi: 10.56975/tijer.v10i5.158181.

[3] S. Malaraju, "Securing Cloud Environments with Bastion Hosts," *Int. J. Multidiscip. Res.*, vol. 7, no. 2, Apr. 2025, doi: 10.36948/ijfmr.2025.v07i02.40257.

[4] R. K. Gadiraju, "Cloud-Native AI Platforms for Scalable Enterprise Machine Learning: Architecture, Challenges, and Best Practices," *Int. J. Intell. Syst. Appl. Eng.*, vol. 9, no. 4, pp. 481–492, Oct. 2021, doi: 10.17762/ijisae.v9i4.8119.

[5] G. C. Kakaraparthi, "Integrating Serverless Architectures and Kubernetes for Scalable and High-Availability AI Workflows," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4, pp. 5896–5905, 2024, doi: 10.6084/m9.figshare.30445046.

[6] S. Chandrappa and S. Paheding, "Exploring the Depth of the KAN Method for Hyperspectral Image Classification," in *2025 Northeast Section Conference Proceedings*, ASEE Conferences, 2025. doi: 10.18260/1-2--55021.

[7] R. Dandigam, "A Multi-Agent Reinforcement Learning System for Autonomous Optimization of Web Infrastructure and Services," *Int. J. AI, BigData, Comput. Manag. Stud.*, vol. 4, no. 3, pp. 146–154, Sep. 2023, doi: 10.63282/3050-9416.IJAIBDCMS-V4I3P115.

[8] M. Kari, "Deep Learning-Based Fault Prediction Models for Enhanced Network Security Monitoring," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 3, no. 3, p. 492, Jun. 2023, doi: 10.48175/IJARSCT-11600I.

[9] N. Kolli, J. W. Sajja, and A. Nerella, "Building Secure AI Agents for Autonomous Data Access in Compliance/Regulatory-Critical Environments," *Comput. Fraud Secur.*, vol. 2024, no. 9, pp. 363–373, 2024, doi: 10.2139/ssrn.5528763.

[10] A. R. Cavalli and E. M. De Oca, "Cybersecurity, monitoring, explainability and resilience," in *2023 Fourteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*, 2023, pp. 1–7.

[11] M. R. Tandi, "AI-Powered Cyber Defense Framework for Advanced Computing Environments and Critical Infrastructure," *Electron. Commun. Comput. Summit*, vol. 1, no. 1, pp. 76–85, 2023.

[12] A. Tanikonda, B. K. Pandey, S. R. Peddinti, and S. R. Katragadda, "Advanced AI-driven cybersecurity solutions for proactive threat detection and response in complex ecosystems," *J. Sci. \& Technol.*, vol. 3, no. 1, 2022.

[13] Ebuka Mmaduekwe Paul, Ugochukwu Mmaduekwe Stanley, Joseph Darko Kessie, and Mukhtar Dolapo Salawudeen, "Adversarial machine learning in cybersecurity: Mitigating evolving threats in AI-powered defense systems," *World J. Adv. Eng. Technol. Sci.*, vol. 10, no. 2, pp. 309–325, Dec. 2023, doi: 10.30574/wjaets.2023.10.2.0294.

[14] S. S. Dari, K. U. Thool, Y. D. Deshpande, M. G. Aush, V. D. Patil, and S. P. Bendale, "Neural Networks and Cyber Resilience: Deep Insights into AI Architectures for Robust Security Framework," *J. Electr. Syst.*, vol. 19, no. 3, 2023.

[15] G. Rjoub *et al.*, "A survey on explainable artificial intelligence for cybersecurity," *IEEE Trans. Netw. Serv. Manag.*, vol. 20, no. 4, pp. 5115–5140, 2023.

[16] S. Tiwari, V. Sresth, and A. Srivastava, "The role of explainable AI in cybersecurity: Addressing transparency challenges in autonomous defense systems," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 9, pp. 718–733, 2020.

[17] I. Ray, S. Sreedharan, R. Podder, S. K. Bashir, and I. Ray, "Explainable AI for prioritizing and deploying defenses for cyber-physical system resiliency," in *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2023, pp. 184–192.

[18] S. Ashfaq, S. Biswas, and T. K. Chowdhury, "Integration Of Artificial Intelligence And Advanced Computing To Develop Resilient Cyber Defense Systems," *J. Sustain. Dev. Policy*, vol. 2, no. 04, pp. 74–107, 2023.

[19] Z. Zhang, H. Al 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. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[20] A. Sharma, D. Kejriwal, and A. K. Pakina, "Adversarial AI and cyber--physical system resilience: Protecting critical," *Int. J. Artif. Intell. Data Res.*, vol. 14, no. 2, 2023.

[21] I. Vaccari, A. Carlevaro, S. Narteni, E. Cambiaso, and M. Mongelli, "eXplainable and reliable against adversarial machine learning in data analytics," *IEEE Access*, vol. 10, pp. 83949–83970, 2022.

[22] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," *Front. Artif. Intell.*, vol. 8, Jan. 2025, doi: 10.3389/frai.2025.1526221.

[23] G. Petihakis, A. Farao, P. Bountakas, A. Sabazioti, J. Polley, and C. Xenakis, "AIAS: AI-ASsisted cybersecurity platform to defend against adversarial AI attacks," in *Proceedings of the 19th International Conference on Availability, Reliability and Security*, New York, NY, USA: ACM, Jul. 2024, pp. 1–7. doi: 10.1145/3664476.3669920.

[24] E. C. Nkoro, J. N. Njoku, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "Zero-Trust Marine Cyberdefense for IoT-Based Communications: An Explainable Approach," *Electronics*, vol. 13, no. 2, p. 276, Jan. 2024, doi: 10.3390/electronics13020276.

[25] M.-D. Nguyen, A. Bouaziz, V. Valdes, A. Rosa Cavalli, W. Mallouli, and E. Montes De Oca, "A deep learning anomaly detection framework with explainability and robustness," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, New York, NY, USA: ACM, Aug. 2023, pp. 1–7. doi: 10.1145/3600160.3605052.

[26] L. Almuqren, M. S. Maashi, M. Alamgeer, H. Mohsen, M. A. Hamza, and A. A. Abdelmageed, "Explainable Artificial Intelligence Enabled Intrusion Detection Technique for Secure Cyber-Physical Systems," *Appl. Sci.*, vol. 13, no. 5, p. 3081, Feb. 2023, doi: 10.3390/app13053081.

[27] S. Neupane *et al.*, "Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, 2022, doi: 10.1109/ACCESS.2022.3216617.