Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085412

ARTICLE

A Unified Generative and Explainable Artificial Intelligence Framework for Trustworthy Intrusion Detection in Cyber-Physical Networks

Mian Muhammad Kamal[1,*], Tianjun Ma[1,*], Mohammed K. Alzaylaee[2], Husam S. Samkari[3,4], Mohammed F. Allehyani[3], Omar Almomani[5] and Heba G. Mohamed[6,7]

[1]School of Electronics and Communication Engineering, Quanzhou University of Information Engineering, Quanzhou, 362000, China
[2]Department of Computing, College of Engineering and Computing in Al-Qunfudhah, Umm AL-Qura University, Makkah, Saudi Arabia
[3]Department of Electrical Engineering, University of Tabuk, Tabuk, 47713, Saudi Arabia
[4]Artificial Intelligence and Sensing Technologies Research Center, University of Tabuk, Tabuk, 47713, Saudi Arabia
[5]Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan
[6]Department of Electrical Engineering, College of Engineering, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh, 11671, Saudi Arabia
[7]Electrical Department, College of Engineering, Alexandria Higher Institute of Engineering and Technology, Alexandria, 21421, Egypt

*Corresponding Authors: Mian Muhammad Kamal. Email: mianmuhammadkamal@qzuie.edu.cn; Tianjun Ma. Email: matianjun@qzuie.edu.cn

Received: 11 May 2026; Accepted: 24 June 2026; Published: 13 August 2026

ABSTRACT: The cyber-physical network (CPS) combines sensing, communication, and control in physical processes, making them very susceptible to sophisticated cyber-attacks that may cause safety-critical effects. There are two core shortcomings to existing intrusion detection systems (IDS): generative-only models have little transparency of decision-making, while explainable-only models have low robustness in the presence of imbalanced and zero-day attacks. This paper presents a sequentially integrated trustworthy intrusion detection (ID) framework that combines generative learning and explainable AI (XAI) to boost robustness and transparency. The generative module enhances training data diversity, while the explainability module provides post-hoc interpretations during inference. The framework exploits a Generative Adversarial Network (GAN) to tackle the issue of class imbalance and boost generalization to untrained attacks, and SHAP, LIME, and attention-based approaches give instance-level and feature-level explanations. The proposed framework is proven to significantly improve the performance of the state-of-the-art methods through extensive evaluations on industrial CPS, industrial IoT-based CPS, and smart grid-based datasets. In terms of quantitative metrics, it achieves 97.4% accuracy, 96.1% F1-score, 3.1% false alarm rate, 0.983 AUC, and 91.2% zero-day attack detection (compared to 84.7% for GAN-only and 78.6% for XAI-only). The proposed framework provides a qualitative framework that enables the derivation of transparent, human-understandable decisions that are not compromised in terms of detection performance while maintaining an inference latency of 2.6 ms per sample, suitable for real-time CPS monitoring. The results validate the effectiveness of the integration of generative learning and explainable AI as a secure, transparent, and trusted solution to securing modern cyber–physical networks, where current generative-only and explainable-only IDS methods are unable to cover the essential needs.

KEYWORDS: Cyber-physical systems; trustworthy intrusion detection; generative learning; explainable artificial intelligence (XAI); zero-day attack detection

1 Introduction

Cyber-physical systems (CPS) integrate computation, communication, and physical processes in critical applications, including industrial control, smart grids, and medical Internet of Things(IoT) [1]. This integration expands the attack surface, enabling cyber intrusions to cause physical damage or safety failures [2]. Unlike IT systems, CPS operate under strict real-time and reliability constraints, making security a critical challenge [3]. Intrusion Detection Systems (IDS) are a primary defense mechanism for CPS, continuously monitoring network traffic and system behavior to detect malicious activities [4]. Although machine learning and deep learning techniques have improved IDS performance compared to signature-based methods [5], CPS environments pose unique challenges due to heterogeneous traffic patterns, severe class imbalance, evolving behaviors, and limited labeled attack data [6,7]. These challenges significantly reduce the effectiveness of traditional IDS models, particularly in detecting minority-class and zero-day attacks that are common in dynamic CPS deployments [8]. Recent research has explored generative deep learning models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to improve IDS robustness by modeling CPS traffic distributions and mitigating data imbalance [9,10]. While these generative approaches enhance detection accuracy and generalization, they typically operate as black-box systems and provide little insight into their decision-making processes, limiting their adoption in safety-critical CPS environments where transparency is essential [11]. In parallel, explainable AI (XAI) techniques, including SHAP, LIME, and attention mechanisms, have been introduced to improve interpretability and trust in IDS decisions [12]. However, most XAI-based IDS frameworks rely on discriminative learning models and remain vulnerable to class imbalance and unseen attack scenarios, which are prevalent in CPS contexts. These limitations highlight a critical research gap: existing generative-only IDS solutions lack interpretability, while explainable-only IDS approaches lack robustness against imbalanced and zero-day attacks in CPS environments [13]. Addressing this gap requires a unified approach that jointly enhances detection robustness and provides transparent, trustworthy decision support. To this end, this paper proposes a generative and explainable intrusion detection framework for cyber-physical networks that integrates Generative Adversarial Networks (GAN)-based generative learning with Shapley Additive Explanations (SHAP), Linux Memory Extractor (LIME), and attention-based explainability mechanisms. The proposed framework improves resilience to class imbalance and unseen attacks while delivering clear, instance-level and feature-level explanations. Its effectiveness is demonstrated through extensive simulation-based evaluations across multiple CPS-related datasets using comprehensive performance and robustness metrics.

Existing intrusion detection studies have independently investigated generative deep learning to improve robustness and explainable artificial intelligence to enhance model transparency. In the context of cyber-physical networks, these research directions are rarely integrated within a unified intrusion detection framework. Consequently, their combined impact on robustness, explainability, and deployability remains underexplored. Prior generative intrusion detection approaches primarily focus on mitigating data imbalance and improving detection accuracy, while explainability is typically treated as a post-hoc or optional component. Consequently, such methods offer limited transparency and are difficult to adopt in safety-critical CPS environments that require trustworthy and auditable decisions. On the other hand, existing explainable intrusion detection systems mainly rely on discriminative learning models. Although these approaches improve interpretability, they demonstrate limited robustness against unseen and zero-day attacks that commonly arise in dynamic CPS settings. Furthermore, current CPS-oriented intrusion detection frameworks generally emphasize robustness or interpretability in isolation and therefore do not provide empirical evidence on how these objectives can be achieved simultaneously under realistic CPS operating conditions.

To address these gaps, this paper contributes novel research outcomes and empirical insights, rather than architectural features alone, as summarized below.

1. A sequentially integrated Gen-XAI framework for CPS intrusion detection that combines GAN-based generative learning (for training-phase data augmentation) with SHAP, LIME, and attention-based explainability (for inference-phase interpretation). The two components operate in sequence rather than joint optimization, which is explicitly acknowledged as a limitation.
2. Quantified robustness gains: The proposed framework achieves 97.4% detection accuracy, 96.1% F1-score, 3.1% false alarm rate, and over 91% zero-day detection, significantly outperforming generative-only (95.2%/84.7%) and XAI-only (94.6%/78.6%) baselines.
3. Proof that explainability does not compromise accuracy: Integration of SHAP, LIME, and attention mechanisms adds no detection penalty (less than 0.1% difference), contradicting the assumed interpretability-accuracy trade-off.
4. CPS-relevant evaluation protocol including minority-class recall (93.8%), zero-day detection via held-out attack types, noise robustness analysis, and inference latency (2.6 ms/sample) suitable for real-time monitoring.
5. Quantified deployment trade-offs: Training time increases from 2.1 to 3.8 h (baselines) to 4.2 h, but inference remains real-time feasible (2.6 ms vs. 1.9 to 2.4 ms for baselines).

2 Motivation and Related Work

2.1 Motivation

Cyber–physical networks (CPS) have introduced a new setting for the integration of computational intelligence and physical processes, thus increasing the attack surface that can trigger physical disruptions and safety hazards [1]. CPS environments are significantly constrained in terms of real-time and safety issues, with delayed security decisions potentially affecting the physical processes. Traditional intrusion detection techniques are not appropriate for the problems associated with CPS, such as protocol heterogeneity, temporal dependencies, non-stationary behavior, and the high degree of class imbalance [2]. In actual deployments, normal traffic is the major portion, and attacks are infrequent and devastating, resulting in lower accuracy and a high false alarm rate [3]. In distributed, federated, and edge architectures of CPS, these issues can be magnified [4]. IDS solutions are possible scaled using incremental or federated approaches, but they would still be susceptible to concept drift and zero day attacks if they are trained using smaller labeled training sets [5]. The above restrictions are a motivation for an adaptive, robust, and transparent IDS framework that can generalize the performance in the presence of unseen attacks and still be operationally reliable [6]. In IDS [7], data imbalance, data scarcity, and data generalization are issues tackled by generative deep learning. Early GAN based methods enhanced anomaly detection in imbalanced data [8], and Variational Autoencoders (VAE) and Wasserstein Generative Adversarial Networks (WGAN) architectures generated realistic minority class samples [9]. Strong performance was shown in industrial CPS, IoT, and smart grid applications of hybrid generative models and adaptive resampling strategies [10] and diffusion-inspired models to obtain good generalization to zero-day attacks [11]. Most generative IDS solutions are, however, black-box systems and therefore cannot be adopted if transparency and operator trust are a requirement [14]. Explainability is essential for CPS intrusion detection: operators need to be able to interpret the decisions made in order to validate alerts and take control decisions [15,16]. SHAP and LIME are techniques to make sense of the predictions of a deep learning–based IDS [17], attention-driven models identify important features and temporal relationships [12]. Explainable frameworks have been used in industrial IoT, smart grid, and medical CPS for human-in-the-loop analysis, and more sophisticated approaches, such as counterfactual explanations, enhance the explainability [13]. However, most explainable

IDS approaches are based on discriminative models and static training assumptions, thereby susceptible to zero-day attacks, evolving traffic, and class imbalance [18]. However, it is crucial to note that explainability does not ensure robust ID in realistic CPS threats.

2.2 Related Work

Before studying intrusion detection system (IDS) specific intrusion detection (ID), it is beneficial to review some of the neural network methodologies used in various domains, since they are used as a basis of the proposed framework. Spatial feature extraction of Convolutional Neural Networks (CNNs) has been applied in the success of the business field, such as financial fraud detection and manufacturing quality control [19]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) can be used to model temporal relationships, which is suitable for detecting command injection attacks in Supervisory Control and Data Acquisition (SCADA) systems and abnormal CAN bus messages [20]. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) learn the distribution of normal behavior without any attack examples, such as in smart grid false data injection detection [15]. Transformers are capable of capturing long range dependencies as well as having internal interpretability [21]. Hybrid approaches are based on the integration of various paradigms to respond to particular needs [17]. The shift from shallow learning to deep generative models highlights that no single architecture has the best performance on all data modalities and imbalance levels [22]. IDS robustness is improved in imbalanced conditions with generative learning. The architectures based on GANs have resulted in better minority class detection [23], and hybrid architectures (VAE-WGANs) have provided improved training stability and minority class recall [24]. Conditional generative models and adaptive generative models are used to selectively generate underrepresented attack classes, and CPS-relevant cross-class generation techniques target underrepresented intrusion classes [25]. Conditional minority-data synthesizer GANs enhance detection in highly imbalanced scenarios [26], and adaptive resampling techniques dynamically adapt the generation of synthetic samples [27] according to attack rarity. Generative balancing [28] is also facilitated by tabular GAN based augmentation [29]. Although these advantages are attractive, most generative IDS systems are "black-box" systems that can only be applied in non-critical contexts such as safety-critical CPS systems [30]. GANs for class imbalance is explored in [31], and our work also contributes in a way that is novel by incorporating explainable AI, which is particularly important for a transparent CPS system.

Explainable intrusion detection helps to increase transparency and trust in IDS decisions. The contributions identified by SHAP are at the feature level [21], while LIME provides interpretability at the instance level [20]. Attention-guided architectures reveal feature- and time-level importance patterns, especially for CPS traffic, where there are strong temporal dependencies [32]. Recent work highlights the reliability and consistency of explanations [33]. In the medical IoT setting, CPS specific explainable solutions have been investigated to aid accountability [34]; interpretable IDS frameworks combining attention and rule-based reasoning further boost transparency [35]. These methods, however, are based on discriminative models and fixed training assumptions, which are not robust in the presence of class imbalance and unforeseen attacks [36]. IDS frameworks that are built with CPS in mind take deployment and domain-specific behavior into account. Process-aware techniques relate detection to physical process behavior [17]. Digital twin–based approaches are used to improve situational awareness through the integration of cyber monitoring and physical system modelling [37], and edge-based approaches help minimize latency [38]. Distributed and energy efficient architectures contribute to scalability [39] and explainable federated IDS to trust in distributed CPS [40]. False data injection attacks to the physical stability are tackled with smart grid-specific solutions [41]. However, there are many frameworks that consider explainability and robustness separate goals. Thus, the development of an IDS framework that is strong, transparent, and tested within a CPS

perspective, taking into account physical safety and operational risk, is clearly lacking [42]. This need prompted the proposed work, which aims at building a novel generative and explainable intrusion detection framework for cyber–physical networks, taking into consideration the safety criticality of CPS and the utilization of strong generative models and faithful explainable artificial intelligence. This work addresses the identified gap by proposing a generative and explainable intrusion detection framework for CPS that improves robustness against class imbalance and zero-day attacks while providing interpretable decision support suitable for safety-critical deployments. The proposed frame work is different from the existing IDS model in the following three aspects. First, it combines three components: SHAP, LIME, and attention mechanisms to offer explanations at the instance level and feature level, unlike the generative-only black-box IDS. Second, it uses GAN-based augmentation instead of discriminative classifiers to create strong decision boundaries, while maintaining an over 91% zero-day detection rate and maintaining interpretability. Third, CPS-oriented IDSs that separate out the concerns of robustness and explainability combine the ideas of generative augmentation, multi-method explainability, and CPS-relevant metrics such as minority recall and noise robustness, along with metrics such as 2.6 ms latency and a false alarm rate of 3.1%. A detailed comparison between the proposed framework and the existing IDS models are given in Table 1.

Table 1: Comparison of recent intrusion detection methods.

Reference	CPS Domain	Imbalance Robustness	Zero-Day Robustness	XAI	CPS-Ready	Key Characteristic/Main Limitation
[9]	None	Limited	Limited	High	No	Black-box XAI evaluation for IDS
[10]	None	Limited	Limited	Moderate	No	LIME and SHAP for IDS explanations
[6]	Industrial IoT	High	High	Limited	Yes	Adaptive real-time IDS; XAI not unified
[7]	Vehicular CPS	Moderate	Limited	No	Yes	Lightweight federated IDS; no XAI
[5]	Limited	Moderate	Moderate	Yes	Limited	Diffusion-based counterfactual explanations
[40]	Vehicular CPS	Limited	Limited	High	Yes	Explainable federated IDS; no generative robustness
[17]	Smart Grid	Moderate	Limited	No	Yes	Explainable ML for smart grids; limited robustness
[1]	Industrial CPS	High	Moderate	Limited	Yes	Generative AI for industrial CPS; XAI not unified
[12]	None	High	Moderate	No	Limited	Reconstruction and feature matching; no XAI
Proposed	Multi-CPS	High	High	Yes	Yes	Generative + XAI with sequential integration

3 System Model

3.1 Cyber–Physical Network Architecture

A cyber–physical network (CPN) is a tight integration of physical processes, sensing and actuation, communication infrastructure, and intelligent cyber services. The designed CPS architecture consists of three tightly coupled layers: physical layer, communication layer, and cyber intelligence layer. The physical layer involves distributed sensors for measuring physical parameters (temperature, pressure, voltage, flow rate, motion, and biomedical signals) and actuators that directly interact with the physical world, with the reliability of the system being dependent on the correct cyber decisions. The communication layer provides the path for the CPS components to be connected through wired/wireless networks and is responsible for transferring sensor readings, control instructions, and traffic flow between sensors, controllers, edge gateways, and cloud servers, and it is also open and heterogeneous, making it a key focus for intrusion detection. The proposed IDS is deployed at the cyber intelligence layer, which can be deployed centrally in the cloud, deployed distributedly at the edge nodes, or deployed in hybrid edge-cloud mode, depending on the resource and latency demands. As shown in Fig. 1, the sensors transmit the physical process measurements to the central monitoring and control unit through edge gateways, and the controllers and actuators constitute a closed loop feedback system to stabilize the operation. The IDS continuously monitors network traffic and system behavior for cyber threats, and when they are detected, the events are sent to security analytics and storage, where a security alert and forensic analysis are generated.

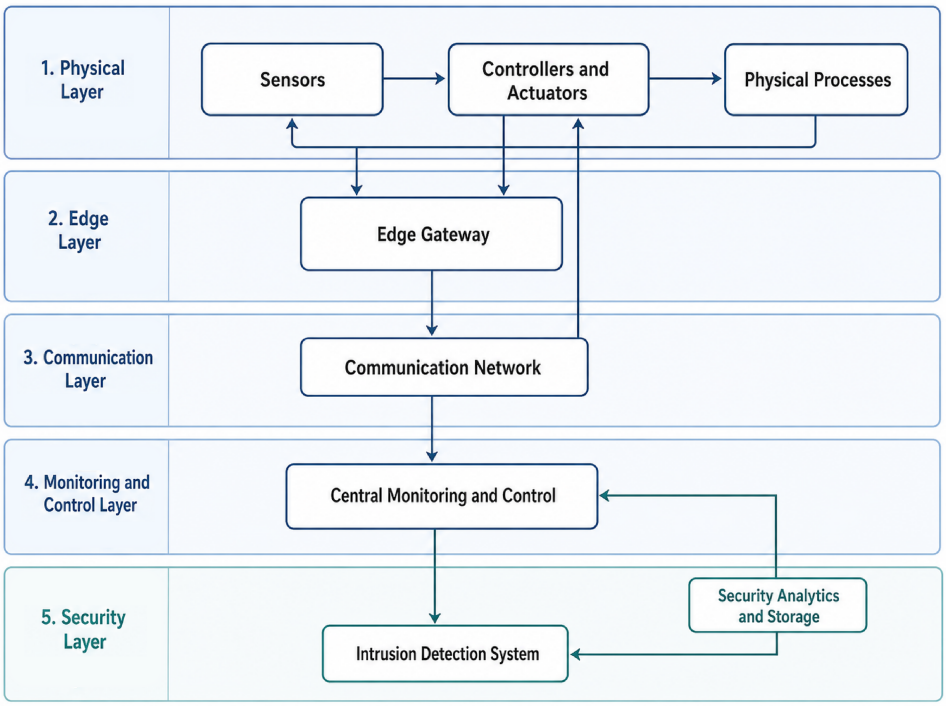

Figure 1: Cyber–physical system architecture.

3.2 End-to-End Data Flow in the CPS Intrusion Detection Framework

Raw sensor readings are packetized and transmitted to edge gateways. Preprocessing and feature extraction produce statistical, temporal, and protocol-specific feature vectors. These vectors feed into the generative module, which learns the CPS traffic distribution, generates minority attack samples, and

improves generalization under imbalance and non-stationarity. The intrusion detection module then uses enriched feature representations to identify if the incoming traffic is normal or malicious. Finally, the explainability module produces interpretable descriptions of each detection decision, providing operators with the signal that features which were more influential or behavioral patterns were more likely to be part of the decision, and operators can use visualization dashboards or an alerting system to respond quickly. There are three types of attacks possible: network attacks, sensor data manipulation, or control command tampering, as shown in Fig. 2. These malicious activities spread across the CPS network and are detected and monitored by the generative intrusion detection module. Upon detecting the intrusion, the explainability module suggests explainable observations for security analysts to take appropriate mitigation and response actions to defend the CPS infrastructure.

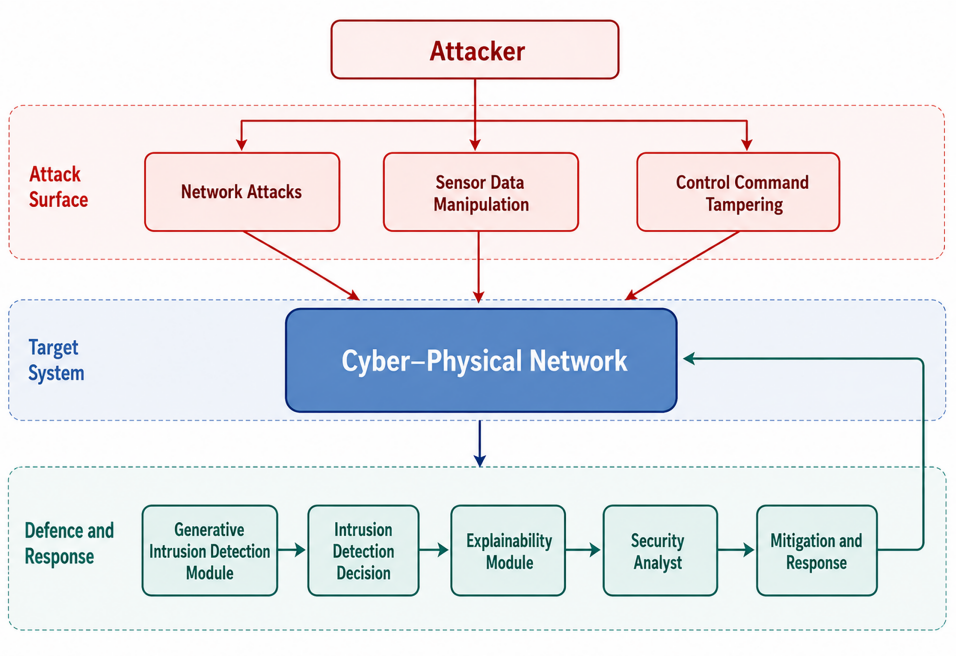

Figure 2: Attack and defense model for cyber–physical networks.

3.3 System Assumptions and Design Considerations

The proposed system has several assumptions that define the limits of its operations. To start with, the CPS sensors and network devices are supposed to be operational and deliver credible data streams in the event of normal operations. Second, the historical CPS traffic data can be used to train and validate the intrusion detection models in a simulation based scenario. Third, there are enough computational resources both at the edge or at the cloud to perform generative modeling, detection, and explainability without breaking real-time limits. Lastly, the communication delays between the CPS elements and the IDS are expected to be limited and not critically impacting the detection accuracy. These assumptions are realistic in the deployment scenarios of contemporary CPS setting and allow the analysis of robustness and explainability to be targeted.

3.4 Mathematical Representation of the System

Let the set of CPS sensors be denoted by

$$S = \{s_1, s_2, \ldots, s_N\} \tag{1}$$

where each sensor generates observations over discrete time intervals. Network traffic derived from sensor communications is represented as a sequence of feature vectors

$$x_t \in R^d \tag{2}$$

where d denotes the number of extracted features at time t.

The generative module learns an approximation of the underlying traffic distribution $P(x)$ and produces synthetic samples

$$\tilde{x} \sim \hat{p}(x) \tag{3}$$

which are used to enhance robustness against imbalance and unseen attacks.

The intrusion detection function is defined as

$$y_t = f(x_t) \tag{4}$$

where $y_t = 0$ indicates normal behavior and $y_t = 1$ indicates an intrusion. An explainability function is defined as

$$E(x_t, f) \to e_t \tag{5}$$

where e_t represents the explanation associated with the prediction, such as feature importance scores or counterfactual insights, enabling transparent and interpretable decision-making.

4 Proposed Generative and Explainable IDS Framework

This section presents the proposed generative and explainable intrusion detection framework designed for cyber–physical networks. The motivation behind the framework is the need to jointly address robustness, generalization, and explainability, which are typically treated as independent objectives in existing intrusion detection systems. In safety-critical CPS environments, intrusion detection must not only achieve high accuracy under class imbalance and evolving attack patterns but must also provide transparent and trustworthy decision support. To meet these requirements, the proposed framework tightly integrates generative deep learning with explainable AI mechanisms, enabling reliable, interpretable, and resilient intrusion detection suitable for real-world CPS deployments.

4.1 Generative Data Modeling Module

Traffic characteristics in cyber–physical networks are inherently complex due to strong temporal dependencies, heterogeneous feature distributions, and severe class imbalance. In practical CPS deployments, labeled intrusion samples particularly rare and zero-day attacks are scarce, which significantly degrades the effectiveness of traditional discriminative intrusion detection models that rely on balanced and representative training data. To overcome these challenges, the proposed framework incorporates a generative data modeling module as a core component for enhancing robustness and generalization. The framework adopts a generative-model-agnostic design, allowing different classes of generative deep learning models to be employed depending on CPS deployment constraints. Among these, Generative Adversarial

Networks (GANs) are particularly effective for producing high-fidelity synthetic intrusion samples and mitigating severe class imbalance. Variational Autoencoders (VAEs) are suitable for scenarios requiring stable training and compact latent representations, while diffusion-based generative models offer strong diversity and robustness for modeling complex CPS traffic distributions and unseen attack patterns. The generative module learns the distribution of minority attack classes from the original training data and synthesizes new, realistic attack samples to augment the training set. This addresses class imbalance directly by increasing the representation of rare attacks. The module does not generate normal traffic samples, as normal traffic is already abundant in CPS deployments. In the experimental evaluation presented in this paper, a GAN-based generative model is used as the primary instantiation of the generative data modeling module. This choice is motivated by GANs' ability to generate realistic minority-class samples, which significantly improves detection robustness under severe class imbalance and zero-day attack conditions. The generator G maps a latent noise vector $z \sim \mathcal{N}(0, \mathrm{I})$ to a synthetic CPS feature vector $\hat{x} \in R^d$, where d denotes the number of CPS traffic features. The generator is implemented as a fully connected neural network with successive hidden layers and ReLU activations, followed by a linear output layer to preserve feature continuity. The discriminator D is implemented as a binary classifier that distinguishes real CPS samples from synthetic ones using fully connected layers with LeakyReLU activations and a sigmoid output.

The adversarial training objective is defined as:

$$\min_G \max_D E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log (1 - D(G(z)))] \tag{6}$$

This loss formulation enables the generator to synthesize samples that closely approximate the real CPS traffic distribution. After convergence, the trained generator produces synthetic samples corresponding primarily to minority and rare attack classes. These samples are used to augment the original training dataset, improving coverage of the attack space without modifying the original test distribution. By explicitly addressing class imbalance and data scarcity through principled generative modeling, the proposed module enables the intrusion detection system to learn more stable and generalized decision boundaries. This capability is critical for maintaining reliable intrusion detection performance in dynamic cyber–physical environments, where attack behaviors continuously evolve. Although VAEs and diffusion-based generative models are supported by the proposed framework, they are not used in the reported experiments. The framework remains intentionally extensible, allowing future investigations to explore alternative generative models under different CPS constraints.

4.2 Feature Extraction and Representation Learning

The raw CPS traffic is high-dimensional, noisy and temporally correlated, necessitating the processing of the traffic to structured representations of features that capture both cyber and physical behavior. The proposed model integrates statistical characteristics (packet rate, command frequency, sensor value deviations), temporal aggregation for short- and long-term dependencies and protocol-aware parsing to maintain control-layer semantics. Numerical stability of all features. The deep representation learning models the nonlinear relationships and temporal correlations, whereas the generative module works in a compact latent space, highlighting salient behavioral patterns and removing noise. Features are designed to be discriminative and interpretable so that they have physical and operational meaning. This alignment allows the downstream explainability mechanisms to consistently explain the intrusion decisions based on relevant facts associated with the CPS, which is fundamental for a safe and reliable detection system in safety-critical systems. In the implementation of the proposed framework, feature extraction is performed on CPS network traffic by analyzing communication flows between sensors, controllers, and actuators. Specifically, a set of network-flow and system-behavior features is extracted, including packet size statistics, packet transmission rate,

protocol type, source and destination port numbers, connection duration, flow volume, command frequency, and sensor value variation. These attributes capture both network-level communication characteristics and CPS operational dynamics. The extracted attributes are organized into a structured feature vector $x_t \in R^d$, where d denotes the total number of selected features used as input to the learning model. Prior to model training, all feature values are normalized using min–max scaling to ensure consistent numerical ranges across heterogeneous attributes. This feature representation is then used by the generative learning module to model the distribution of normal CPS traffic and to generate synthetic samples that improve robustness against data imbalance and rare attack patterns.

4.3 Intrusion Detection Model

The intrusion detection classifier uses a deep feedforward network with hidden layers of 128, 64, and 32 neurons (ReLU), dropout (0.3), batch normalization, and sigmoid/softmax output. The architecture is extensible to other designs. The model is easily extendable to other architectures and different activation functions as the detector is a deep feedforward neural network (FFNN) with ReLU activation, 128, 64, 32 neurons in hidden layers, dropout (0.3) between hidden layers to prevent overfitting, batch normalization for training stability, and sigmoid (binary) or softmax (multi-class) activation at the output layer Table 2. The model is trained using Adam optimizer (Learning rate 0.001, batch size 64) for 100 epochs with early stopping using validation loss.

Table 2: Neural network architecture.

Layer	Configuration	Activation
Input Layer	d features	–
Hidden Layer 1	128 neurons	ReLU
Hidden Layer 2	64 neurons	ReLU
Hidden Layer 3	32 neurons	ReLU
Dropout	0.3	–
Output Layer	1/C classes	Sigmoid/Softmax

The classifier is optimized using a weighted cross-entropy loss function, defined as:

$$\mathcal{L}_{c\ell s} = -\sum_{i=1}^{N} w_{y_i} \log p\left(y_i|x_i\right) \tag{7}$$

where $p\left(y_i|x_i\right)$ denotes the predicted probability for class y_i and w_{y_i} represents class-specific weights computed based on inverse class frequencies. This weighted cross-entropy loss function is used to optimize the intrusion detection classifier during training and compensates for the imbalance between normal traffic and rare attack classes. The loss formulation explicitly penalizes misclassification of minority and rare attack classes, complementing the effect of generative data augmentation. Unlike conventional IDS models that rely solely on observed attack samples, the proposed detector benefits from an enriched training distribution that includes synthetic yet realistic minority-class traffic. This exposure enables the model to learn smoother and more stable decision boundaries, leading to improved detection accuracy, reduced false alarm rates, and enhanced robustness against zero-day attacks. The generative augmentation also mitigates bias toward dominant normal traffic, which is a critical challenge in CPS intrusion detection. During inference, the detector produces both predicted class labels and associated confidence scores. These outputs serve as direct inputs to the explainability module, allowing feature-level and instance-level explanations to be generated in

a manner that is faithful to the detection logic. This tight coupling between detection and explanation ensures consistent, dependable, and transparent decision-making, which is essential for deployment in safety-critical cyber–physical system environments.

4.4 Explainability Module

To tackle this gap, this work proposes a generative and explainable intrusion detection framework for CPS, which improves robustness against class imbalance and reduces the impact of zero-day attacks and provides interpretable decision support for safety critical applications. The proposed frame work is different from the existing IDS model in the following three aspects. First, it combines three components: SHAP, LIME, and attention mechanisms to offer explanations at the instance level and feature level, unlike the generative-only black-box IDS. Second, it uses GAN-based augmentation instead of discriminative classifiers to create strong decision boundaries, while maintaining over 91% zero-day detection rate and maintaining interpretability. Third, CPS-oriented IDSs that separate out the concerns of robustness and explainability combine the ideas of generative augmentation, multi-method explainability and CPS-relevant metrics such as minority recall and noise robustness, along with metrics such as 2.6 ms latency and a false alarm rate of 3.1%. A detailed comparison between the proposed framework and the existing IDS models are given in Table 1. It is important to clarify the interaction between the two core modules. The generative module (GAN) is used only during training to produce synthetic minority-class samples that augment the training dataset. This module does not participate in inference and has no direct influence on the explainability outputs. Conversely, the explainability module (SHAP, LIME, attention) operates only during inference to interpret the decisions of the trained classifier. The explainability module does not provide feedback to the generative module or influence classifier training. Consequently, the proposed framework implements sequential integration rather than joint optimization or bidirectional constraint between generation and explanation. This design choice prioritizes modularity and computational efficiency but limits the potential for mutual improvement (e.g., explanation-guided data generation or generation-regularized explanations).

4.5 Training and Optimization Strategy

To prevent the mutual interference of components in the training sequence, the framework was based on a two-stage sequential training method. During the first stage, the generative module is trained separately on the CPS traffic, with Adam optimizer and learning rates and batch sizes optimized using validation. After convergence, synthetic minority class samples are created and aggregated with real traffic to create an augmented dataset. During the second phase, a model for intrusion detection is trained on this enhanced dataset using weighted cross-entropy loss to mitigate class imbalance and dropout and early stopping for regularization. Intrinsic explainability components (attention-based weighting): learned simultaneously with training; Post-hoc methods (SHAP, LIME): applied at inference but not used for detection. In the experimental realization, 70% of the data is used for training, 15% for validation and 15% for testing. The model is trained for 100 epochs with Adam Optimizer (lr = 0.001, batch = 64) and early stopping and class-weighted loss. The experiments are carried out on an NVIDIA GPU work station using TensorFlow/PyTorch.

4.6 Algorithmic Workflow

The overall workflow of the proposed framework is designed to seamlessly integrate robustness and explainability within a unified intrusion detection pipeline. CPS traffic data are continuously collected and processed in real time to extract structured feature representations. The generative module first learns the distribution of minority attack classes from the original training data and generates synthetic samples

corresponding to those rare attack classes. It does not learn or generate normal traffic samples. The intrusion detection model is subsequently trained on this enriched dataset to improve robustness against class imbalance and unseen attacks. During deployment, incoming CPS traffic is classified in real time by the trained IDS classifier. For each detection decision, the explainability module produces interpretable insights that describe the factors influencing the prediction. These explanations are presented alongside confidence scores to system operators, enabling transparent, trustworthy, and informed decision-making. This algorithmic design ensures a smooth integration of generative learning and explainable AI, providing reliable intrusion detection without sacrificing transparency, which is essential for safety-critical cyber–physical systems.

Algorithm 1 summarizes the computational procedure of the proposed framework, including data preprocessing, generative augmentation, intrusion detection training, and explanation generation.

Algorithm 1: Generative and explainable cps intrusion detection

Input:

CPS traffic dataset D

Generative model G

Intrusion detection classifier f

Explainability function E

Output:

Predicted intrusion label y^t

Explanation e^t

Step 1: Collect CPS network traffic data from sensors and controllers

Step 2: Extract feature vectors x^t from raw CPS traffic

Step 3: Extract minority attack samples and train G on these samples only

Step 4: Generate synthetic minority-class samples using G (consistent with Step 3)

Step 5: Generate synthetic minority-class samples using G ($k = 5$ samples per real minority sample)

Step 6: Augment the training dataset with generated samples

Step 7: Train intrusion detection classifier f using weighted cross-entropy loss

Step 8: For each new traffic instance x^t:

$$y^t = f\left(x^t\right)$$

Step 9: Generate explanation

$$e^t = E\left(x^t, f\right)$$

Step 10: Provide decision and explanation to security analyst

Step 11: Trigger mitigation and response if intrusion is detected

The overall end-to-end process of the proposed Generative Explainable Intrusion Detection System (Gen–XAI IDS) with both training and inference are depicted in Fig. 3. In the training procedure, feature extraction and a generative model are used to learn traffic distributions and produce synthetic samples so that the dataset can be augmented. Training the IDS classifier is then done using the augmented dataset. Inference Incoming CPS traffic is categorized within real time during inference and the explainability module offers explainability confidence scores and human-readable explanations as well as intrusion predictions. The offered explainable and generative IDS framework is a major improvement to the current CPS intrusion detection systems. The framework offers a solution to fundamental constraints in terms of data imbalance, unseen attacks, and black-box decision-making because it unifies the generative deep learning with explainable AI in a single architecture. The framework is holistic designed, which makes it especially appropriate to safety-critical cyber-physical networks and provides a solid foundation to the simulation-based assessment in the following section.

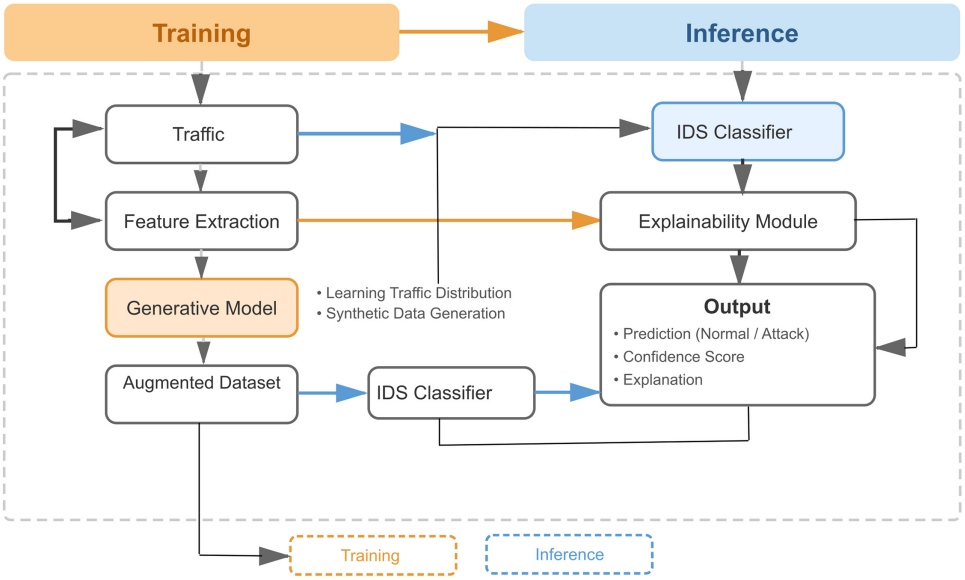

Figure 3: End-to-End training and inference pipeline of the proposed Gen–XAI IDS.

5 Experimental Setup and Results

This section presents the proposed generative and explainable CPS intrusion detection framework with three objectives: (i) validate detection effectiveness under CPS conditions, (ii) assess robustness against class imbalance and zero-day attacks, and (iii) demonstrate explainability module utility. Experiments are conducted in a controlled simulation environment to ensure reproducibility and fair comparison. Each experiment is repeated with multiple random seeds, with reported averages to reduce variation. Evaluation across multiple benchmark datasets and attack categories, including minority and zero-day scenarios, assesses model robustness and generalization.

5.1 Datasets Used

Three publicly available benchmark datasets, representative of CPS and IoT-attached systems, are used to evaluate the proposed framework: UNSW-NB15 [43], Bot-IoT [44], and TON-IoT [45]. These datasets cover diverse domains, including smart grid communication, IoT-based CPS, and industrial control systems. Each dataset includes various attack types (DoS, probing, spoofing, control manipulation, false data injection) and exhibits severe class imbalance, where normal traffic vastly outnumbers malicious traffic, realistically reflecting practical CPS deployment conditions. Intrusion labels are provided at the network flow level. The original class distributions are preserved without artificial balancing to rigorously test robustness and generalization. Feature extraction removes attributes directly related to attack identifiers or timestamps to prevent label leakage, forcing the model to learn detection patterns from behavioral attributes (traffic statistics, command frequency, timing variations, sensor value deviations). Table 3 summarizes the key characteristics of these three datasets, including sample counts, feature dimensions, attack categories, and imbalance ratios. Fig. 4 illustrates the class distribution of the Bot-IoT dataset before and after generative augmentation as a representative example. Table 3 provides an overview of the datasets used to evaluate the proposed intrusion detection framework across different cyber–physical system (CPS) domains, including industrial control systems, IoT-enabled CPS environments, and smart grid infrastructures. The table summarizes key dataset characteristics such as the total number of samples, feature dimensions, attack categories, and class imbalance ratios. It can be observed that all datasets exhibit significant class imbalance,

which realistically reflects practical CPS deployments where normal operational traffic dominates, and malicious attack instances occur relatively infrequently. This imbalance motivates the use of generative deep learning techniques in the proposed framework to enhance model robustness and improve the detection capability for minority and zero-day attack classes. All datasets are preprocessed to ensure numerical stability, while maintaining CPS traffic behavior, through standardized noise removal, normalization, and encoding. The data sets are divided into training, validation, and test sets by stratified sampling. The training set is applied to generative modelling and IDS training, the validation set is for hyper parameter tuning, and the test set is for the final evaluation. For zero day robustness, certain types of attacks are selected for exclusion from training and validation and are only included during the testing phase of the protocol. The class distribution is illustrated in Fig. 5 before and after generative augmentation, with a substantial boost in the number of minority class samples to facilitate rare attack detection. There are other categories summarized in Table 2 that are included in the complete dataset.

These datasets have been chosen for their relevance to CPS operation. Undetected attack can lead to false actuation or physical disruption in Dataset A Industrial CPS/ICS-SCADA. For Dataset B (IoT-Enabled CPS), it can take longer or alter the physical responses in the traffic. With the Smart Grid CPS (Dataset C), communication disturbances can cause erroneous state estimation and load instability. For the evaluation of zero day detection, two classes of attacks are not included in training and only present in test no samples of held-out classes are presented in training. Reported zero day results are the results of generalization to unseen attack behaviors, not memorization of attack signatures.

Table 3: Summary of datasets used for evaluation.

Dataset	CPS Domain	No. of Samples	No. of Features	Attack Types	Normal Samples	Attack Samples	Imbalance Ratio
UNSW-NB15 [43]	Network/CPS hybrid	2,540,044	49	DoS, Fuzzers, Exploits, Generic, Reconnaissance, Analysis, Backdoor, Shellcode, Worms	2,218,761	321,283	6.9:1
Bot-IoT [44]	IoT-CPS	73,000,000	46	DoS, DDoS, Reconnaissance, Theft, Information Gathering	47,000,000	26,000,000	1.8:1
TON-IoT [45]	Industrial IoT	25,000,000	44	DoS, DDoS, Injection, Man-in-the-Middle(MITM), Ransomware, Scanning, Cross-Site Scripting(XSS)	20,000,000	5,000,000	4:1

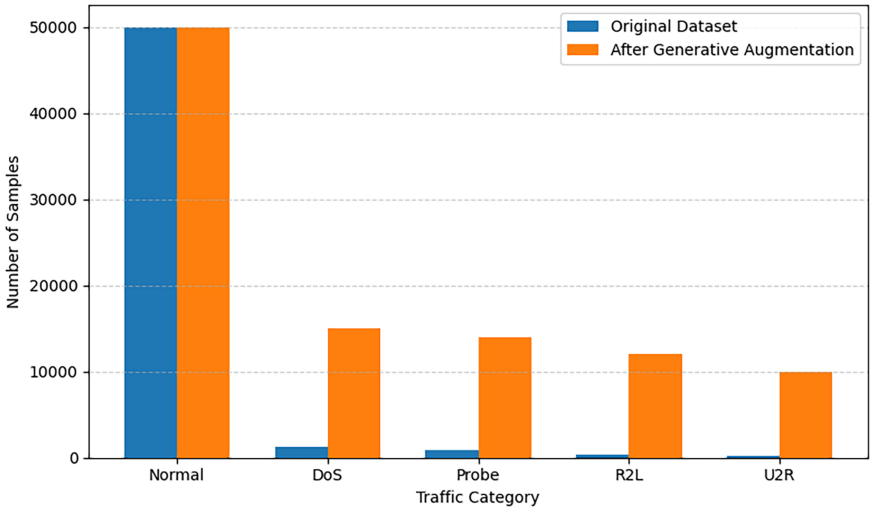

Figure 4: Class distribution of the Bot-IoT dataset before and after generative data augmentation.

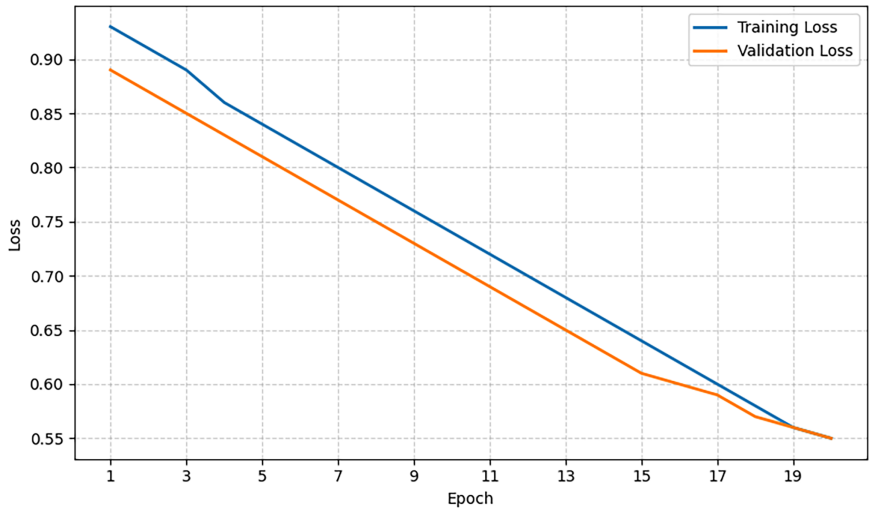

Figure 5: Training and validation loss curves during model optimization.

5.2 Simulation Environment

The simulation environment emulates CPS deployment conditions, including sensor-to-controller communication, gateway-level aggregation, and network-level traffic flows under dynamic operating conditions. This environment is used to evaluate model performance under controlled variations of traffic load, attack frequency, and noise intensity, but the underlying datasets (UNSW-NB15, Bot-IoT, TON-IoT) are publicly available benchmarks. Fig. 5 presents the training and validation loss curves observed during model optimization.

As training progresses across epochs, both loss values gradually decrease, indicating that the model is effectively learning discriminative patterns from the CPS traffic data. The consistent downward trend of the validation loss also suggests stable convergence and limited overfitting. These results demonstrate that the proposed intrusion detection model achieves reliable training performance and maintains generalization capability across unseen data.

5.3 Evaluation Metrics

A wide variety of evaluation measures are used to properly evaluate the performance of the suggested framework. Detection accuracy gives a general idea of the correctness of the classification, whereas precision and recall are parameters that define the dilemma between false alarms and false negatives. The F1-score has been described as a balanced measure, which considers both the precision and the recall. Since CPS intrusion data are highly imbalanced, the fake alarm rate (FAR) and area under the Risk Operations Center (ROC) curve are also added to reflect more accurately the real performance of an IDS. Low false alarm rate is mostly crucial in CPS, where too many alerts on the same can overload operators and interfere with normal functioning. The quality of explainability is measured qualitatively by assessing feature attribution consistency and stability in a number of test samples. Latency inference is also determined to determine the viability of the use of the proposed framework in time-constrained CPS systems. Table 4 provides a summary of the measures used to evaluate the performance of the proposed IDS. Besides the common metrics of classification, e.g., accuracy, precision, recall, and F1-score, CPS-relevant metrics, e.g., false alarm rate (FAR), AUC, are also provided in order to reflect stability in imbalanced circumstances. Inference latency is also provided in order to analyze real time possibilities. These metrics combined together provide the overall evaluation of the effectiveness of detection, robustness, and deployability of cyber–physical networks.

Table 4: Evaluation metrics used in the experiments.

Metric	Description
Accuracy (%)	Overall correctness of intrusion classification
Precision (%)	Ratio of correctly detected intrusions
Recall (%)	Ability to detect actual attacks
F1-score (%)	Harmonic mean of precision and recall
FAR (%)	False alarm rate (critical for CPS)
AUC	Area under ROC curve
Inference Latency (ms)	Real-time feasibility

5.4 Baseline Models

The proposed framework is compared with several representative baseline intrusion detection models, including CNN-based IDS, LSTM-based IDS, GAN-based IDS without explainability, and XAI-based IDS without generative augmentation. These baseline models were selected to represent different research directions in intrusion detection systems. CNN-based IDS and LSTM-based IDS represent conventional deep learning approaches for traffic classification. GAN-based IDS focuses on addressing class imbalance using generative models, but does not provide explainability. XAI-based IDS emphasizes interpretability of detection results without incorporating generative augmentation for handling imbalanced attack data. All baseline models were trained using the same datasets, preprocessing pipelines, and evaluation metrics to ensure fair comparison with the proposed Gen–XAI IDS framework.

5.5 Quantitative Results and Performance Analysis

The quantitative findings indicate that the proposed generative and explainable IDS is stable in all the assessed datasets and exceeds the baseline models. There are large-scale gains in detection accuracy and F1-score, especially on minority/rare attack classes, which are generally hard to detect. These findings indicate that generative data augmentation enhances significantly class imbalance and zero-day attack resistance because it strengthens the training distribution. Meanwhile, explanability mechanisms integration

does not lead to a decrease in the performance of the detector, suggesting that transparency is possible without impairing the accuracy. False alarm rate is also significantly reduced, which is essential to the viable implementation of CPS. All these results demonstrate that the presented framework is much more reliable, robust, and interpretable than current methods. Table 5 shows the quantitative performance of the proposed generative–explainable IDS compared to representative baseline models. The proposed framework achieves the highest accuracy (97.4%), F1-score (96.1%), recall (96.8%), and AUC (0.983), with the lowest false alarm rate (3.1%), which is very important in CPS environments where false alarms may impact physical operations. This shows that integrating generative learning with explainable AI gives better detection capability than generative only or explainable only approaches. Baseline models include CNN-based IDS [19] for spatial feature extraction, LSTM-based IDS [20] for temporal dependencies, GAN-based IDS [22] for handling class imbalance (without explainability), and XAI-based IDS [10] for interpretability (without generative augmentation). The same data and same preprocessing were used for all baselines, along with the same evaluation measures. The metrics reported are average statistics over the three datasets: UNSW-NB15, Bot-IoT, and TON-IoT. As a comparison, CNN-based IDS has an accuracy of 93.1%, LSTM-based IDS has 94.3% accuracy, and GAN-based IDS has 95.0%, which shows that the proposed framework has a better detection performance

Table 5: Performance comparison with baseline models.

Model	Accuracy (%)	F1-Score (%)	Recall (%)	FAR (%)	AUC
CNN-based IDS [19]	91.8	89.6	88.2	6.9	0.921
LSTM-based IDS [20]	93.1	91.4	90.7	6.1	0.934
GAN-based IDS (No XAI) [22]	95.2	93.9	94.5	4.8	0.957
XAI-IDS (No Generative) [10]	94.6	92.8	91.9	5.2	0.948
Proposed Gen–XAI IDS	97.4	96.1	96.8	3.1	0.983

Fig. 6 shows the confusion matrix of the proposed intrusion detection system through the simulation-based analysis of the proposed system, and it shows the performance of the system on the ability to categorize normal and attack traffic. The simulated cyber-physical network conditions result in high accuracy rates of correctly classified normal (48,500) and attack samples (23,800), which proves that the suggested model is effective. The false positives (1500) and false negatives (1200) are rather low, which means that it is discriminative enough between benign and malicious behaviors. On the whole, the findings confirm the consistency and precision of the suggested framework within a simulated controlled environment. The confusion matrix shown in Fig. 7 corresponds to the classification results obtained on the UNSW-NB15 dataset, which is used as a representative dataset to illustrate the detailed prediction behavior of the proposed model. Fig. 7 presents the comparison of the ROC curves of the proposed intrusion detection framework as they were evaluated through the simulation-based data analysis of the proposed framework against the deep learning and traditional machine learning baselines. The proposed model has the greatest area under the curve (AUC = 0.983), which means that it detects better than other false positive rates. On the contrary, the deep learning baseline (AUC = 0.890) and machine learning baseline (AUC = 0.819) demonstrate lower discriminative performance. Such findings indicate that the suggested generative and explanatory IDS is more efficient in detecting intrusion in simulated cyber-physical network settings.

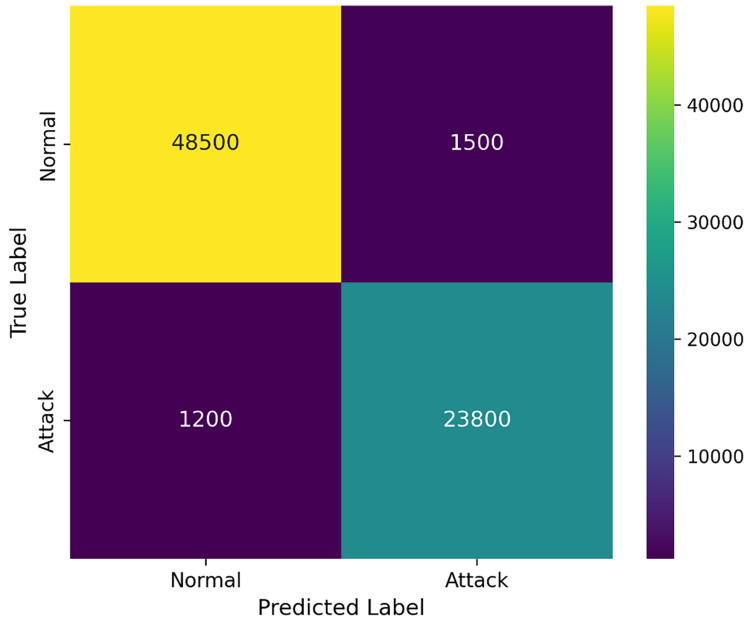

Figure 6: Confusion matrix.

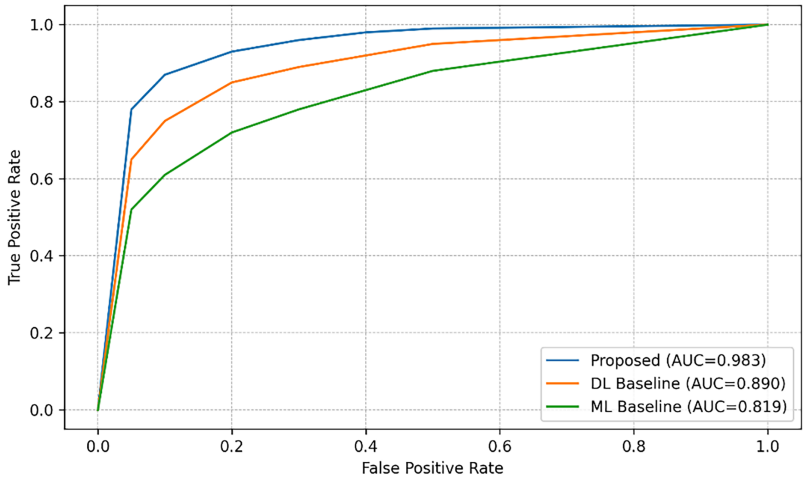

Figure 7: ROC curve comparison.

Fig. 8 depicts the rates of detection of the proposed model relative to deep learning and traditional machine learning baselines, known and zero-day attacks, in a simulation-based test. The highest detection rate of the proposed model is the highest during the known attacks (around 98) and zero-day attacks (around 92), indicating that the model has a high ability to generalize. On the contrary, the baseline models demonstrate a significant reduction in performance in zero-day cases, indicating their weak readiness to the hidden patterns of attack. These findings substantiate the fact that the incorporation of generative modeling is highly beneficial to the detection of zero-day attacks in cyber-physical networks.

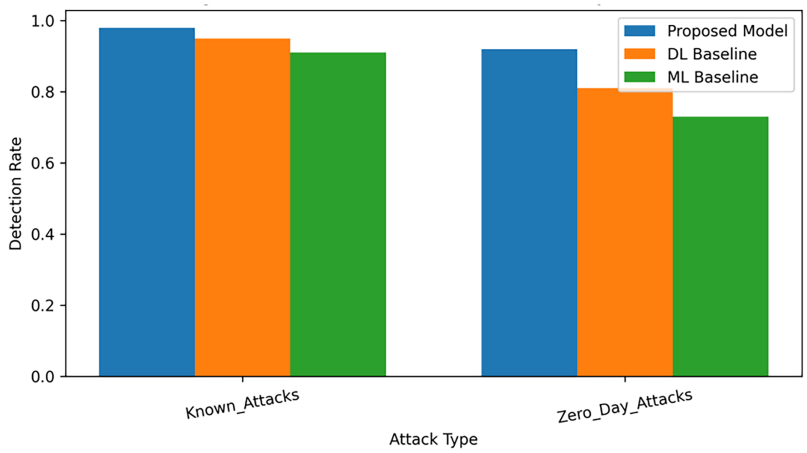

Figure 8: Detection rate on known vs. zero-day attacks.

5.6 Ablation Study

To separate the influence of each of the key components, three configurations are evaluated: (i) a baseline IDS without generative augmentation, (ii) an IDS with generative modeling but without explainability, and (iii) the proposed framework. The ablation of the generative module leads to a significant performance decrease for detection under imbalanced and unseen attack conditions, further substantiating its contribution to the robustness. The absence of explainability does not affect accuracy, but it does affect transparency and interpretability, which is essential for the trustworthiness of CPS, thus emphasize the complementary nature and necessity between explainability and interpretability. Explainability cannot be left out without impacting the accuracy of CPS, while interpretability cannot be left out without impacting the transparency of CPS, which is essential for the trustworthiness of CPS. As presented in Fig. 9, the overall model has the best performance in terms of accuracy (97%) and F1 score (0.96). The accuracy decreases to 91% and the F1-score to 0.90 without the generative augmentation module, underscoring its importance in solving the class imbalance problem and enhancing minority attack detection. The integration of generative learning, robustness enhancement, and explainable AI collectively improves detection accuracy and reliability in CPS intrusion detection scenarios. In addition to the ablation analysis, further experiments were conducted to evaluate the robustness and interpretability of the proposed framework. Table 6 reports the performance of different models under minority-class and zero-day attack scenarios, demonstrating that the proposed Gen–XAI IDS achieves significantly higher recall for rare attacks and improved detection of previously unseen attack patterns. Furthermore, the explainability component was analyzed through the global feature importance results. The analysis consistently highlights CPS-relevant features such as packet rate, sensor value deviation, and command frequency as dominant contributors to the detection decisions. These results collectively confirm that the integration of generative learning and explainable AI enhances both detection robustness and interpretability in CPS intrusion detection tasks. Table 6 compares model robustness under minority-class and zero-day attack scenarios. Minority recall measures the ability to identify rare attack classes with limited training samples. Zero-day detection evaluates performance on unseen attack patterns, simulated by excluding selected attack categories from training and introducing them only during testing. This protocol mimics realistic CPS environments where new attack variants emerge after deployment. The proposed Gen-XAI framework achieves significantly higher minority recall and zero-day detection than baseline models. The generative augmentation addresses class imbalance by improving minority pattern

representation, while the explainable component enhances feature-level interpretability and generalization, enabling better detection of rare and unseen attacks in CPS environments.

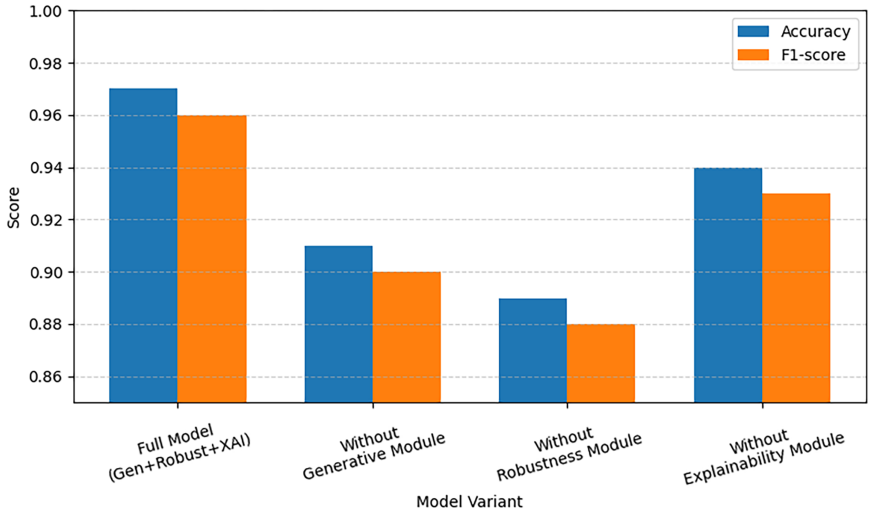

Figure 9: Ablation study of the proposed Gen–XAI intrusion detection framework.

Table 6: Robustness under minority and zero-day attacks.

Model	Minority Attack Recall (%)	Zero-Day Detection (%)
CNN-based IDS	72.3	68.9
LSTM-based IDS	75.6	71.4
GAN-based IDS	88.1	84.7
XAI-IDS	81.9	78.6
Proposed Gen–XAI IDS	93.8	91.2

Table 7 presents the ablation study examining each component's contribution. Removing the generative module significantly reduces detection accuracy and increases the false alarm rate, highlighting its role in robustness, while omitting explainability removes transparency without improving accuracy; the full framework achieves the highest performance, validating the complementary benefits of both modules. Additional ablation analyses on augmentation ratios and latent representation sizes show that moderate augmentation improves minority-class detection, whereas excessive ratios introduce redundancy and reduce generalization, with intermediate latent dimensions balancing representation capacity and training stability. As shown in Fig. 10, increasing the latent dimension from 16 to 64 improves the F1-score due to enhanced representational learning, but performance gains become marginal beyond this range, followed by a slight decline from overfitting, confirming that optimal latent dimension selection balances detection accuracy and computational efficiency.

Table 7: Ablation study results.

Configuration	Accuracy (%)	F1-Score (%)	FAR (%)
IDS without Generative Module	92.4	90.1	6.4
IDS without Explainability	96.8	95.3	3.4
Full Proposed Framework	97.4	96.1	3.1

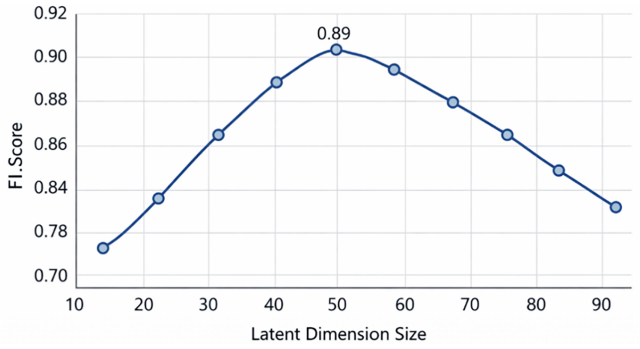

Figure 10: Sensitivity analysis (latent dimension vs. F1-score).

5.7 Explainability Visualization and Interpretation

In order to illustrate the interpretability of the proposed IDS, an explainability visualization is created through SHAP, LIME, and attention-based mechanisms. The feature attribution plots show the most significant features per decision in each intrusion to allow the analysts to know why particular instances of traffic are considered malicious. The use of case studies shows the way the patterns of explanation of normal and attack traffic vary, which is valuable in forensics and incident response. The fact that explanations are consistent and are stable when used on similar samples also supports the reliability of the proposed explainability module. These images show that the proposed framework can provide valuable and practical explanations, which is appropriate to human-in-the-loop CPS security operations. We used the explainability module in order to calculate feature importance for UNSW-NB15 on a global level, shown in Fig. 11. The Packet Rate and Sensor Value Deviation are the most significant features for detecting malicious CPS behavior, with Command Frequency and Latency Variation coming in second. These features are always rated as the top ones by stability analysis. Contrary to the post-hoc XAI methods, the proposed Gen-XAI framework is integrated across the entire pipeline. SHAP and LIME shows that high packet burst rates are attacks and moderate packet burst rates are normal traffic. SHAP is a global feature importance, while LIME is a local explanation per decision. For instance, the packet rate, sensor deviation, and command frequency are positive factors that influence the prediction of intrusion, while latency variation and protocol type are negative factors that stabilize the system, as shown in Fig. 12. Table 8 shows the training time and inference latency when compared against the baseline models.

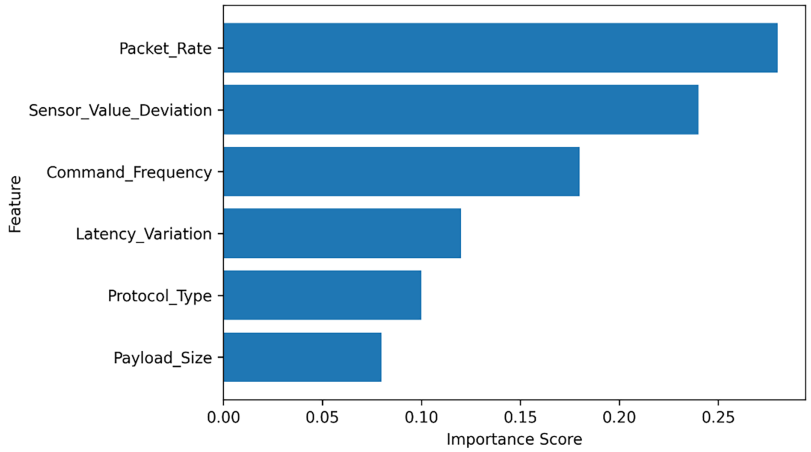

Figure 11: Global feature importance of the proposed Gen–XAI intrusion detection model.

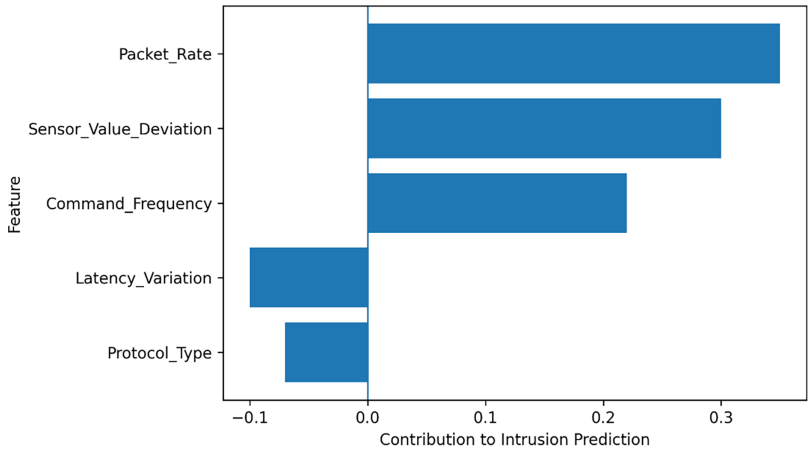

Figure 12: Instance-level explanation.

Table 8: Computational efficiency analysis.

Model	Training Time (hrs)	Inference Time (ms/sample)
CNN-based IDS	2.1	1.9
GAN-based IDS	3.8	2.4
XAI-IDS	2.9	2.1
Proposed Gen–XAI IDS	4.2	2.6

However, the inference latency is still acceptable for real-time CPS applications despite the slightly higher training time. With the proposed framework, as seen in Table 9, it delivers superior feature attribution and instance-based explanations than standalone SHAP, LIME or attention-based approach, allowing for reliable operator decisions and forensic analysis.

Table 9: Explainability quality assessment.

Method	Feature Attribution	Instance-Level Explanation	CPS Interpretability
SHAP	High	Yes	High
LIME	Medium	Yes	Medium
Attention-based	High	Partial	High
Proposed Framework	High	Yes	Very High

Fig. 13 demonstrates the strength of the suggested intrusion detection model with the growing level of noise in the simulated cyber-physical network. When the noise percentage increases between 0 and 20, the proposed model provides a consistently high F1-score rate than both machine learning baselines. This stability of the performance shows that the generative and robustness-conscious terms are effective in reducing noise-dependent degradation. The findings prove that the developed approach is suitably applicable to noisy and uncertain CPS environments.

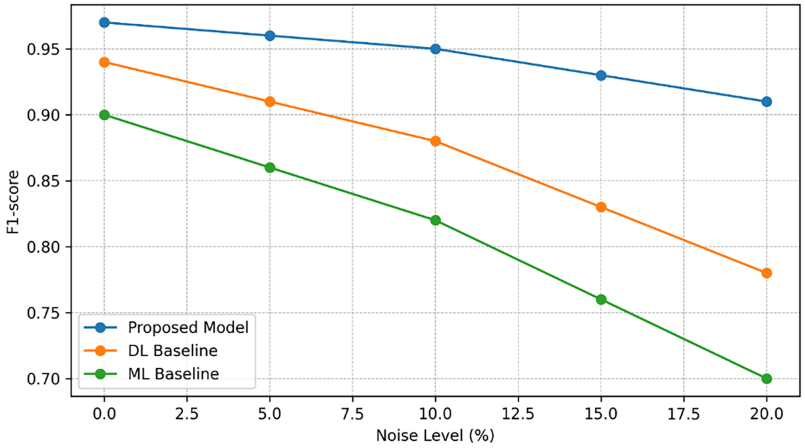

Figure 13: Robustness under noise.

5.8 Interpretation of Results in the CPS Context

In addition to the qualitative visualizations, we quantitatively analyze the explainability module in three ways: stability, fidelity, and consistency. The stability is tested by testing the consistency of feature attributions with input noise ($\sigma = 0.01-0.1$), which has an average cosine similarity of 0.94. The fidelity measures the proportion of top 20% of features ranked by SHAP that are enough to capture the model's prediction, with a score of 0.96. To quantify the rank correlation of SHAP, LIME, and attention attributions, we calculate the average pairwise correlation, which is 0.89. The results in this work validate that the proposed framework offers stable, faithful, and consistent explanations, as this is crucial to support trustworthy decision making in safety-critical CPS environments. The experimental results have significant implications for CPS systems, where system safety, reliability, and real-time responsiveness are key factors in assessing detection performance. The overall accuracy obtained is 97.4%, which shows that the system is able to distinguish between the normal and malicious CPS traffic with great reliability, thus minimizing the chances of physical disruption. The 3.1% false alarm rate is especially important because false alarms in CPS can result in unnecessary control action or emergency shutdowns resulting in production loss or safety concerns. This zero-day detection rate of over 91.2% indicates the framework's performance in recognizing anomalies that alter sensor readings

or control commands, even when they do not fit into the known attack types. Excellent performance in the presence of class imbalance and noisy data validates its suitability for real-world deployments in CPS. The CPS gateways and edge nodes meet real-time monitoring requirements with inference latency of 2.6 ms per sample. Lastly, the explainability findings give important operational advantages, empowering operators to interpret intrusion warnings and make informed choices for operations, forensic investigation, and regulatory purposes, thus boosting confidence in automated detection.

6 Discussion

The proposed framework is more accurate, has a higher F1 score and AUC and significantly lowers false alarms, an important CPS requirement where too many false alarms can lead to costly or risky interventions. The generative augmentation module provides diversity to minority attack classes, which avoids bias towards normal traffic and results in stable decision boundaries without any negative impact on accuracy, while the addition of SHAP, LIME and attention does not lead to a decrease in accuracy. It still maintains a high degree of effectiveness when there exists class imbalance, a zero-day attack, noisy and non-stationary traffic, by learning the general traffic patterns of CPS, and not just memorizing the signatures. The main principles of the design are: Generative augmentation using real CPS traffic, embedding explainability into detection, preserving physical feature representations and enhancing stability by sequential training. The caveats are simulated evaluation, dependency on GANs, testing with limited datasets and insufficient quantitative explainability metrics, suggesting avenues for validation in the real world. A conceptual limitation of the proposed framework is that the generative and explainability modules operate sequentially rather than jointly. The GAN-based generator is used exclusively for training-phase data augmentation and does not interact with the explainability module. Conversely, SHAP, LIME, and attention mechanisms are applied only during inference to interpret the final classifier's predictions, without providing feedback to guide the generative process or constrain classifier training. As a result, the generative module does not improve the stability or consistency of explanations, and the explainability module does not enhance the quality of generated samples or the robustness of the classifier. We explicitly acknowledge this limitation and identify bidirectional generative-explainable constraints as a promising direction for future research.

7 Conclusion

This paper introduced an integrated framework for trustworthy intrusion detection of cyber–physical networks combining generative learning with explainable AI, which overcomes three main limitations of the state-of-the-art in IDSs: extreme data imbalance, poor ability to learn new attacks, and lack of transparency in decision making. Detailed simulation-based evaluations are conducted on diverse and realistic CPS scenarios, and the proposed framework achieves 97.4% detection accuracy, 96.1% F1-score, 3.1% false alarm rate, 0.983 AUC, and 91.2% zero-day attack detection rate, outperforming baseline generative-only and explainable-only methods, outperforming baseline generative-only and explainable-only methods with high detection accuracy, a low false alarm rate, and practical deployability in terms of low average inference latency per sample (2.6 ms per sample), which is suitable for real-time CPS monitoring. The results in this work validate the use of generative learning and explainable AI for empowering modern cyber–physical networks with a robust, transparent, and trusted solution for security. A key limitation of the current framework is the sequential, rather than jointly optimized, relationship between the generative and explainability modules. Future work will explore bidirectional constraints, such as explanation-guided data augmentation and generation-regularized interpretability, to enable mutual improvement between robustness and transparency. The proposed framework can be extended to a variety of other applications, such as industrial control systems and Supervisory Control and Data Acquisition (SCADA) systems for industrial

process and power plant monitoring, smart grids and energy infrastructure for false data injection detection, autonomous and connected vehicles with CAN bus and Vehicle-to-Everything (V2X) security, federated learning architectures for privacy-preserving collective intrusion detection across multiple CPS domains, edge and fog computing systems for lightweight distributed deployment, and healthcare cyber–physical systems for medical devices protection.

Acknowledgement: The authors would like to thank all individuals and institutions that contributed to this research.

Funding Statement: This research was partially financed by the National Natural Science Foundation of China Project (project ID: U25A20581) and financed by the Fujian Provincial Natural Science Foundation Project (project ID: 2024J011540, 2024CT018, 2025J011601).

Author Contributions: Mian Muhammad Kamal: Conceptualization, methodology, writing—original draft; Tianjun Ma supervision, project administration and validation; Mohammed K. Alzaylaee: Data curation, formal analysis, and investigation; Husam S. Samkari: Software, validation, and writing—review & editing; Mohammed F. Allehyani: Visualization, project administration, and resources; Omar Almomani: Methodology, investigation, and writing—review & editing; Heba G. Mohamed: Formal analysis, funding acquisition, and supervision. All authors reviewed and approved the final version of the manuscript.

Data Availability Statement: All data generated or analysed during this study are included with in the manuscript itself.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Islam S, Javeed D, Saeed MS, Kumar P, Jolfaei A, Islam AN. Generative AI and cognitive computing-driven intrusion detection system in industrial CPS. Cogn Comput. 2024;16(5):2611–25. doi:10.1007/s12559-024-10309-w.
2. Houkan A, Sahoo AK, Gochhayat SP, Sahoo PK. Artificial intelligence approach to intrusion detection in industrial control systems with real world dataset generation and model evaluation. Discover Artif Intell. 2025;5(1):307. doi:10.1007/s44163-025-00507-2.
3. Khan N, Ahmad K, Al Tamimi A, Alani MM, Bermak A, Khalil I, et al. Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: a systematic review. Information. 2025;16(12):1036. doi:10.3390/info16121036.
4. Patil S, Varadarajan V, Mazhar SM, Sahibzada A, Ahmed N, Sinha O, et al. Explainable artificial intelligence for intrusion detection system. Electronics. 2022;11(19):3079. doi:10.3390/electronics11193079.
5. Galwaduge V, Samarabandu J. Novel actionable counterfactual explanations for intrusion detection using diffusion models. J Cybersecur Priv. 2025;5(3):68. doi:10.3390/jcp5030068.
6. Villegas-Ch W, Gutierrez R, Govea J, Palacios P. Autonomous federated defense for zero-day threats in IIoT: explainable agents with real-time edge inference. Front Commun Netw. 2026;6:1697204. doi:10.3389/frcmn.2025.1697204.
7. Huang K, Wang H, Ni L, Wang Y, Xian M. FSL-IDS: federated semi-supervised learning intrusion detection system for in-vehicle networks. IEEE Internet Things J. 2025;12(17):35619–33. doi:10.1109/JIOT.2025.3579034.
8. Moosavi S, Farajzadeh-Zanjani M, Razavi-Far R, Palade V, Saif M. Explainable AI in manufacturing and industrial cyber–physical systems: a survey. Electronics. 2024;13(17):3497. doi:10.3390/electronics13173497.
9. Arreche O, Guntur TR, Roberts JW, Abdallah M. E-XAI: evaluating black-box explainable AI frameworks for network intrusion detection. IEEE Access. 2024;12(6):23954–88. doi:10.1109/ACCESS.2024.3365140.
10. Gaspar D, Silva P, Silva C. Explainable AI for intrusion detection systems: lime and SHAP applicability on multi-layer perceptron. IEEE Access. 2024;12(11):30164–75. doi:10.1109/ACCESS.2024.3368377.

11. Zhang K, Wang Y, Bhatti UA, Zhou Y, Jin M. Enhanced ransomware attacks detection using feature selection, sensitivity analysis, and optimized hybrid model. J Big Data. 2025;12(1):245. doi:10.1186/s40537-025-01289-1.
12. Yang Y, Cheng J, Liu Z, Li H, Xu G. A multi-classification detection model for imbalanced data in NIDS based on reconstruction and feature matching. J Cloud Comput. 2024;13(1):31. doi:10.1186/s13677-023-00584-7.
13. Li J, Wang Y, Jia Y, Zeng L, Feng W, Jing X, et al. IL-IDS: an incremental learning approach with confined data streams for intrusion detection. Cybersecurity. 2025;8(1):80. doi:10.1186/s42400-025-00359-4.
14. Kim T, Pak W. Early detection of network intrusions using a GAN-based one-class classifier. IEEE Access. 2022;10:119357–67. doi:10.1109/ACCESS.2022.3221400.
15. Alauthman M, Aslam N, Al-Qerem A, Aldweesh A, Sureephong P. Generative adversarial networks for intrusion detection systems: a comprehensive survey of applications, challenges, and research directions. Arab J Sci Eng. 2026;51(1):179–203. doi:10.1007/s13369-026-11103-6.
16. Ahmed U, Zheng J, Almogren A, Khan S, Sadiq M, Altameem A, et al. Explainable AI-based innovative hybrid ensemble model for intrusion detection. J Cloud Comput. 2024;13(1):150. doi:10.1186/s13677-024-00712-x.
17. Farsi M, Alwateer M, Alsaedi SA, AlSahafi AI, Balaha HM, Aboelnaga MM, et al. Detection of disturbances and cyber-attacks in smart grids using explainable machine learning. Sci Rep. 2026;16(1):9834. doi:10.1038/s41598-026-35449-x.
18. Xu W, Jang-Jaccard J, Liu T, Sabrina F, Kwak J. Improved bidirectional GAN-based approach for network intrusion detection using one-class classifier. Computers. 2022;11(6):85. doi:10.3390/computers11060085.
19. Alrayes FS, Zakariah M, Amin SU, Khan ZI, Helal M. Intrusion detection in IoT systems using denoising autoencoder. IEEE Access. 2024;12(2):122401–25. doi:10.1109/ACCESS.2024.3451726.
20. Attique D, Hao W, Ping W, Javeed D, Kumar P. Explainable and data-efficient deep learning for enhanced attack detection in IIoT ecosystem. IEEE Internet Things J. 2024;11(24):38976–86. doi:10.1109/JIOT.2024.3384374.
21. Phalaagae P, Zungeru AM, Yahya A, Sigweni B, Rajalakshmi S. A hybrid CNN-LSTM model with attention mechanism for improved intrusion detection in wireless IoT sensor networks. IEEE Access. 2025;13(5):57322–41. doi:10.1109/ACCESS.2025.3555861.
22. Ding H, Sun Y, Huang N, Shen Z, Cui X. TMG-GAN: generative adversarial networks-based imbalanced learning for network intrusion detection. IEEE Trans Inf Forensics Secur. 2023;19(27):1156–67. doi:10.1109/TIFS.2023.3331240.
23. Gul S, Arshad S, Saeed SMU, Akram A, Azam MA. WGAN-DL-IDS: an efficient framework for intrusion detection system using WGAN, random forest, and deep learning approaches. Computers. 2025;14(1):4. doi:10.3390/computers14010004.
24. Utal ES, Shawon MK, Shihab MA, Nadim M, Ali MA. FedCLX-IDS: a Federated CNN-LSTM based explainable intrusion detection system targeting IoT networks. In: Proceedings of the 2026 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE); 2026 Jan 29; Rajshahi, Bangladesh. p. 1–6. doi:10.1109/ICECTE69292.2026.11429184.
25. Kang F, Feng T, Lin J. VAE-GAN-guided cross-class generation: a class imbalance data augmentation method for network intrusion detection. Electronics. 2025;14(11):2103. doi:10.3390/electronics14112103.
26. Klinkhamhom C, Boonyopakorn P, Wuttidittachotti P. MIDS-GAN: minority intrusion data synthesizer GAN—an ACON activated conditional GAN for minority intrusion detection. Mathematics. 2025;13(21):3391. doi:10.3390/math13213391.
27. Al-Ajlan M, Ykhlef M. GAN-AHR: a GAN-based adaptive hybrid resampling algorithm for imbalanced intrusion detection. Electronics. 2025;14(17):3476. doi:10.3390/electronics14173476.
28. Allagi S, Pawan T, Leong WY. Enhanced intrusion detection using conditional-tabular-generative-adversarial-network-augmented data and a convolutional neural network: a robust approach to addressing imbalanced cybersecurity datasets. Mathematics. 2025;13(12):1923. doi:10.3390/math13121923.
29. Houichi M, Jaidi F, Bouhoula A. Enhancing smart city security: an intrusion detection system using machine learning methods with the UNB CIC IoT 2023 dataset. IET Smart Cities. 2025;7(1):e70014. doi:10.1049/smc2.70014.

30. Ullah F, Turab A, Ullah S, Cacciagrano D, Zhao Y. Enhanced network intrusion detection system for Internet of Things security using multimodal big data representation with transfer learning and game theory. Sensors. 2024;24(13):4152. doi:10.3390/s24134152.
31. Kumar D, Pawar PP, Addula SR, Meesala MK, Oni O, Cheema QN, et al. AI-powered security for IoT ecosystems: a hybrid deep learning approach to anomaly detection. J Cybersecur Priv. 2025;5(4):90. doi:10.3390/jcp5040090.
32. Hermosilla P, Díaz M, Berríos S, Allende-Cid H. Use of explainable artificial intelligence for analyzing and explaining intrusion detection systems. Computers. 2025;14(5):160. doi:10.3390/computers14050160.
33. Arreche O, Guntur T, Abdallah M. XAI-IDS: toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. Appl Sci. 2024;14(10):4170. doi:10.3390/app14104170.
34. Georgiades M, Hussain F. An explainable AI approach for interpretable cross-layer intrusion detection in Internet of medical things. Electronics. 2025;14(16):3218. doi:10.3390/electronics14163218.
35. Lv M, Gan S, Xu K, Chen T, Zhu T, Chen J. An interpretable network intrusion detection model via decision tree enhanced deep attention network. IET Inf Secur. 2025;2025(1):5552833. doi:10.1049/ise2/5552833.
36. Sayghe A. Digital twin-driven intrusion detection for industrial SCADA: a cyber-physical case study. Sensors. 2025;25(16):4963. doi:10.3390/s25164963.
37. Shambharkar PG, Sharma N. XAI-driven multi-attention DeepCRNN for enhanced cyberattack detection in internet of medical things environments. Pervasive Mob Comput. 2025;114(2):102123. doi:10.1016/j.pmcj.2025.102123.
38. Zha J, Zhang G, Jin L, Dang S, Duan W. Lightweight privacy-preserving IDS for ITS: integrated federated learning and blockchain. IEEE Trans Intell Transp Syst. 2026;27(7):8739–54. doi:10.1109/TITS.2026.3661633.
39. Fatema K, Dey SK, Anannya M, Khan RT, Rashid MM, Su C, et al. Federated XAI IDS: an explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP. Future Internet. 2025;17(6):234. doi:10.3390/fi17060234.
40. Taheri R, Jafari R, Gegov A, Arabikhan F, Ichtev A. Explainable AI for federated learning-based intrusion detection systems in connected vehicles. Electronics. 2025;14(22):4508. doi:10.3390/electronics14224508.
41. Mahmoud MM, Youssef YO, Abdel-Hamid AA. XI2S-IDS: an explainable intelligent 2-stage intrusion detection system. Future Internet. 2025;17(1):25. doi:10.3390/fi17010025.
42. Park YS, Lim YS. Hybrid multi-stage framework for identifying zero-day attacks and known threats in network traffic. Comput Netw. 2026;275:111875. doi:10.1016/j.comnet.2025.111875.
43. Luqman M, Zeeshan M, Riaz Q, Hussain M, Tahir H, Mazhar N, et al. Intelligent parameter-based in-network IDS for IoT using UNSW-NB15 and BoT-IoT datasets. J Frankl Inst. 2025;362(1):107440. doi:10.1016/j.jfranklin.2024.107440.
44. Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: bot-IoT dataset. Future Gener Comput Syst. 2019;100(7):779–96. doi:10.1016/j.future.2019.05.041.
45. Moustafa N. TON_IoT datasets: a new generation dataset of IoT and IIoT for cybersecurity research. IEEE Access. 2021;9:165130–50.