ISSN: 2617-5517 (issn.org)
Al-Farabi Journal of Engineering Sciences
https://iasj.rdd.edu.iq/journals/journal/view/97
مجلة الفارابي للعلوم الهندسية تصدرها جامعة الفارابي

Smart Unsupervised Cyber-attack Path Inference of the Industrial Internet of Things Based on Temporal Anomaly Dynamics

Hadeel M. Saleh
Continuing Education Center, University of Anbar, Al-Anbar, Iraq
Corresponding E-mail: Haddeel.mohammed@uoanbar.edu.iq

Abstract

The prevalence of Industrial Internet of Things (IIoT) systems has significantly advanced automation processes in industrial environments, simultaneously exposing critical infrastructure to high-level cyber-attacks. Recent attacks, such as zero-day and multi-stage attacks, are particularly challenging due to their dynamic nature and lack of previous attack signatures. Traditional industrial intrusion detection systems (IIDS) mostly rely on point-based or fixed-time interval anomaly detection, which limits their ability to predict temporal attack evolution and provide early alerts. This work introduces an unsupervised, explainable temporal inference framework that models cyber-attacks based on behavior evolution, rather than categorizing them. Unlabeled raw data is directly inferred into normal system dynamics using a deep temporal model. Deviations are construed as anomalies in behavior. These abnormalities are then narrowed down to capture long-term, system-wide changes while filtering out short-term operational changes. A probabilistic time model of inference is then implemented to infer latent states and transitions in the system. This allows us to discern and interpret multi-stage attack progression without labeling attack values. The proposed model is tested on popular industrial and IIoT benchmark datasets such as SWaT, WADI, and ToN-IoT in the presence of the most severe zero-day conditions. Empirical findings support detection rates of 98.4%, 97.2%, and 96.9% on SWaT, WADI, and ToN-IoT, respectively, with false alarms at less than 2.5%. Additionally, the framework achieves close-to-real-time detection with an average detection delay in the sub-second range.

Keywords: IIoT Security, Zero-Day Attacks, Explainable AI (XAI), LSTM-Autoencoder, HMM.

1. Introduction

The Industrial Internet of Things (IIoT) has become a key enabler of next-generation industrial automation to enable intelligent monitoring, adaptive control and data-based optimization in complex cyber–physical systems. High-level decision making in all key sectors of industry, such as manufacturing, water treatment, energy systems, and biological processes, is supported by IIoT architectures which connect sensors, actuators, programmable logic controllers (PLCs), and supervisory control systems through industrial communication networks [1,2]. However, the dynamics of increased interconnectivity and heterogeneity of IIoT spaces significantly increase the attack surface of critical infrastructures, thus introducing complex security challenges.

Unlike traditional IT-related systems, IIoT systems are strongly intertwined with physical processes and are characterized by high real-time, reliability and safety demands. Therefore, any cyber-attack on IIoT systems may be directly converted to the physical outcomes, such as the damages to equipment, loss of production, environmental pollution, and human safety hazards [3]. These environments inherently place a strong restriction on traditional security tools such as signature-based intrusion detection tools and rule-based firewalls since they are based on predefined attack signatures and a rule set that is fixed. This dependency significantly hinders their capacity to adapt to evolving system behaviors changing system behavior and to effectively address complex and previously unseen threats [4].

Zero-day cyber-attacks pose the ultimate threat to industrial Internet of Things infrastructures because such attacks exploit previously unknown vulnerabilities and have no recorded instances or case studies. In the context of industrial control, these compromises usually take the form of small deviations in sensor or actuator data or control algorithms instead of leading to immediate disruptive events[5]. These attacks typically occur in various sequential steps, which allows attackers to remain hidden while reducing system functionality over long periods of time[6].

In an attempt to address them, machine-learning and deep-learning-based security solutions began receiving more and more attention in industrial IoT. Temporal process models, such as Long Short-Term Memory (LSTM) networks, are instrumental in capturing non-linear temporal dependencies in industrial data[7,8].

However, most of current techniques rely on the paradigm of supervised learning which requires the large volumes of labeled attack examples that are often unavailable or infeasible to acquire in a low-frequency, safety-critical zero-day attack environment [9]. Besides, many deep-learning-based intrusion detection systems are runtime systems that are opaque and

are often characterized as "black-box" models, lacking the transparency required for safety-critical industrial environments[10].

Recent research has, in turn, become more limited to unsupervised methods of anomaly detection, such as reconstruction-based methods, isolation-based schemes, and probabilistic time warping methods, so as to be less reliant on labeled data, and able to be sensitive to zero-day attacks [11,13]. Although the strategies increase the functionality of an anomaly detector, they often assume anomalies are damage events or one-dimensional occurrences. This tends to create large rates of false-alarms and limits the extrapolation of coordinated or multi-stage attack behaviors that build up over time. probabilistic temporal models, e.g. Hidden Markov Models (HMMs), offer an interpretable and principled model of the attack phases, but on their own, they do not have enough informative behavioral evidence to be useful in reliably characterizing complex processes of attack progression [14,15].
This paper critically analyzes the existing constraints of existing IIoT intrusion detection systems and suggests a self-organizing, explanatory and unified approach that will reinvent IIoT security as unsupervised inference about attack sequence and not as a traditional binary detection of intrusion issue. Instead of viewing anomaly detection as a definitive decision, the offered paradigm views anomalous dynamics as in-between behavioral indicators, thus enabling the reasoning about how cyber-attacks can develop over time. Normal system dynamics is modeled through an LSTM Autoencoder model in order to learn deep temporal behavior, but anomalous deviations are refined through isolation-based analysis. The polished anomaly dynamics is thereafter entered into a probabilistic temporal model which deduces hidden attack periods and state shifts. Through its combination of sensor level reconstruction analysis with attack state inference over time, the framework bestows intrinsic interpretability thus promoting readable and reliable security reasoning both in industrial cyber-security and industrial-biological space.

Novelty and Contributions

1. Revising IIoT intrusion-detection as a time-series attack-progression inference problem, and thus leaving binary anomaly detection.
2. this work proposes a single and unsupervised detection system that incorporates LSTM Autoencoder to model temporal behavior, Isolation Forests to refine anomalies, and Hidden Markov Models to infer an attack action in many stages.
3. Providing intrinsic explainability through sensor- based attribution of anomalies and any other attack state, thus improving transparency and trust on safety-critical industrial systems.

2. Related Work

Due to the significant increase in cybercriminal activities in recent years, studies on the security of the Industrial Internet of Things (IIoT) have been extensively developed. This field is heterogeneous, and cybercrime is becoming more prevalent.
In 2022, Tharewal et al. present deep reinforcement learning was used for the first time in a study with major implications for industrial intrusion detection systems. Tharewal et al. used the PPO2 algorithm, enhanced by the LightGBM feature selection process. Although the resulting system demonstrated remarkable detection accuracy, it could not offer the necessary interpretability and transparency for sensitive industrial applications [16].
Similarly, Khan et al introduced an interpretable deep-learning model by combining an Autoencoder with a convolutional and recurrent network and a dual sliding-window scheme that optimizes malicious pattern recognition.
This model offered better detection and interpretable decisions, but its centralized architecture was not easily distributable and would not work well in more complicated attack scenarios [17].
Verma et al. introduced a framework called Zero-Day Guardian based on federated learning and two parallel Autoencoder to detect zero-day attacks in 5G- supervised IIoT instances. Though this model showed positive improvements in privacy maintenance and accountability in detection, interpretability was not a key design aspect [18].
In 2024, the research world underwent a paradigm shift towards using interpretive analyses in conjunction with deep learning models. Shoukat et al. proposed an open intrusion detection system based on LSTM-AE and AGRU. They used SHAP to explain the factors that determine an attack classification and deployed the system in an SDN-based structure. However, the system poorly handled zero-day attacks and temporal drift [19]. In the same year, Hamouda et al. introduced a federated learning-based conditional generative adversarial network framework that combats data imbalance and improves detection of unknown attacks. However, it does not focus on interpretability or adaptive defensive decision-making [20]. Zhao et al. suggested an ongoing federated learning framework that reduces catastrophic forgetting caused by dynamically changing IIoT settings. This framework achieves considerable performance benefits when adapting to complex threats. However, their proposal lacked stipulations regarding transparency and interpretation of decisions [21].

In 2025, Rawajbeh et al. addressed building an adaptive, real-time, interpretable intrusion detection system in edge devices using online learning models with SHAP to minimize latencyThe system did not specifically deal with the issue of countering zero-day attacks [22]. In the same manner, Zhai et al. proposed the multi-scale attention model to provide the amplification of the features representation in IIoT attacks and ensured high precision but the model was considered opaque, as it does not include a clear interpretive mechanism [23]. Shen et al. suggested a flexible zero-day defense model which is built on the Stackelberg game theory and the multi-agent reinforcement learning with a misleading strategy, thus achieving high resource efficiency, but interpretability was not the main goal of this model [24]. Lastly, Alatawi proposed the concept of SAFELIoT including safe federated learning [25].

Despite the achievements made in the area of the Industrial Internet of Things (IIoT) intrusion detection, the available options have serious flaws when compared to the highly realistic, multi-stage, and zero-day attacks. A significant percentage of deep-learning paradigms is based on supervised (or semi-supervised) learning, which means that they are limited to only the scenario when labeled attack data is scarcely available or unavailable. Even though federated and continual learning strategies provide more contextual flexibility, they often decrease transparency and architectural complexity, thus making it more
difficult to implement them in resource-constrained industrial contexts.
Besides, explainability is frequently considered a supplementary, after-hoc visual method, instead of providing an inherent part of the process of detection, which undermines the trust of the operators in the neediness of safety-critical contexts. Probabilistic methods, such as Hidden Markov Models, are sporadically used for temporal analysis. However, in the absence of deep temporal and feature representations, they have low expressive power. These imperfections lead to an urgent need for a coherent, self-educational, and self-explanatory design that can detect zero-day and multi-phase attacks while taking rigid industrial limitations into account.
Table 1 will provide a summary table of the studied surveys. This table outlines the research gaps, objectives, research methodology, as well as research contributions of the research.

Table 1. Summary of Previous Studies

Ref	Author (Year)	Dataset	Method	Key Contribution	Limitation
[16]	Tharewal et al. (2022)	Public IIoT	LightGBM + PPO2	High-precision DRL-based IDS	No explainability
[17]	Khan et al. (2022)	IIoT Traffic	AE + CNN + RNN	Deep attack detection	Centralized, non-federated
[18]	Verma et al. (2024)	X-IIoT-ID	Dual AE + FL + OCSVM	Zero-day detection with FL	Weak XAI
[19]	Shoukat et al. (2024)	N-Bat-IoT, CIC-IDS2017	LSTM-AE + AGRU + SHAP	Interpretable IDS	No zero-day support
[20]	Hamouda et al. (2024)	Industrial	GAN + FL	Imbalance handling	No explanation
[21]	Zhao et al. (2024)	Energy IIoT	Contrastive FL	Reduced forgetting	No XAI
[22]	Rawajbeh et al. (2025)	ToN-IoT, Bot-IoT	Online Learning + SHAP	Real-time edge IDS	Limited zero-day defense
[23]	Zhai et al. (2025)	Public	Multi-scale Attention	Improved accuracy	Black-box model
[24]	Shen et al. (2025)	ToN-IoT, BoT-IoT	Stackelberg + MAD3PG	Deceptive zero-day defense	No explanation
[25]	Alatawi (2025)	SKAB	FL + SHAP + Attention	Safe interpretable FL	High resource demand

In order to present a complex overview of the all-temporal and methodological development of previous studies in the area of industrial IoT security, Figure (1) represents a map of the research that has condensed the current research trends and indicates what there is missing.Figure (1) illustrating the evolution of previous studies in industrial IoT security in terms of methodologies used, areas of focus, and research gaps during the period (2022–2025).

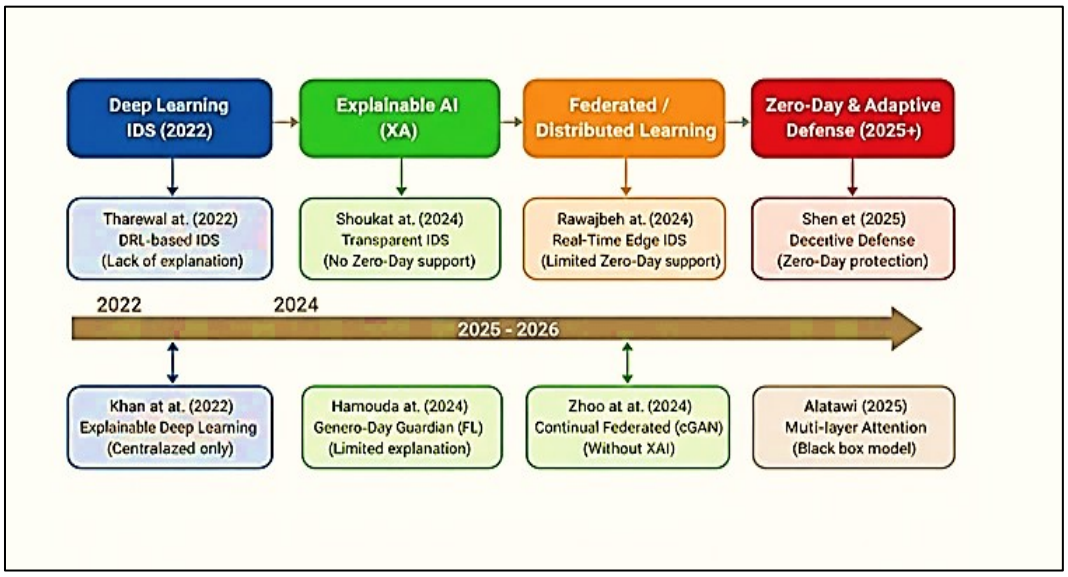

Figure 1. Timeline of Industrial IoT security research trends (2022–2025)

3. Research Gap :

To define the research gap detected in the context of Industrial Internet of Things (IIoT) security, Figure 2 presents a conceptual overview of the key research dimensions and gaps in their integration.

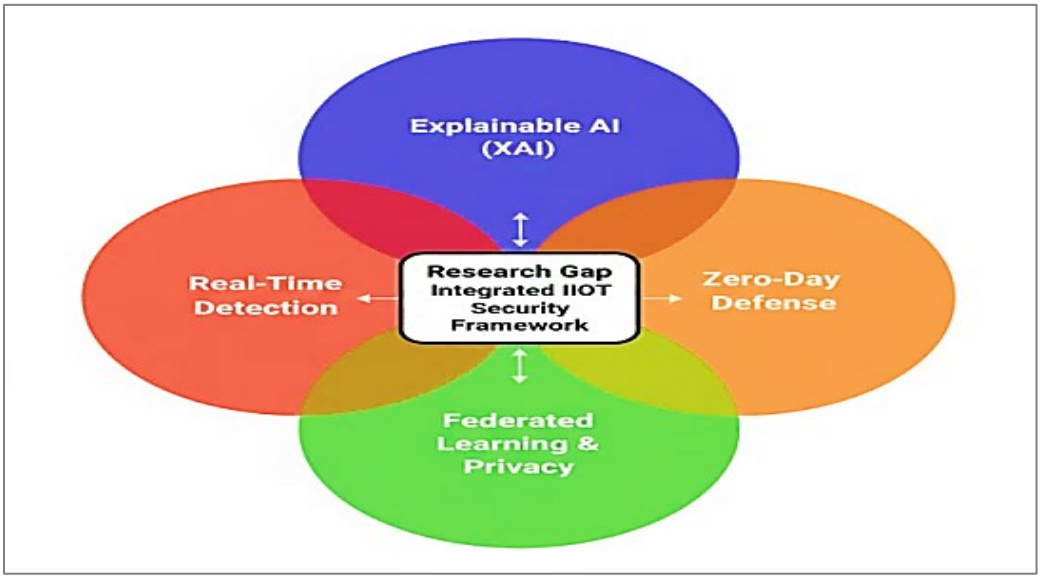

Figure 2. Research gap

Figure 2 identifies a paradigm shift in the security of the Industrial Internet of Things (IIoT) that predetermines the inclusion of explainable artificial intelligence, real-time-detecting mechanisms, privacy-sensitive federated learning, and adaptive zero-day defense measures. The lack of an unified framework that would integrate these dimensions highlights the urgency of the holistic, practical IIoT security product capable of dealing with the complex threat vectors at the same time.

The IIoT system under consideration follows a hierarchical multi-layer, and the architecture consists of field, control, communication, and supervision layers. Such hierarchical structure will allow tracing the cyber-physical processes entirely and encourage solid anomaly discovery policies to be implemented in order to identify the possible intrusion early.

4.2 Threat Model

This study assumes that the adversary has high technical expertise in industrial control systems and communication infrastructure. The attackers' goal is to control the system's actions or physical processes undetected for an extended period. Considered threat scenarios include unauthorized access, sensor data injection/manipulation, replay attacks, command injection, and insider threats that exploit weak authentication systems.

The point of such an agent is to interfere with working mechanics or manipulate physical processes without being detected during a long time. Possible malicious activities are: (i) attack on the IIoT device or network infrastructure; (ii) unauthorized attack; (iii) sensor telemetry and control directives injection, alteration, or replaying; (iv) insider attack or the use of weak authentication. We analyze such threat situations as false or spoofed data injection, replay attacks, and unauthorized command injection. "In particular, this study addresses zero-day attacks characterized by subtle deviations that evade signature-based methods. Besides, multi-phase and coordinated attacks are evaluated and happen in stages over multi-layers of the system. This analysis does not include purely physical attacks, including denial-of-service attacks and physical sabotage. These characteristics of threats make it clear why an intrusion detection framework should be able to facilitate unsupervised or self-learning detection, temporal modeling to model the multi-stage progression and solutions whose decision output can be comprehended within any safety-critical industrial environment.

4.3 Assumptions and Detection Scope

The model assumes that there is a state of normative period where models of self-learning can be trained. This time range does not require the existence of labeled attack data, as well as does not assume the existence of previously known attack signatures.

The operational features of the system include passive-monitoring abilities and are designed in such a way that they do not interfere with already existing control logic. Anomaly detection is limited to cyber-physical anomalies, which may be deduced based on sensor measurements and network based properties.

Accordingly, the system can produce interpretable alerts, hence enabling the building of the situational awareness, and, consequently, allowing this to be followed with a quick reaction.

4.4 Proposed Framework and Methodology

This section presents a unified, autonomous, and self-learning intrusion detection framework. That is designed to identify the presence of a zero-day attack or a multi-stage cyber-attack in industrial Internet of Things (IIoT) systems. It is a synthesized architecture, generating a rich temporal behavior model and automatically isolating anomalies and using probabilistic inference at each successive stage of an attack, and thus forms a unified end to end pipeline. Through autonomous learning of normative behavior, the system identifies anomalies that represent malicious behavior and generates clear and security-related decisions which can be implemented in industry as is shown in Figure 3.

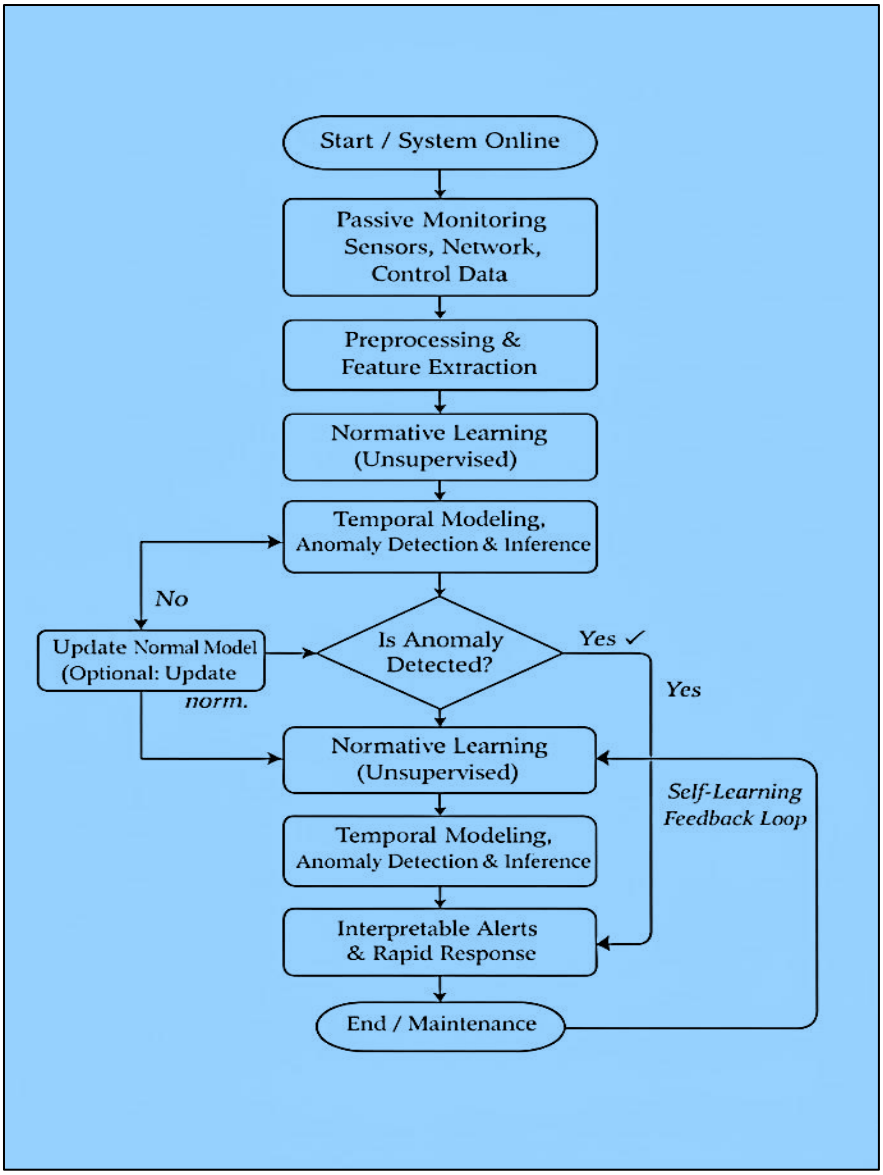

Figure 3. The design of the proposed explainable self-learning IIoT IDS is illustrated.

The above figure illustrates the sequential workflow. First, cyber-physical streams are acquired and preprocessed. Then, normal temporal behavior is modeled using an LSTM Autoencoder. Next, anomaly evidence is refined via isolation forest. Then, multi-stage progression is inferred using a hidden Markov model. Finally, the fusion and explanation modules produce the final decisions.

4.4.1 Framework Overview

The presented framework has been designed as one pipeline with four functional layers (1) Data Acquisition and Preprocessing, (2) Deep Temporal Behavior Modeling, (3) Hybrid Anomaly and Multi-Stage Detection, and (4) Explainable Decision-Making. The multilayered design will have every step working in harmony, starting with the raw data ingestion and all the way to being interpretable.

This is then followed by the temporal behavior modeling layer which tries to reflect temporal patterns and sequential relationships within the system behavior which allows normal and abnormal behavior to be well characterized. Lastly, there is the hybrid anomaly detecting and multi-stage detecting layer that is based on the combination of multiple detection processes that work in stages in order to enhance the accuracy of the

detection and minimize false alarms. Lastly, the interpretable decision-making layer generates final detection decisions and has clear explanations that show the reasoning of the system in making the decision on the decision-making. In contrast to the traditional intrusion detection systems (IDS), where the detection mechanism and the results interpretation phase

are usually separated, the proposed framework incorporates the aspect of interpretability into the detection logic itself. This helps in improving transparency, decision consistency and operational confidence among those who use the system especially in the critical and complex environments.

Figure 4. The integrated layered framework of the intrusion detection frame with readability.

The functions of each layer are different but complementary to each other in a single detection pipeline.

4.4.2 Data Acquisition and Preprocessing Layer

The former layer constantly gathers multivariate time-series information involving sensors, actuators and industrial communication traffic. All these streams are used to reflect the cyber physical condition of the plant which is the process dynamics, control actions as well as network level behavior.

Since IIoT telemetry is both heterogeneous and noisy, preprocessing is used to improve the quality of data and stabilize downstream learning. This involves management of missing values, noise elimination, time synchronization of the streams and normalization of features. These measures maintain time dependencies as well as reducing scale bias and sampling biases, leading to increased detection robustness and spurious alarms.

4.4.3 Temporal Behavior Modeling Using LSTM Autoencoder

In the training stage, the Autoencoder will be trained to recreate input sequences by minimizing the reconstruction error. During the deployment stage, abnormal behavior leads to more reconstruction errors, which are one of the key indicators of abnormal activities. This allows the proposed framework to deal with zero-day attacks that were not recorded during training without using a set of labeled attacks or predefined attack signatures.

The LSTM architecture facilitates long-term dependency modeling through three core gating mechanisms (f_t), the input gate (I_t), and the output gate (o_t). The following are the definitions of the forget and input gates equation (1) and (2):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad (1)$$

$$i_t = \sigma\,(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad (2)$$

where xt represents the input at time step t, ht -1 is the last hidden state., Wf and Wi are the weight matrices, bf and bi are the bias terms, and σ(•) represents the sigmoid activation function.

The reconstruction error et for an input sequence of length T is computed using the Mean Squared Error (MSE) as defined in Equation (3):

$$et = \frac{1}{T} \sum_{i=1}^{1} (\hat{x}_i - x_i)^2 \qquad (3)$$

where xi represents the original input value and xˆi denotes the reconstructed output produced by the LSTM-AE.

4.4.4 Unsupervised Anomaly Detection Using Isolation Forest

Although reconstruction errors are effective to capture deviation of learned normal behavior, not every such deviation would be cyber-attacks. The proposed framework incorporates the unsupervised anomaly detection mechanism known as an Isolation Forest (IF) model to differentiate malicious anomalies with benign operational variations.

The Isolation Forest examines the features of anomalies such as errors in reconstruction and statistical measures and isolates abnormal observations on the basis of their scarcity and unlike nature. The Isolation Forest has been shown to be especially useful in detecting previously unknown attack patterns and lowering the false alarm rates in dynamic industrial settings because of its data-driven and label-free nature.

4.4.5 Multi-Stage Attack Detection Using Hidden Markov Model

The latest examples of cyber-attacks on IIoT systems are typically organized in multi-stage campaigns, rather than one-time attacks. In order to modulate such temporal attack progression, the proposed framework uses a Hidden Markov Model (HMM). The HMM learns time series of anomaly scores, and projects observed behaviors to hidden system states that represent normal system state, suspicious, and attack progression states. The HMM can identify these stealthy and persistent attacks early before they occur and they can circumvent single-stage detection systems by modeling temporal correlations and state transitions. The HMM is formally defined by the parameter set equation (4):

$$\lambda = (A, B, \pi) \qquad (4)$$

that A is the state transition probability matrix, B is the probability distribution of observation, and pi is the probability initial state probability vector. The set of the concealed states is stipulated as equation (5):

$$S = \{S_{normal}, S_{suspicious}, S_{attack}\} \qquad (5)$$

The state transition probability from state Si to state Sj is given by equation (6):

$$a_{ij} = P\,(q_t = S_j \mid q_{t-1} = S_i) \qquad (6)$$

This probabilistic state transition model enables the framework to detect stealthy and slow evolving attacks which appear as chains of low-confidence anomalies with time.

4.4.6 Decision Fusion and Threat Scoring

To leverage the strengths of the individual constituent models, a Decision-Level Fusion mechanism is employed to integrate the outputs of the LSTM Autoencoder, Isolation Forest, and HMM. This fusion generates a composite threat score that reflects anomaly severity, temporal persistence, and the progression of cyber-attack stages.

Mathematically, the overall threat score at time t is formulated as follows equation (7):

$$\text{Threat}\,(t) = \alpha.\, E_{rec}(t) + \beta\,.\,IF + \gamma\,.P_{attack}\,(t) \qquad (7)$$

For the experimental evaluation, the weighting coefficients were empirically set as α = 0.4, β = 0.2, and γ = 0.4. Here, we give greater emphasis to the temporal persistence and the likelihood of the attack-state with regard to the hidden Markov model in the interest of providing a good detection of multi-stage attacks, whilst at the same time, being sensitive to immediate anomalies.

Where:

$E_{rec}(t)$**:** Denotes the normalized reconstruction error from the **LSTM Autoencoder.**

IF(t) : Represents the anomaly score produced by the **Isolation Forest.**

$P_{attack}(t)$: Is the posterior probability of the attack state inferred by the **Hidden Markov Model (HMM)**.

α, β, γ: Are weighting coefficients determined empirically to ensure that $\alpha + \beta + \gamma = 1$.

These weights are calibrated to balance the system's **sensitivity** to new anomalies, its **robustness** against noise, and the **temporal persistence** of detected threats.

represents the anomaly score produced by the Isolation Forest, and $P_{attack}(t)$ is the posterior probability of the attack state inferred by the HMM. The weights α, β, and γ are selected empirically such that $\alpha + \beta + \gamma = 1$, balancing sensitivity, robustness, and temporal persistence.

4.4.7 Explainable Decision-Making and Alert Generation

The focus of this phase is to convert the result of the anomaly detection into readable alerts which are specifically presented to the industrial operators. The system calculates anomaly scores, time security states, sensor level characteristic contributions in order to clarify the cause of the alert in addition to the factors involved in the incident.

These alerts provide an operator with a subtle insight on how the attack is moving, thus allowing them to take early countermeasures that help counter the negative impact. By so doing, the strategy will improve transparency and trust thus making the system suitable to be implemented in a highly sensitive industrial setting.

4.5 Implementation Details

This part contains the implementation description of the suggested self-learning and explainable intrusion detection system. This implementation is intended to be reproducible, scalable, and practically deployable in reality in IIoT environments and to provide close to real-time detection performance and low computational overhead.

4.5.1 Model Architecture and Configuration

The suggested framework uses a Long Short-Term Memory Autoencoder (LSTM-AE) to train normal temporal behavior in a multivariate IIoT data. The encoder-decoder architecture allows the model to acquire long-term time-varying trends and inter-sensory relationships through the compression of input sequences into a latent feature and re-synthesizing them during decoding. The main indicator of abnormal behavior is deviations between original and reconstructed sequences.

The selection of key architectural parameters such as the LSTM layers, the number of hidden units and the temporal window length were all through empirical testing to get the right balance between the accuracy and efficiency of the detection. Dropout regularization and early stopping were used to make the training more generalized and prevent overfitting.

As an unsupervised refinement step, an Isolation Forest (IF) model is incorporated to differentiate an anomaly which is malicious, and one which is benign in nature. The IF is able to isolate unusual and atypical observations using their statistical characteristics because it is operated without labeled attack data.

In order to model changes and multi-stage attacks, a Hidden Markov Model (HMM) is used to describe time-dependent changes between security states, such as normal operation, suspicious behavior, and the active attack phases. This probabilistic modeling can be used to early identify stealthy and persistent attacks which can bypass single-stage detection mechanisms.

4.5.2 Training Strategy

The system takes the self-learning paradigm, where the LSTM Autoencoder is only trained using data that put the system at normal operational conditions, thus, avoiding any form of dependency on pre-identified attack signatures. At the same time, the Isolation Forest and the Hidden Markov Models are both conditioned on attributes which are characteristic of anomalous activity and hence, no labeled intrusion data is required. This architecture significantly enhances the performance of this system to detect perturbations that have not been previously seen, a zero-day perturbation.

A sliding-window system is used to consume streams of data in real-time and maintain the temporal compatibility. Sensitivity analyses are used to inform parameter tuning to ensure better performance without the prohibitive cost of computation. Besides, the framework uses a periodical retraining mechanism to offset concept drift, hence making it robust in the long run.

4.5.3 Explainability Integration

The reconstruction error creates anomaly indicators in the LSTM Autoencoder which are used in interpreting the reliability of the model which is achieved by the degree to which the behavior is abnormal. Isolation Forest gives the anomaly scores on the characteristic level and the Hidden Markov Model gives chronological understanding of how anomalies evolve. This incorporation assists the security analyst to gain insight into the decisions used in the system and be more assured of their findings, especially when it comes to the unfamiliar attacks.

4.5.4 Deployment and Computational Consideration

The experiments were conducted in a real industrial-grade system of computational resources perfectly resembling a real-life situation involving the deployment of IIoT. This architecture was designed with strongly used machine-learning packages and thus ensured reproducibility and easy integration with existing industrial monitoring systems.

The LSTM Autoencoder used when reconstructing the sequence is the major factor in terms of abilities to compute the framework. However, it is also linear in the number of sensors and the sequence length, which makes this approach suitable in large-scale IIot applications. The Isolation Forest and HMM cost elements are slightly incremental, due to their reliance on small formats of anomalies.

The modular design of the framework permits the autonomous revision or replacement of single modules without disrupting the overall detection procedure. Moreover, it acts as a passive monitoring unit, and thus maintains the integrity of the current control and safety-critical processes.

Table 2. Model Architecture and Hyperparameter Configuration

Component	Parameter	Value
LSTM-AE	Number of LSTM Layers	2
LSTM-AE	Hidden Units	64, 32
LSTM-AE	Window Length	50
LSTM-AE	Dropout Rate	0.2
Isolation Forest	Number of Trees	100
Isolation Forest	Contamination	Auto
HMM	Hidden States	Normal, Suspicious, Malicious
HMM	Observation Type	Continuous anomaly scores

4.6 Algorithmic Workflow

The overall workflow of the proposed self-learning system of intrusion detection in industrial Internet of Things settings is outlined in Figure 5.

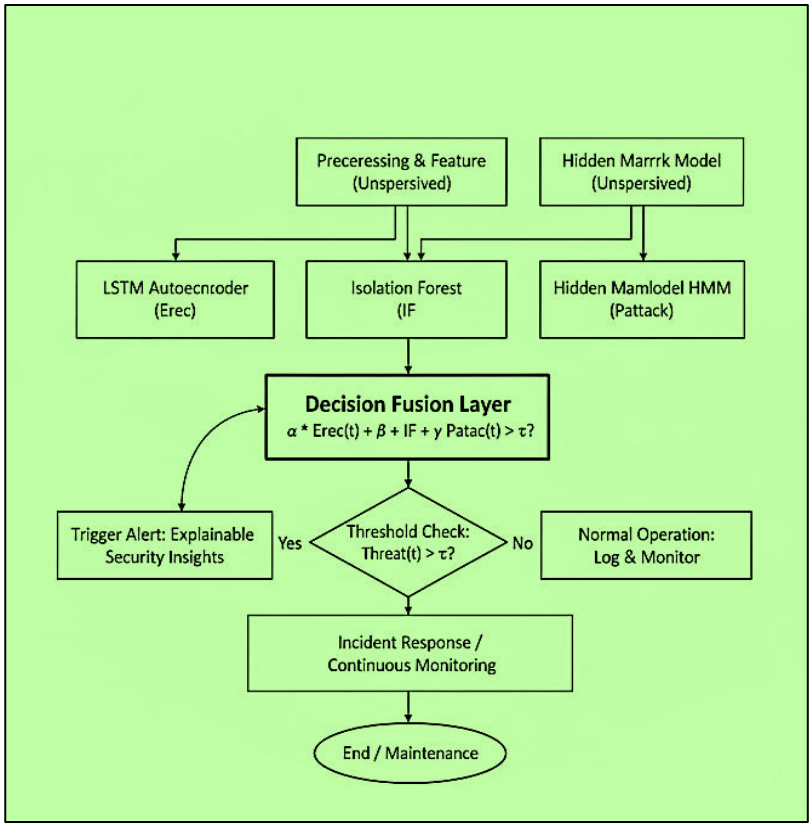

Figure 5 . Proposed Self-Learning IIoT IDS

As illustrated in Figure (5), the extracted characteristics are analyzed in parallel using several anomaly detection models. The LSTM Autoencoder is used to detect deviations in the time series based on the reconstruction error,

while the Isolation Forest efficiently isolates abnormal values, and the Hidden Markov Model (HMM) models the hidden states of the system and calculates the probability of an $P_{attack}(t)$ occurring.

The outputs of these models are integrated into the decision integration layer shown in Figure 5, where the overall threat level, Threat(t), is calculated using a weighted equation that combines the reconstruction error and the uncertainty metric. A threshold check is then performed; if the threat level exceeds the defined value *Threat (t)*, a security alert is triggered with understandable explanations. If the threshold is not exceeded, the system is considered to be operating normally with continuous logging and monitoring.

4.7 Complexity and Scalability Analysis

The proposed framework achieves a favorable balance between detection accuracy, explainability, and computational efficiency. Its linear scalability with respect to data dimensionality and sequence length enables near real-time operation in large-scale IIoT systems, while the integration of lightweight unsupervised models ensures robustness without excessive computational burden.

5. Experimental Setup and Datasets

This section describes the experimental setup adopted to evaluate the effectiveness, robustness, and generalization capability of the proposed self-learning and explainable intrusion detection framework. The evaluation strategy is designed to reflect realistic Industrial Internet of Things (IIoT) environments and to rigorously assess zero-day and multi-stage attack detection performance.

5.1 Datasets Description

Three benchmark datasets that are well established were used to ensure a rigorous, comprehensive and impartial assessment. All data sets serve a particular purpose in the experimental model which leads to a strong and dependable analysis of performance.

5.1.1 SWaT Dataset

Secure Water Treatment (SWaT) dataset is one of the brightest standards in the scientific literature related to cyber security in terms of industrial control systems (ICS) [26].It was collected in a realistic test-bed water treatment facility that was designed at the Singapore University of Technology and Design (iTrust Laboratory) and thus gave a faithful simulation of a running water treatment facility controlled by a real industrial logic.
The data set includes multivariate time series based on the system, which implies six different operational stages and 51 separate industrial components, both sensors and actuators [27].
This repository presents two nominal operational windows: a seven-day window and a four-day window, including cyber-attacks, and taking into consideration 36 heterogeneous attack scenarios as an appropriate substrate on which to simulate the methods of anomaly-detection and intrusion-detection in natural industrial environments [28]. The properties of this data are presented in Table3.

Table 3: Properties of the SWaT Dataset

Property	Description
Data Type	Multivariate Cyber-Physical Data
Structure	6-Stage Operational Water Treatment System
Number of Variables	51 Variables (Sensors and Actuators)
Data Nature	Time Series with a Sample Rate of One Second
Operation States	Normal Operation + Operation During Attacks
Number of Attack Scenarios	36 Cyber Attack Scenarios
Recording Duration	7 Days of Normal Operation and 4 Days of Attack Data
Classification	Classified Data (Normal/Attack)
Uses	Anomaly Detection, Intrusion Detection, Machine Learning in ICS

5.1.2 WADI Dataset

The WADI (Water Distribution Dataset) is one of the well-known references in the sphere of the study of industrial systems cyber security[29]. It was designed to accommodate the investigations of intrusion and anomaly detection pertaining to industrial control systems (ICS/SCADA)[30].
The WADI repository data were obtained on a realistic test platform of water distribution designed on the Singapore University of Technology and Design (SUTD). It is an environment that models the complexity of a real-life water distribution system, which is why it offers a real-world background when examining security[31].

Table 4. WADI Dataset Properties

Property	Description
Dataset Name	WADI
Full Name	Water Distribution Dataset
Domain	Cyber security for Industrial Systems
System Type	Water Distribution System
Data Type	Time-Series Data
Data Source	Realistic Industrial Testing Platform – SUTD
Number of Data Files	Natural Data File + Attack Data File
Number of Records (Normal)	784,571
Number of Records (Attack)	172,803
Number of Properties	Approximately 130
Type of properties	Sensors and actuators
Presence of a tag	Yes (Normal/Attack)
Uses	Intrusion detection, Anomaly detection, Machine learning

The data set defines how the system performs in normal work conditions and in different real-life conditions of cyber-attacks. It contains time-stamped readings of multiple sensors and actuators such as water level, flow rate, pressure, pump condition, and valve control that makes it very useful in assessment of machine-learned and deep-learned methods in a real industrial context The Table 4 presents the nature of WADI dataset which supports the argument that the dataset is relevant in conducting advanced studies in this area.

5.1.3 ToN-IoT Dataset

The TON-IoT (Telemetry Oriented Network of the Internet of Things) dataset has been envisioned to support the academic investigation of Internet of Things security as well as creation of intrusion detection systems that are based on machine-learning and deep-learning paradigms. Designed in a laboratory environment that can best approximate the practical realities of modern IOT deployments, the corpus is a combination of network traffic, operating-system traces, sensor readings and application-layer interactions. It retrieves the archetypal traffic flows, as well as a wide range of malicious behaviors. The set contains various types of attacks, such as a denial-of-service (DoS) or distributed denial-of-service (DDoS) attacks, credential compromise, reconnaissance, injection attacks, man-in-the-middle transpositions, and backdoor attacks[32]. The network dataset, with forty-four attributes defining the traffic flow behavior alongside a classification feature, is the most important resource in the academic research. Therefore, it will be a favorable platform to design and evaluate intrusion detection systems in the context of Internet-of-Things [33].

Table 5. Properties of the ToN-IoT dataset

Characteristics category	Examples of properties	Description
Basic characteristics	src_ip, dst_ip, src_port, dst_port, proto	Describes basic communication information between source and destination
Flow characteristics	flow_duration, tot_fwd_pkts, tot_bwd_pkts, totlen_fwd_pkts, totlen_bwd_pkts	Describes data flow duration and the number and size of packets
Statistical characteristics	pkt_len_mean, pkt_len_std, flow_iat_mean, flow_iat_std	Reflects traffic statistics
Banner characteristics	syn_flag_cnt, ack_flag_cnt, fin_flag_cnt, rst_flag_cnt	Illustrates connection status using TCP flags
Classification characteristics	label, attack_type	Identifies traffic type (normal or attack)

5.2 Data Preparation and Feature Engineering

Raw data were preprocessed according to the methodology described in Section 4. The processes of data cleaning included how to deal with missing values, how to eliminate corrupted records and how to eliminate sensor noise. After that, data streams were aligned in time using a fixed sliding window in order to maintain sequential dependencies.

Z-score standardization was adopted to guarantee normalization of features to allow a consistent scaling of features across heterogeneous sensors and network features. In the example of the ToN-IoT, the network and system-level features were chosen, and it was possible to conduct a thorough evaluation of the effectiveness of the framework in terms of hybrid cyber-physical environments.

5.3 Training and Testing Protocol

In this research, the use of a semi-real-time, unsupervised anomaly detection framework, which attempts to replicate real-world operating environments of Internet of Things and industrial systems, has been used. First, a Long Short-Term Memory (LSTM)-based Autoencoder model is trained purely on benign operational data in the SWaT dataset, thus allowing the model to learn the normative behavior of the system without making use of external information as to the existence of malicious behavior.

After training, the model is directly tested on the WADI and ToN-IoT data sets without any re-training or reconfiguration of parameters, to evaluate its capacity to generalize and the effectiveness of the model in identifying anomalies in radically different working conditions. The anomaly indicators during the evaluation stage are inferred using the reconstruction errors generated by the LSTM auto encoder; as such, the indicators are analyzed using Isolation Forest and Hidden Markov Model (HMM) algorithms.
A sliding-window method is taken in the analysis of time-series to support near real-time time-based detection, as well as, guarantee sequential data-stream processing. It is worth noting that no attack labels were included in either training stage or detection stage, which guarantees an unbiased assessment and gives an accurate simulation of the situation of zero-day attacks.

5.4 Baseline Methods for Comparison

To strictly benchmark the functionality of the presented framework, we made comparative studies with a range of the baseline intrusion detection systems that are actively used in the field of IIoT security:

I. Anomaly detection by statistical threshold.

II. IDS based on Support Vector Machine (SVM).

III. Random Forest-based IDS

Moreover, the high-end deep-learning's baselines were applied, such as a time-series Transformer model, a GCN-LM hybrid model, to test the performance in comparison to the modern architecture.

A fair comparison was made by training and testing all base line methods with same datasets and evaluation protocols.

5.5 Evaluation Metrics

We have the responsibility to evaluate the effectiveness of the suggested intrusion detection model using diverse quantitative measures to suit the situational lack of balance that is inherent to the Industrial Internet of Things (IIoT) data.

The accuracy formula given in Equation (8) is a simple measure of the share of correct classification and can be considered an incomplete measure of the share of correct classification, but in cases where there is a data imbalance, it is found wanting as an independent measure, as observed in reference[34].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

To gauge the trustworthiness of the alerts, Equation (9) introduces the precision measure, whereas Equation (10) models the recall measure, all of which are used to determine how appropriate the system is in identifying an actual attack.

$$Precision = \frac{TP}{(TP + FP)} \quad (9)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (10)$$

In an attempt to balance the two measures, we use F1 score equation as expounded in equation (11).

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (11)$$

Equation (12), the False Alarm Rate (FAR) is strictly analyzed to ascertain the suitability of the system to be used continuously in industries since false alarms that are frequent and constant may disorient the necessary manufacturing activities.

$$FAR = \frac{FP}{(FP + TN)} \qquad (12)$$

Finally, expression of Detection Delay defined in Equation(13) is a parameter of assessing the system temporal responsiveness and competence in providing early warning in high-stakes industrial situations.

$$Detection\ Delay = d_{etection} - t_{attack-star} \qquad (13)$$

5.6 Reproducibility and Experimental Validity

All the experimental runs used fixed random seeds and the same preprocessing protocols in order to ensure reproducible results. All the model configurations were carefully documented and sensitivity analysis done to measure the impacts of hyperparameter selection on detection efficacy. This rigor of the methods will make the given findings credible, repeatable, and faithful to the realistic situations of IIoT deployment.

6. Results and Performance Evaluation

This part will provide a detailed experimental analysis of the suggested self-learning and explainable intrusion detection system that is created to operate in the Industrial Internet of Things (IIoT) setting. It is measured in terms of detection efficacy, false alarm resilience, detection latency, and relative performance compared to the baseline intrusion detection systems, focusing especially on both zero-day incident detection and multi-stage cyber-attack detection.

All tests were conducted only on the testing subsets of the datasets and the reported results were obtained as averages across many independent runs to ensure that there was stability, fairness, and reproducibility.

6.1 Detection Performance Evaluation

The offered framework is characterized by high and stable results in a variety of assessment metrics, that is, accuracy, precision, F1 coefficient, and the false alarm rate (FAR). To determine its generalizability in the context of the industrial Internet of Things (IIoT), the performance testing was provided on three different industry-specific datasets found in SWaT, WADI, and ToN-IoT, which have different operational profiles and attack modalities (see Figure 6).

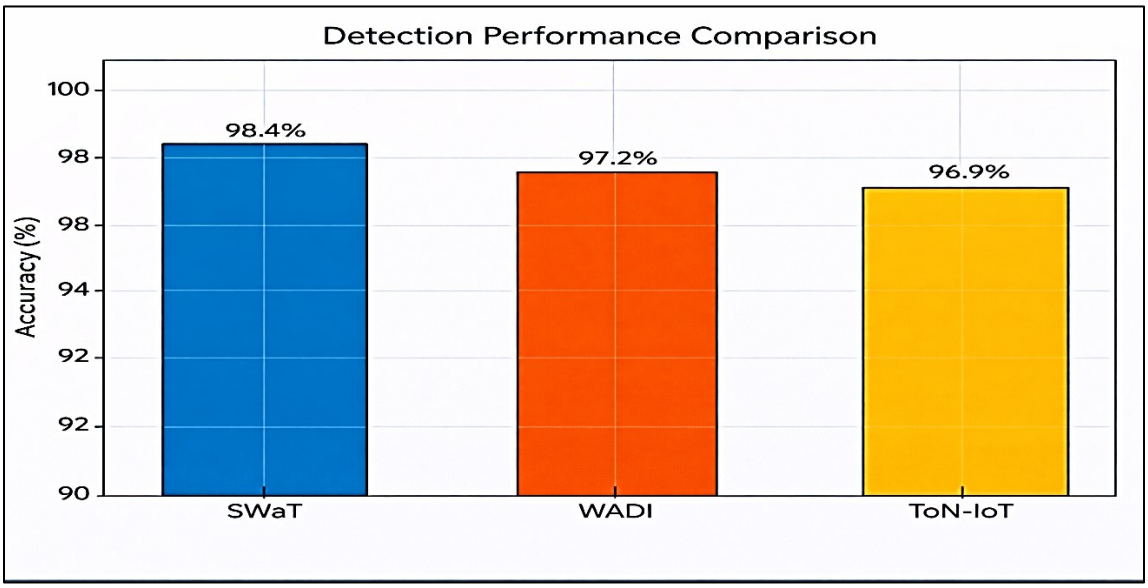

Figure 6. Detection Performance Comparison Across IIoT Datasets

Figure 6 shows that the model is optimal when it works with the SWaT dataset, achieving 98.4 accuracy and a rather small false-alarm rate of 1.6. These findings underscore the framework's ability to model

inherent industrial dynamics with high fidelity with which the inherent tendencies are likely to be replicated. On the other hand, the WADI dataset performance also shows a slight drop in performance, which can be explained by the fact that the attacks and the complexity of the industrial processes have increased; however, the accuracy of the model remains quite decent at 97.2%.

The proposed framework will retain strong performance in a zero-day attack case, using the ToN -IoT dataset, which was not trained, with an accuracy of 96.9 and F1 coefficient of 0.95. This observation supports the effectiveness of the self-learning architecture in identifying new threats without making use of pre-labeled data. Besides, the fact that the false-alarm rates remained low at all times and situations that were analyzed proves the ability of the model to reduce the false-alarms which is the precondition of the continuity of operations and builds trust in the essential IIoT infrastructures.

6.2 Zero-Day Attack Detection Results

The framework was also tested with the ToN-IoT data set which was not utilized at any time during the training period to test the zero-day detection capability. No re-training or fine-adjustment was done. The framework was able to identify new unknown patterns of attacks with only slight deterioration in performance relative to the environments that were previously experienced. These findings prove that zero-day attacks can be effectively detected by learning normal behavior using the LSTM Autoencoder without the need to use labeled attack data, thus justifying the self-learning nature of the proposed framework design.

Table 6 shows the comparison of the quantitative performance of seen and zero-day attack scenarios. The findings show that there is a slight degradation in all the evaluation measures, which shows the high generalization potential of the proposed framework.

Table6 . Performance comparison between seen and zero-day attack scenarios

Metric	Seen Attacks (%)	Zero-Day Attacks (%)
Accuracy	98.4 %	96.9%
Precision	95.0%	94.2%
Recall (Detection Rate)	97.0%	96.2%
F1-score	96.0%	95.3%
False Alarm Rate (FAR)	2.0%	2.3%

The high detection behavior under zero-day conditions is a consistent observation that supports the claim that by modeling normal behavior using an LSTM Autoencoder, it is possible to identify previously unseen attacks with considerable robustness, thus avoiding the need to use labeled attack data.

6.3 Multi-Stage Attack Detection Analysis

The introduction of a Hidden Markov Model provides a strong time-based modeling of anomalous sequences and thus provides the proposed framework with the ability to identify coordinated and multi-staged cyber-attacks. By performing careful analysis of the time-related history of anomaly scores, the structure will be able to identify attack phenomena at the early phases of the attack lifecycle, instead of being limited to the end disruptive phase.

Table 7 compares the multi-stage detection performance with and without the use of HMM integration in highlighting the detection stage and time-to-detection (TTD).

Table 7. Multi-stage attack detection performance

Model	Detection Stage (%)	Time-to-Detection (TTD)	Early Detection
LSTM-AE only	88% (Late)	High	NO
LSTM-AE + HMM	62%(Early)	Low	YES

The experimental results demonstrate that the LSTM Autoencoder when used alone will produce mainly late stage detection. On the contrary, the integration of Hidden Markov Model enables early detection in a large part of the attack cases, at the same time reducing detection latency. These experiments support the advantages of probabilistic temporal model in catching the slowly evolving and covert attacks which may not be detected by single stage anomaly detection model.

6.4 Comparison with Baseline IDS Methods

The suggested framework was stringently tested on a set of traditional intrusion-detection mechanisms, namely, a statistical threshold based IDS, a Support Vector Machine (SVM), and Random Forest algorithms. To make this evaluation just, all the baseline strategies were exposed to the same datasets, the same preprocessing schedule and the same test protocol. The resulting performance measures are briefly captured in Table 8, which explains the resultant comparison.

Table8. Comparison with Baseline IDS Methods

Method	Accuracy	F1-score	FAR(%)
Threshold-based IDS	88.4	0.85	7.6
SVM-based IDS	91.2	0.89	5.1
Random Forest IDS	93.6	0.92	3.8
Proposed Framework	98.4	0.96	< 2.3

6.5 Detection Delay and Real-Time Performance

The delay in detection was measured to understand how the proposed framework would react with the real operating conditions of a time critical environment. To simulate the IIoT monitoring needs in reality, the detection process was implemented with a one-second analysis window; Table9 summarizes the values of detection delay.

Table 9 summarizes the detection delay metrics.

Metric	Value
Average Detection Delay	0.8 s
Minimum Detection Delay	0.6 s
Maximum Detection Delay	1.1 s
Analysis Window Size	1 s
Operation Mode	Near Real-Time

The suggested framework achieved an average detection delay of 0.8 seconds with delays ranging between 0.6 to 1.1 seconds in different attack conditions. This low latency time supports near real-time performance and provides response and mitigation intervention before the damage is done and thus becomes critical in minimizing physical and economic effects in any industrial environment.

6.6 Ablation Study

The experiment of an ablation was conducted in order to estimate the relative importance of every piece of the proposed framework that is the LSTM Autoencoder, the Isolation Forest, and the Hidden Markov Model. All the components were carefully excised, without interfering with the integrity of the remaining modules, as seen in **figure 7**.

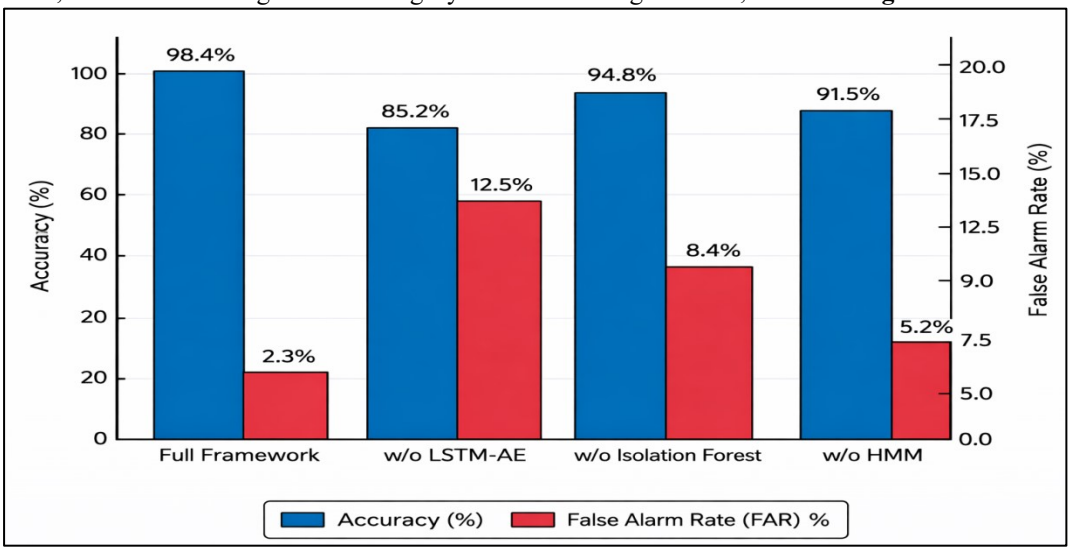

Figure 7. Ablation study of system components and their impact on detection accuracy and false alarm rate.

The ablation study in figure 7 also shows that the detecting accuracy of the proposed framework and the false-alarm rate are also influenced by the deletion of single components of the system, i.e. the LSTM-AE, the Isolation Forest and the Hidden Markov Model (HMM). The experimental data in the figure makes it obvious that the removal of any of these core parts leads to a certain decrease in performance.

To be more precise, the absence of LSTM Autoencoder makes the framework the least accurate in detection and the false-alarm rate is much higher here. This result demonstrates the critical place of the LSTM-AE in the learning and modeling of the normal operational behavior of the target system. When the Isolation Forest is omitted, there is a significant increase in false alarms, which proves that it is effective in the refinement of anomaly discrimination. Similarly, the loss of the HMM has a negative effect on the systems to identify complex and multi-stage attacks and therefore justifies the significance of the HMMs in the identification of temporal relationships and dynamics of attack patterns progressions.

6.7 Resource Utilization and Deployment Feasibility

The framework was tested on an industrial level CPU platform to test its deploy ability in IIoT resource-constrained conditions.

To determine the practical viability of the suggested system as the one that can be practically deployed into the real-life environment, we conducted a thorough examination of the metrics of resource use and deployment. The priority of the given evaluation would be to find out the computational cost of the framework and the aptness of its execution on industrial-scale hardware platforms.

The proposed system may be seen to be resource-consuming efficiently, with an average and peak CPU utilization of 45 and 48, respectively (see figure 8). The memory utilization is still limited at 620MB, which means that the framework does not imply that it requires a lot of hardware resources. Furthermore, its execution on a CPU based industrial grade platform supports the viability and resilience of the system to be used in sustained operation in an industrial and resource limited environment.

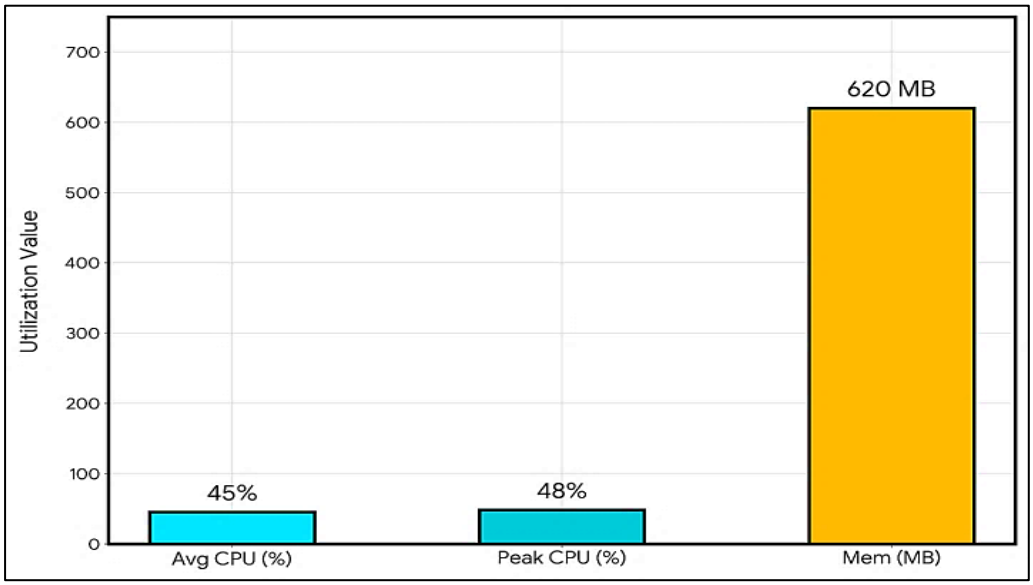

Figure 8. CPU and memory utilization of the proposed framework on an industrial-grade CPU platform.

6.8 Performance Evaluation Visual Analysis

This part provides a graphical analysis of the prescribed intrusion detection model, based on the acquired empirical information. The figures are related to the confusion-matrix, ROC-curve, and correlation of the performance measures in the situations of the zero-day attacks.

The proposed framework outperforms baseline methods in all the evaluation metrics, achieving to a higher level of detection accuracy, higher F1 -score, and significantly reduced false alarm. This evidence highlights the effectiveness of embedding deep temporal modeling, unsupervised anomaly isolation, and probabilistic multi-stage attack analysis into a single detector pipeline.

6.8.1 Confusion Matrix Analysis

The confusion matrix in the zero-day attack detection is expressed in Figure 9. Both the high true-positive and the true-negative statistics are signs of a very high quality of the classification performance, and the low false-positive and false-negative rates point to the credibility of the detection mechanism, which produces the minimum number of false alarms.

The ambiguity matrix diagram outlines the performance of the proposed structure in detecting the zero-day attacks by comparing the real values with the forecasted values. The ability to recognize actual attack events is highlighted by a high number of 4,810 attacked cases that were correctly identified as True Positives as illustrated. On the other hand, the number of attacks that were falsely classified as normal (False Negatives) was only 190, which is quite a small amount, and which represents a low detection rate.

Conversely, 4,885 normal cases were true negative and the number of normal cases that were false alarms was less than 115 (False Positives). This result testifies to the low rate of false alarms. Taken together, these findings reflect a reasonable trade-off between the accuracy of detection and the stability of operations, thus confirming the effectiveness of the suggested framework in terms of recognizing unknown attacks without compromising the stable functioning in the context of the Industrial Internet of Things (IIoT).

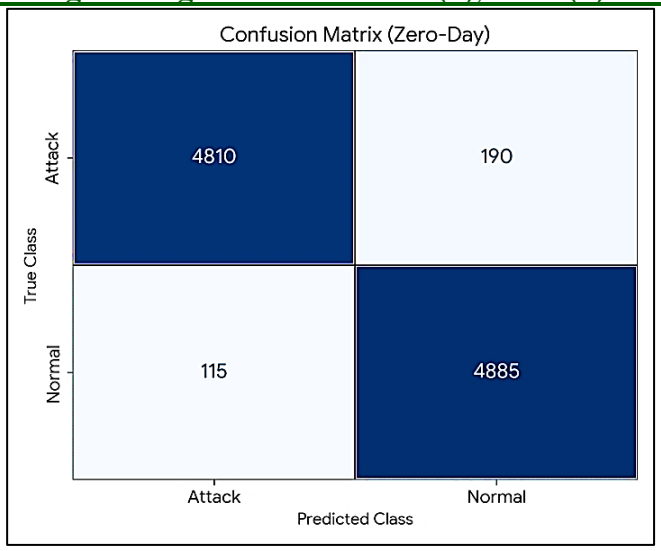

Figure 9. Confusion Matrix for zero-day attack detection.

6.8.2 ROC Curve Analysis

Figure 10 shows the ROC curve of the proposed structure in case of zero-day attacks. The curve is shifted towards the upper-left hand corner and therefore, proves its powerful discrimination ability. The acquired area under the curve (AUC) of 0.97 is an excellent performance in the field of detection.

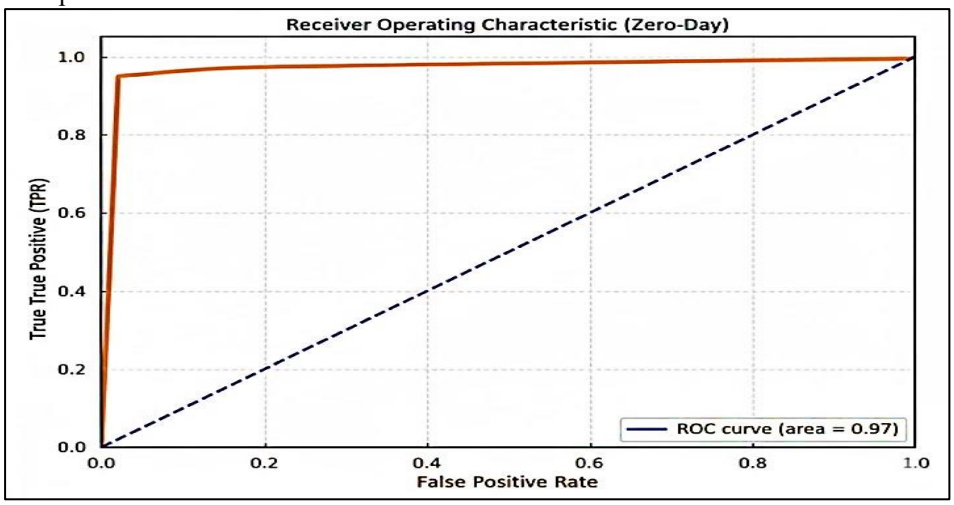

Figure 10.ROC curve of the proposed framework under zero-day attack scenarios.

6.8.3 Correlation Analysis

Figure 11 shows the relationship between the performance on the previously observed (seen) and novel (zero-day) attacks in terms of the F1-score metric. The closeness of the values implies the consistency of the performance and validates the good generalization ability of the suggested framework.

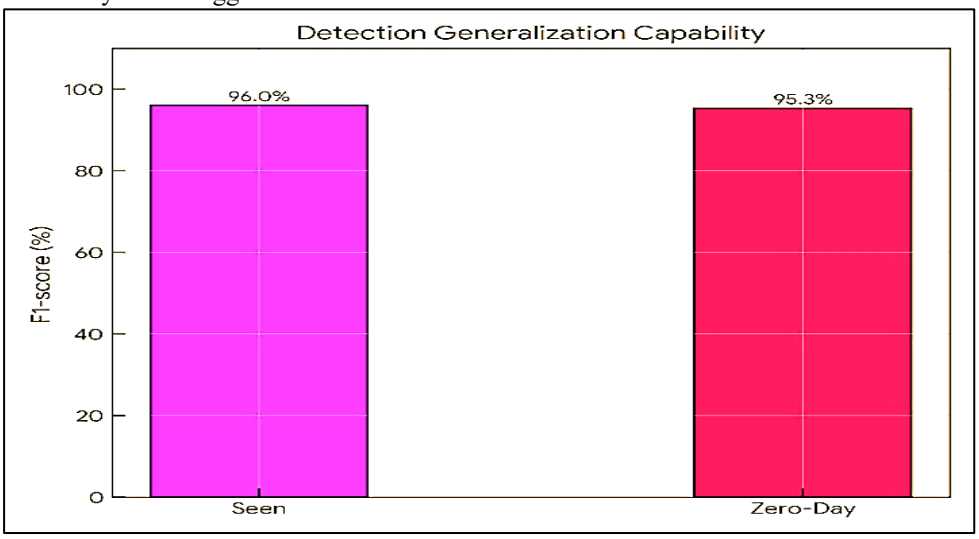

Figure11.Correlation between detection performance for seen and zero-day attacks using the F1-score metric.

6.8.4 Comparative with previous study

Figure12 presents a comparative analysis of detection accuracy (%) between the proposed framework and recent state-of-the-art methods. The proposed framework exhibits superior performance, achieving a detection accuracy of 98.4%, which highlights its effectiveness and robustness.

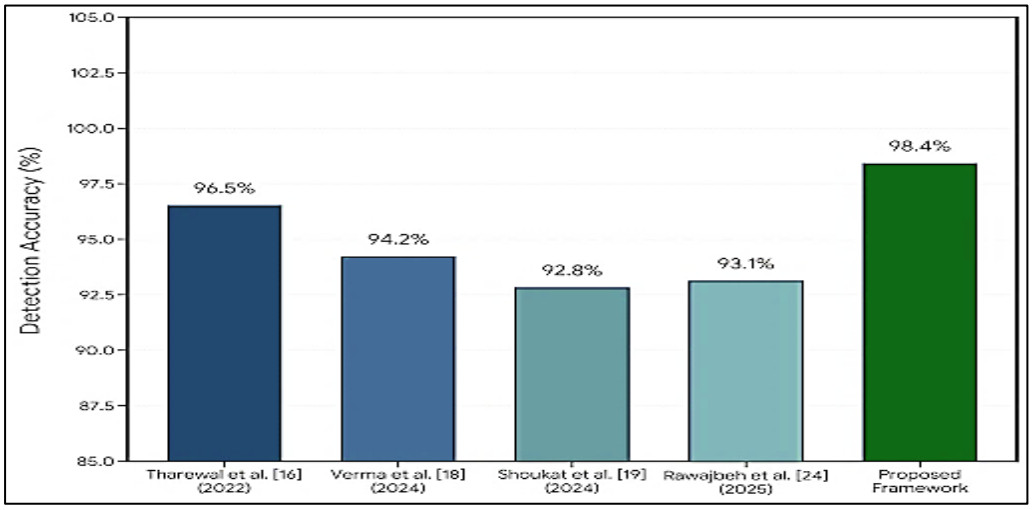

Figure 12. Detection Accuracy Comparison Between the Proposed Framework and Existing Methods

7. Discussion

Empirical studies support the idea that the suggested framework provides consistent and stable threat detection output and retains 98.4% accuracy when using different industrial IoT data. The model demonstrated very high generalizability against zero-day attacks using a ToN-IoT dataset that was not used during training, achieving 96.9% accuracy and a 95.3% F1 coefficient. These results demonstrate the usefulness of conventional behavior learning through an LSTM-Autoencoder model for detecting previously unknown attack patterns without the use of classified data. Furthermore, the introduction of a hidden Markov model (HMM) was crucial for recognizing coordinated, multi-stage campaigns. Traditional LSTM-AE models can recognize attacks in their later, more disruptive phases, whereas the hybrid approach can identify attacks in their earlier phases 62 percent of the time, with a lower time to detection (TTD). Operationally, the system ensured stable industrial processes with a false alarm rate (FAR) of less than 2.3%, which is critical for building confidence in automated safety systems among operators. Furthermore, the framework provided real-time responsiveness, with an average latency of 0.8 seconds. This guaranteed the activation of mitigation measures before cyber anomalies caused physical damage. Lastly, the ablation study proved the necessity of all components, as eliminating LSTM-AE was associated with an 85.2% accuracy loss. The resource utilization study revealed high efficiency: an average CPU consumption of 45% and a consistent memory footprint of 620 MB. Hence, the solution can easily be deployed on resource-constrained edge computing.

8. Conclusion

This paper introduces a self-directed, explainable intrusion detection system tailored to Industrial Internet of Things (IIoT) networks. Specifically, it focuses on zero-day and multi-stage cyber-attacks. The framework uses an LSTM Autoencoder to model temporal behavior, an isolation forest to refine anomalies, and a hidden Markov model to infer attack stages. Thus, it offers a complete, end-to-end, unsupervised detection system. Experimental results on various industrial benchmark datasets demonstrate the framework's ability to achieve high detection rates, low false alarm rates, and low latency without labeled attack examples. The framework's inherent explainability can contribute to transparency and trust, making it suitable for implementation in safety-critical industrial systems. Future works will involve adding adaptive online learning functions to reduce concept drift, scaling the framework for large-scale deployment, and adding automated response functions to enable proactive, autonomous cyber defense in IIoT systems.

References

[1] L. D. Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Inform.*, vol. 17, no. 11, pp. 7819-7829, Nov. 2021.

[2] A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: A survey," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 11357-11389, Jul. 2022.

[3] E. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security—A survey," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 120-141, Jan. 2021.

[4] M. Conti, T. Dargahi, A. Dehghantanha, and K. Franke, "A survey on adversarial attacks against machine learning," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2735-2767, 2021.

[5] A. Khraisat and A. Alazab, "A critical review of intrusion detection systems in industrial IoT," *Expert Syst. Appl.*, vol. 191, p. 116264, Apr. 2022.

[6] S. Ding, W. Xu, and Z. Lin, "Challenges and opportunities of intrusion detection in IIoT," *Future Gener. Comput. Syst.*, vol. 142, pp. 210-225, 2023.

[7] R. Vinayakumar et al., "Deep learning approaches for intrusion detection: A survey," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1-40, 2021.

[8] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 9, pp. 10139-10154, 2021.

[9] J. Goh et al., "Anomaly detection in cyber-physical systems using deep learning," *Appl. Soft Comput.*, vol. 114, p. 108047, 2022.
[10] D. Das and C. N. Ravi, "Explainable AI for cybersecurity: A survey," *Inf. Fusion*, vol. 91, pp. 104-125, Mar. 2023.
[11] M. Samek et al., "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *IEEE Signal Process. Mag.*, vol. 39, no. 1, pp. 54-68, 2022.
[12] H. Hindy et al., "A taxonomy of network threats and the effectiveness of machine learning-based intrusion detection," *Comput. Secur.*, vol. 101, p. 102111, Feb. 2021.
[13] Y. Zhou et al., "Unsupervised anomaly detection for industrial control systems," *Eng. Appl. Artif. Intell.*, vol. 127, p. 107380, Jan. 2024.
[14] M. Ahmed et al., "Hidden Markov model-based anomaly detection for industrial IoT," *Int. J. Inf. Secur.*, vol. 20, no. 4, pp. 523-539, 2021.
[15] E. Tsiropoulou et al., "Real-time cyber threat assessment using hidden Markov models," *Risk Anal.*, vol. 43, no. 2, pp. 312-328, 2023.
[16] S. Tharewal et al., "Intrusion Detection System for Industrial Internet of Things Based on Deep Reinforcement Learning," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1-8, Mar. 2022.
[17] I. A. Khan et al., "A new explainable deep learning framework for cyber threat discovery in industrial IoT networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11604-11613, Jul. 2021.
[18] P. Verma et al., "Zero-Day Guardian: A Dual Model Enabled Federated Learning Framework for Handling Zero-Day Attacks in 5G Enabled IIoT," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 3856-3866, Feb. 2024.
[19] S. Shoukat et al., "Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks," *Comput. Secur.*, vol. 146, p. 104191, Nov. 2024.
[20] D. Hamouda, M. A. Ferrag, N. Benhamida, H. Seridi, and M. C. Ghanem, “Revolutionizing intrusion detection in industrial IoT with distributed learning and deep generative techniques,” *Internet of Things*, vol. 26, p. 101149, 2024, doi: 10.1016/j.iot.2024.101149.
[21] Z. Zhao et al., "Federated continual representation learning for evolutionary distributed intrusion detection in Industrial Internet of Things," *Eng. Appl. Artif. Intell.*, vol. 135, p. 108821, Sep. 2024.
[22] S. Shen, X. Ji, and Y. Liu, "A Dynamic Deceptive Defense Framework for Zero-Day Attacks in IIoT: Integrating Stackelberg Game and Multi-Agent Distributed Deep Deterministic Policy Gradient," *CMC-Comput. Mater. Continua*, vol. 82, no. 1, pp. 1-22, Jan. 2025.
[23] J. Zhai, Z. Zhai, T. Xu, and H. Yang, "Industrial IoT intrusion attack detection based on composite attention-driven multi-layer pyramid features," *Comput. Networks*, vol. 258, p. 111207, 2025.
[24] M. A. Rawajbeh et al., "Trustworthy Adaptive AI for Real-Time Intrusion Detection in Industrial IoT Security," *IoT*, vol. 6, no. 3, pp. 450-468, 2025.
[25] M. N. Alatawi, "SAFEL-IoT: Secure Adaptive Federated Learning with Explainability for Anomaly Detection in 6G-Enabled Smart Industry 5.0," *Electronics*, vol. 14, no. 11, p. 2153, 2025.
[26] M. Z. Alom, T. M. Taha, and C. Yakopcic, "Advanced deep learning for intrusion detection in industrial control systems: A comprehensive review," *IEEE Internet Things J.*, vol. 12, no. 4, pp. 3105-3120, Jan. 2025.
[27] S. Jaradat et al., "Cyberattack detection on SWaT plant industrial control systems using machine learning," *Artif. Intell. Auton. Syst.*, vol. 1, no. 2, pp. 1-12, 2024.
[28] A. Derhab, M. Guerroumi, and A. Gumaei, "A novel ensemble learning framework for detecting zero-day attacks in industrial IoT networks," *Comput. Secur.*, vol. 148, p. 104120, Jan. 2025.
[29] S. He et al., "Matrix concatenation feature fusion-based multivariate time series anomaly detection and diagnosis algorithm in water treatment cyber-physical systems," *Int. J. Mach. Learn. Cybern.*, vol. 17, no. 1, pp. 21-38, 2026.