



AI-DRIVEN CYBERSECURITY FOR INDUSTRIAL IOT: COMBINING GRAPH NEURAL NETWORKS, FEDERATED LEARNING AND EXPLAINABLE AI

Dr. J. Sobhana Dev, V. Vidhya Lakshmi

Centre for Japanese Studies, School of Language, Literature and Culture Studies, Jawaharlal Nehru University

ABSTRACT

The Industrial Internet of Things (IIoT) has transformed modern industrial environments by connecting sensors, programmable logic controllers, supervisory control systems, machines, gateways, and cloud platforms into highly interconnected cyber-physical ecosystems. Although this connectivity improves operational efficiency, predictive maintenance, automation, and real-time decision-making, it also expands the attack surface and creates new cybersecurity challenges. Conventional intrusion detection approaches often struggle with heterogeneous device behavior, evolving attack patterns, distributed data, limited computational resources, and the requirement for rapid and interpretable security decisions. This paper proposes an integrated artificial intelligence framework for IIoT cybersecurity that combines Graph Neural Networks (GNNs), Federated Learning (FL), and Explainable Artificial Intelligence (XAI). The proposed architecture represents industrial communication environments as dynamic graphs, allowing GNNs to capture relationships among devices, communication flows, and operational entities. Federated learning enables multiple industrial sites or edge nodes to collaboratively train a security model without directly exchanging sensitive raw telemetry. XAI mechanisms provide human-interpretable explanations of detected anomalies and attack classifications, thereby improving analyst trust and facilitating incident investigation. The framework additionally incorporates secure aggregation, client-level anomaly monitoring, adaptive graph construction, and risk-aware decision-making. A comprehensive evaluation methodology is proposed using accuracy, precision, recall, F1-score, false-positive rate, detection latency, communication overhead, computational cost, privacy leakage, robustness against poisoning attacks, and explanation fidelity. The paper argues that the integration of relational learning, decentralized intelligence, and interpretability provides a promising direction for trustworthy IIoT security, while recognizing unresolved challenges involving non-IID data, adversarial model updates, concept drift, resource constraints, and explanation reliability.

Keywords: Industrial Internet of Things, IIoT Security, Graph Neural Networks, Federated Learning, Explainable AI, Intrusion Detection, Edge Computing, Cyber-Physical Systems, Anomaly Detection, Privacy-Preserving Machine Learning

INTRODUCTION

The Industrial Internet of Things (IIoT) represents the convergence of industrial control technologies, networked sensors, intelligent machines, edge computing, cloud platforms, and artificial intelligence. Modern manufacturing plants, energy systems, transportation infrastructures, water-treatment facilities, and process industries increasingly depend on interconnected cyber-physical systems. This connectivity enables organizations to monitor equipment, optimize production, identify failures, and automate operational processes.

However, the same interconnectivity that creates operational benefits also introduces cybersecurity risks. Industrial environments may contain legacy equipment, heterogeneous protocols, resource-constrained devices, remote connections, third-party systems, and safety-critical control processes. A compromise of one component can potentially propagate through connected devices and influence physical operations.



NIST emphasizes that IoT cybersecurity must account for the diversity of connected devices and the cybersecurity and privacy risks they introduce throughout their lifecycles. Its updated 2026 guidance also emphasizes cybersecurity activities that manufacturers can incorporate into the development and support of IoT products.

Traditional intrusion detection systems generally rely on signatures, statistical thresholds, or machine-learning models trained using centralized datasets. These approaches have limitations in IIoT environments. Signature-based mechanisms may fail against previously unseen attacks, while centralized machine learning requires organizations to transfer potentially sensitive operational data to a common location. Moreover, many conventional machine-learning models treat network observations independently and therefore fail to exploit the structural relationships among industrial devices.

Three emerging technologies provide complementary capabilities for addressing these limitations. First, **Graph Neural Networks (GNNs)** can model industrial networks as graphs in which devices, users, services, and communication sessions are represented as nodes and edges. This allows security models to capture both individual device behavior and relationships between entities.

Second, **Federated Learning (FL)** allows distributed participants to train a shared model while retaining their local training data. Research on federated cybersecurity for IoT has identified privacy preservation, communication efficiency, resource limitations, and malicious participants as important challenges.

Third, **Explainable Artificial Intelligence (XAI)** can expose the factors contributing to an AI-generated security decision. Explainability is particularly important in industrial environments because security analysts may need to understand why a device or communication session has been classified as malicious before taking corrective action.

Recent research increasingly considers federated learning, XAI, adversarial machine learning, concept-drift adaptation, and resource-efficient intelligence as interconnected requirements for trustworthy IoT and cyber-physical systems.

This paper therefore proposes an integrated framework that combines these three technologies rather than treating them as independent security mechanisms.

1.1 Research Problem

The central research problem addressed by this study is:

How can IIoT security systems simultaneously exploit network relationships, preserve the privacy of distributed industrial data, and provide interpretable cybersecurity decisions?

Existing approaches frequently optimize one or two of these objectives while neglecting the others. A centralized GNN may achieve strong relational modeling but require sensitive data aggregation. Federated learning can reduce the need to centralize raw data but may suffer from heterogeneous clients and malicious model updates. An accurate deep-learning detector may identify attacks but provide limited explanations to security analysts.

1.2 Research Objectives

The objectives of this research are:

1. To develop a graph-based representation of IIoT communication environments.
2. To design a GNN-based intrusion and anomaly detection mechanism.
3. To employ federated learning for privacy-preserving collaborative training.
4. To incorporate explainable AI for interpretable security decisions.



5. To design mechanisms for identifying anomalous or malicious federated clients.
6. To develop an evaluation methodology covering detection performance, privacy, efficiency, robustness, and explainability.
7. To identify challenges and future research directions for trustworthy AI-driven IIoT cybersecurity.

1.3 Research Contributions

The main conceptual contributions are:

- An integrated **GNN-FL-XAI architecture** for IIoT cybersecurity.
- A dynamic graph representation that captures device and communication relationships.
- A federated training mechanism designed for geographically or organizationally distributed industrial environments.
- An explainability layer for communicating the reasons behind security predictions.
- A security-oriented evaluation framework covering both machine-learning performance and operational constraints.
- A discussion of adversarial, privacy, heterogeneity, and concept-drift challenges.

BACKGROUND AND RELATED WORK

2.1 Cybersecurity Challenges in IIoT

IIoT environments differ from conventional enterprise networks because cybersecurity incidents can affect physical processes. A malicious modification to a control command may have consequences beyond information confidentiality and could influence production, equipment operation, or safety.

Important IIoT security challenges include:

- heterogeneous devices and protocols;
- legacy industrial equipment;
- limited computational resources;
- large volumes of telemetry;
- highly dynamic communication patterns;
- distributed ownership of data;
- insider and external threats;
- zero-day attacks;
- adversarial manipulation of machine-learning systems;
- strict latency requirements;
- privacy restrictions.

NIST's IoT cybersecurity guidance emphasizes that IoT products and their associated ecosystems require risk-based cybersecurity considerations rather than assuming a single security model applies universally.

2.2 Artificial Intelligence for IoT Security

Machine learning and deep learning have increasingly been investigated for intrusion detection, anomaly detection, malware analysis, authentication, and attack mitigation. A 2025 IEEE review describes AI-based IoT security approaches including anomaly and intrusion detection, authentication, access control, DDoS detection, and malware analysis.

However, conventional deep-learning models often process network observations as independent feature vectors. This assumption is problematic when an attack is characterized by relationships among multiple devices.

For example, consider a manufacturing network containing:

- PLC-A,
- PLC-B,



- sensor cluster S1,
- engineering workstation W1,
- gateway G1.

If W1 suddenly communicates with PLC-A and PLC-B using an unusual sequence of connections, the relationship between these entities may provide stronger evidence of compromise than the features of a single connection.

This motivates graph-based learning.

GRAPH NEURAL NETWORKS FOR IIOT SECURITY

3.1 Graph Representation

An IIoT environment can be represented by a graph:

$$G=(V,E,X)G=(V,E,X)$$

where:

- VV represents industrial entities;
- EE represents communication or interaction relationships;
- XX represents node and edge attributes.

A node may represent:

- sensor;
- PLC;
- actuator;
- gateway;
- workstation;
- server;
- industrial robot;
- network switch;
- edge device.

Edges may represent:

- TCP/UDP communication;
- industrial protocol communication;
- authentication;
- control commands;
- data exchange;
- device-to-device interaction.

A node feature vector may be defined as:

$$x_i=[p_i,b_i,r_i,c_i,t_i,s_i]x_i=[p_i,b_i,r_i,c_i,t_i,s_i]$$

where p_i represents packet characteristics, b_i bandwidth behavior, r_i protocol statistics, c_i communication characteristics, t_i temporal behavior, and s_i system-state information.

3.2 GNN-Based Detection

A basic graph convolution operation can be expressed as:

$$H^{(l+1)}=\sigma(D^{-\frac{1}{2}}A^{-\frac{1}{2}}H^{(l)}W^{(l)})H^{(l+1)}=\sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

where:

- $\tilde{A}=\tilde{A}+I$ is the adjacency matrix with self-loops;
- $\tilde{D}$ is the degree matrix;
- $H^{(l)}$ represents node embeddings at layer l ;
- $W^{(l)}$ is a trainable weight matrix;



- σ is an activation function.

The resulting node embeddings contain information not only about the node itself but also about its neighboring entities.

This is valuable for IIoT because attacks can produce relational abnormalities such as:

- unexpected device communication;
- lateral movement;
- unusual communication paths;
- compromised gateway behavior;
- coordinated scanning;
- abnormal command propagation.

3.3 Dynamic Graph Construction

Industrial networks are not static. Devices can join or leave, communication patterns change, and operational states vary.

Therefore, the proposed system uses time windows:

$$G_t = (V_t, E_t, X_t) \quad G_t = (V_t, E_t, X_t)$$

where t represents an observation interval.

The detector compares current graph behavior with historical patterns to identify anomalous structural and behavioral changes.

FEDERATED LEARNING ARCHITECTURE

4.1 Motivation

Centralizing industrial telemetry can introduce privacy, governance, and security concerns. Manufacturing plants may be unwilling or unable to transfer raw operational data outside their local environments.

Federated learning addresses this problem by allowing local clients to train models and send model updates rather than raw data.

Assume there are K industrial clients. Client k has local dataset D_k . The global optimization objective can be written as:

$$\min_{\theta} F(\theta) = \sum_{k=1}^K |D_k| F_k(\theta) \quad F(\theta) = \sum_{k=1}^K \frac{|D_k|}{|D|} F_k(\theta)$$

where F_k represents the local objective function.

A conventional federated aggregation process can be expressed as:

$$\theta_{t+1} = \sum_{k=1}^K \frac{n_k}{N} \theta_{t+1}^k \quad \theta_{t+1}^k = \sum_{k=1}^K \frac{n_k}{N} \theta_{t+1}^k$$

where:

- θ_t is the global model;
- θ_{t+1}^k is the locally trained model;
- n_k is the number of local training samples;
- N is the total number of samples.

4.2 IIoT Federated Architecture

The proposed architecture contains:

1. Industrial data sources
2. Local preprocessing module
3. Local graph construction
4. Local GNN training
5. Secure update generation
6. Federated aggregation server
7. Global model distribution



8. **XAI explanation engine**
9. **Security analyst interface**

Raw industrial telemetry remains within the participating organization.

This approach is consistent with the broader direction of federated cybersecurity research, where distributed learning is used to avoid unnecessary centralization of IoT security data.

PROPOSED GNN-FL-XAI FRAMEWORK

5.1 Overall Architecture

The proposed framework consists of five major layers.

Layer 1: Industrial Sensing Layer

This layer contains:

- sensors;
- PLCs;
- actuators;
- robots;
- industrial controllers;
- smart meters;
- gateways.

Telemetry is collected locally.

Layer 2: Edge Security Layer

Each industrial site contains an edge security node responsible for:

- traffic preprocessing;
- feature extraction;
- graph construction;
- local inference;
- local GNN training.

Layer 3: Federated Intelligence Layer

The federated coordinator:

- selects participating clients;
- distributes the global model;
- receives model updates;
- performs secure aggregation;
- identifies suspicious updates;
- generates a new global model.

Layer 4: Explainability Layer

The XAI component analyzes the GNN prediction and identifies important:

- nodes;
- edges;
- features;
- communication relationships;
- temporal behaviors.

Layer 5: Security Operations Layer

Security analysts receive:

- threat classification;
- confidence score;



- affected devices;
- important communication relationships;
- contributing features;
- recommended investigation priority.

PROPOSED METHODOLOGY

6.1 Data Preprocessing

Raw network traffic is transformed into structured observations.

Possible features include:

- source/destination IP;
- source/destination port;
- protocol;
- packet count;
- byte count;
- flow duration;
- packet-rate statistics;
- connection frequency;
- protocol commands;
- failed authentication attempts;
- temporal statistics.

Continuous variables can be normalized using:

$$x' = \frac{x - \mu}{\sigma}$$

Categorical features can be encoded using suitable embedding techniques.

6.2 Local Graph Generation

For each time window, a graph is generated.

Nodes represent devices, while edges represent communication events.

The edge weight may be defined as:

$$w_{ij} = \alpha f_{ij} + \beta b_{ij} + \gamma r_{ij}$$

where:

- f_{ij} is communication frequency;
- b_{ij} is bandwidth-related information;
- r_{ij} represents protocol or relationship characteristics;
- α, β, γ are weighting parameters.

6.3 Local GNN Model

The local model generates embeddings:

$$z_i = \text{GNN}(x_i, N_i)$$

where N_i represents the neighborhood of node i .

The classification function becomes:

$$\hat{y}_i = \text{Softmax}(W z_i + b)$$

Possible classes include:

- normal;
- denial-of-service;
- reconnaissance;
- scanning;
- brute force;
- spoofing;
- malware-related activity;



- lateral movement;
- command manipulation.

The exact classes should be determined by the selected benchmark dataset.

FEDERATED TRAINING

7.1 Local Training

Each industrial client receives the current global parameters:

θ_t

The client performs local optimization:

$$\theta_{k,t+1} = \theta_t - \eta \nabla F_k(\theta_t) \quad \theta_{k,t+1} = \theta_t - \eta \nabla F_k(\theta_t)$$

where η is the learning rate.

Multiple local epochs may be executed before model aggregation.

7.2 Secure Aggregation

Rather than exposing individual client updates to the coordinator, secure aggregation can be incorporated.

The objective is to obtain:

$$\theta_{t+1} = \text{Aggregate}(\theta_1, \theta_2, \dots, \theta_K) \quad \theta_{t+1} = \text{Aggregate}(\theta_1, \theta_2, \dots, \theta_K)$$

without revealing each individual update.

This reduces the exposure of individual clients, although federated learning itself should not be interpreted as automatically guaranteeing privacy. Recent research specifically highlights risks involving model-update leakage, poisoning, backdoors, Sybil behavior, and explanation instability in federated IoT intrusion detection.

7.3 Malicious Client Detection

An attacker could attempt to poison the global model by submitting manipulated updates.

Therefore, the framework introduces update anomaly detection.

For client update u_k , an anomaly score can be calculated as:

$$A_k = \lambda_1 D(u_k, \bar{u}) + \lambda_2 L_k + \lambda_3 C_k \quad A_k = \lambda_1 D(u_k, \bar{u}) + \lambda_2 L_k + \lambda_3 C_k$$

where:

- $D(u_k, \bar{u})$ measures deviation from the population update;
- L_k represents local model-loss behavior;
- C_k represents consistency across training rounds.

Clients with persistently abnormal updates can be flagged for additional validation.

EXPLAINABLE AI COMPONENT

8.1 Need for Explainability

A cybersecurity analyst cannot always act on an unexplained statement such as:

"Device 172.16.x.x is malicious."

Instead, the system should provide evidence such as:

- unusual communication with PLC-04;
- sudden increase in connection frequency;
- abnormal protocol usage;
- communication with previously unseen nodes;
- deviation from historical behavior.

XAI therefore becomes an operational component rather than simply a visualization feature.

8.2 Feature-Level Explanation

For prediction $f(x)$, a feature attribution method can estimate:

$$f(x) \approx \phi_0 + \sum_{j=1}^m \phi_j f(x) \quad f(x) \approx \phi_0 + \sum_{j=1}^m \phi_j f(x)$$

where ϕ_j represents the contribution of feature j .

Potentially important features may include:

- packet rate;
- connection duration;
- destination-port behavior;
- protocol commands;
- failed authentication;
- communication frequency.

8.3 Graph-Level Explanation

Because the proposed detector is graph-based, explanations should also identify important relationships.

For a detected anomaly, the system can identify:

$$E^* \subseteq E \wedge E^* \subseteq E$$

where E^* represents the subset of communication edges contributing strongly to the prediction.

Similarly:

$$V^* \subseteq V \wedge V^* \subseteq V$$

represents the important nodes.

An explanation can therefore state that an alert was strongly associated with a particular group of devices and communication relationships.

INTEGRATED ALGORITHM

Algorithm 1: Federated GNN-Based IIoT Intrusion Detection

Input: Local IIoT datasets $D_1, \dots, D_K$

Output: Global security model θ^*

1. Initialize global GNN parameters θ_0 .
2. Register participating industrial clients.
3. For each federated round t :
 1. Select participating clients.
 2. Send θ_t to selected clients.
 3. At each client:
 - collect local telemetry;
 - preprocess network data;
 - construct local graph G_k ;
 - train the GNN locally;
 - calculate local model update;
 - calculate local update-quality indicators.
 4. Send protected model updates to the federated coordinator.
 5. Evaluate update consistency.
 6. Identify potentially malicious client updates.
 7. Perform secure aggregation.
 8. Generate θ_{t+1} .
 9. Distribute θ_{t+1} .
4. At inference time:
 - construct the current graph;
 - obtain GNN prediction;
 - calculate threat confidence;
 - generate XAI explanation.
5. Present alert and explanation to the security analyst.
6. Continue monitoring for concept drift.

7. Repeat the federated training process periodically.

SECURITY THREAT MODEL

The proposed framework considers threats at three levels.

10.1 Network-Level Threats

Examples include:

- reconnaissance;
- scanning;
- denial-of-service;
- spoofing;
- unauthorized access;
- lateral movement;
- malicious command injection.

10.2 Model-Level Threats

Potential attacks include:

- data poisoning;
- model poisoning;
- backdoor attacks;
- adversarial examples;
- evasion attacks;
- model inversion.

10.3 Federated-System Threats

Federated environments introduce additional risks:

- malicious clients;
- Sybil clients;
- compromised edge nodes;
- update manipulation;
- communication interception;
- privacy leakage.

A security framework therefore needs to protect not only the industrial network but also the machine-learning pipeline.

EXPERIMENTAL EVALUATION FRAMEWORK

Because actual experimental results require implementation and execution on a defined dataset, this paper proposes an evaluation protocol rather than fabricating numerical findings.

11.1 Dataset Selection

The framework can be evaluated using one or more publicly available IIoT/ICS cybersecurity datasets.

Dataset selection should consider:

- industrial realism;
- attack diversity;
- network topology information;
- temporal information;
- class imbalance;
- availability of device identifiers;
- compatibility with graph construction.

The dataset should be divided into training, validation, and testing partitions without allowing temporal leakage between training and testing.

11.2 Baseline Models

The proposed GNN-FL-XAI system should be compared with:

1. Random Forest
2. Support Vector Machine
3. Multilayer Perceptron
4. CNN
5. LSTM
6. centralized GNN
7. federated MLP
8. federated CNN
9. federated GNN
10. proposed GNN-FL-XAI framework

The comparison should be performed under equivalent data and preprocessing conditions.

EVALUATION METRICS

12.1 Classification Metrics

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Because industrial security datasets can be highly imbalanced, accuracy alone should not be used as the primary metric.

12.2 False-Positive Rate

$$\text{FPR} = \frac{FP}{FP + TN}$$

A low false-positive rate is important because excessive alerts can overwhelm security operators.

12.3 Detection Latency

$$DL = T_{\text{alert}} - T_{\text{attack}}$$

where T_{attack} is the attack occurrence time and T_{alert} is the detection time.

12.4 Communication Overhead

For federated learning:

$$CO = \sum_{t=1}^R \sum_{k=1}^K S_{k,t}$$

where $S_{k,t}$ represents the size of the model update transmitted by client k during round t .

12.5 Explainability Metrics

The XAI component should be evaluated using:

- explanation fidelity;
- feature stability;
- explanation consistency;
- sparsity;



- analyst interpretability;
- computational overhead.

A useful explanation should identify influential features or graph structures without producing excessive irrelevant information.

12.6 Privacy and Robustness

Additional evaluation should consider:

- information leakage;
- membership-inference resistance;
- model poisoning resistance;
- backdoor attack success;
- robustness to malicious clients;
- robustness under non-IID data.

ABLATION STUDY

An important part of the proposed research is an ablation study.

The following configurations can be evaluated:

Configuration	GNN	FL	XAI
A	No	No	No
B	Yes	No	No
C	No	Yes	No
D	No	No	Yes
E	Yes	Yes	No
F	Yes	No	Yes
G	No	Yes	Yes
H	Yes	Yes	Yes

Configuration H represents the complete proposed architecture.

The purpose of the ablation study is not simply to maximize a single classification metric, but to determine the contribution of each component to detection, privacy, robustness, and interpretability.

HANDLING NON-IID INDUSTRIAL DATA

One of the most important challenges in federated IIoT learning is non-independent and non-identically distributed data.

Different industrial plants may have:

- different devices;
- different protocols;
- different production schedules;
- different attack distributions;
- different network topologies;
- different operating conditions.

Therefore:

$$P_k(X, Y) \neq P_j(X, Y) \quad P_k(X, Y) \neq P_j(X, Y)$$

for two different clients k and j .

A global model trained using ordinary averaging may therefore perform poorly for some clients.

Possible solutions include:

- FedProx-style optimization;
- personalized federated learning;



- clustered federated learning;
- adaptive aggregation;
- client-specific model heads;
- graph-level domain adaptation.

This issue is especially relevant because recent research on distributed IoT intelligence identifies data heterogeneity, robustness, energy efficiency, and secure real-time inference as important unresolved challenges.

CONCEPT DRIFT

Industrial networks change over time.

For example:

$$P_t(X, Y) \neq P_{t+1}(X, Y) \quad P_t(X, Y) \neq P_{t+1}(X, Y)$$

may occur because of:

- production changes;
- new equipment;
- firmware updates;
- maintenance;
- seasonal workloads;
- changes in network topology;
- newly emerging attacks.

A static security model can therefore become outdated.

The proposed architecture incorporates continuous monitoring of:

- prediction confidence;
- feature distributions;
- graph topology;
- communication patterns;
- client-level performance.

When significant drift is detected, federated retraining can be triggered.

EXPLAINABILITY AND HUMAN-CENTERED SECURITY

Explainability should be designed for security analysts rather than machine-learning researchers.

An alert should ideally contain:

Threat: Suspicious lateral movement

Affected device: PLC-07

Confidence: Model-generated confidence score

Important features:

- abnormal connection frequency;
- unusual destination;
- unexpected protocol activity.

Important relationships:

- workstation → gateway;
- gateway → PLC-07.

Temporal evidence:

- behavior deviated from historical baseline.

This structure enables the analyst to investigate the alert rather than merely accepting an opaque AI prediction.



NIST's smart-manufacturing roadmap identifies trustworthy, explainable, and reliable AI as important requirements for industrial environments, alongside challenges involving industrial data and heterogeneous sensing and control systems.

ZERO-TRUST INTEGRATION

The proposed GNN-FL-XAI architecture can operate alongside a zero-trust security architecture. NIST's zero-trust model rejects implicit trust based solely on network location and emphasizes explicit authentication and authorization for subjects and devices.

The proposed system can therefore provide behavioral risk information to a zero-trust policy engine.

For example:

$$\text{Risk}(d) = w_1A_d + w_2B_d + w_3G_d + w_4T_d$$

where:

- A_d = authentication anomaly;
- B_d = behavioral anomaly;
- G_d = graph anomaly;
- T_d = temporal anomaly.

A high-risk device may then be subjected to:

- additional authentication;
- restricted communication;
- network segmentation;
- increased monitoring;
- temporary isolation.

The AI model should inform such decisions rather than automatically overriding established industrial safety controls.

ADVANTAGES OF THE PROPOSED FRAMEWORK

The integrated architecture provides several potential advantages.

18.1 Relational Awareness

GNNs can capture relationships between industrial entities rather than analyzing isolated observations.

18.2 Data Privacy

Federated learning reduces the requirement to centralize raw industrial telemetry.

18.3 Distributed Intelligence

Detection can occur close to the data source, reducing dependence on a centralized processing location.

18.4 Explainability

XAI can help analysts understand why an alert was generated.

18.5 Adaptability

Federated retraining can support multiple geographically distributed industrial sites.

18.6 Defense-in-Depth

The architecture can complement conventional controls such as authentication, segmentation, monitoring, and zero-trust mechanisms.

LIMITATIONS

DOI: [10.22020/JOL-46.2.1](https://doi.org/10.22020/JOL-46.2.1)



Despite its potential, the proposed approach has important limitations.

First, GNNs can be computationally expensive for large and highly dynamic industrial graphs.

Second, federated learning introduces communication overhead and synchronization requirements.

Third, federated learning does not automatically guarantee privacy. Model updates themselves can leak information or be manipulated.

Fourth, non-IID industrial data can reduce global-model performance.

Fifth, explainability techniques may produce unstable explanations, particularly when models and local data distributions change.

Sixth, an attacker aware of the detection architecture could deliberately manipulate graph structures or model inputs.

Seventh, real industrial deployment requires compatibility with existing OT protocols, safety requirements, regulatory requirements, and operational constraints.

These limitations mean that the framework should be treated as a component of a broader industrial cybersecurity architecture rather than as a replacement for conventional security controls.

DISCUSSION

The central argument of this research is that AI-based IIoT cybersecurity should not optimize detection accuracy alone.

A practically useful system should simultaneously consider:

$\text{Security} = f(\text{Detection}, \text{Privacy}, \text{Robustness}, \text{Explainability}, \text{Latency}, \text{Resource Cost})$
 $\text{Security} = f(\text{Detection}, \text{Privacy}, \text{Robustness}, \text{Explainability}, \text{Latency}, \text{Resource} \setminus \text{Cost})$

A model achieving high classification accuracy but producing excessive false alarms may be operationally difficult to use. Similarly, a federated system that protects raw data but is highly vulnerable to poisoned updates does not provide sufficient security.

The GNN component addresses the structural nature of industrial networks. The FL component addresses distributed data and organizational privacy. The XAI component addresses transparency and human oversight.

Their combination therefore creates a multi-dimensional security architecture.

Recent literature supports the increasing importance of combining AI-driven detection with privacy-aware distributed learning and trustworthy explainability in IoT security. A 2026 survey specifically examining secure and explainable federated learning for IoT intrusion detection identifies model-update threats, heterogeneous clients, explanation instability, and privacy mechanisms as interconnected concerns.

The research direction is also consistent with broader developments in AI-driven IoT security, where lightweight, adaptive, privacy-preserving, and explainable models are increasingly emphasized.

FUTURE RESEARCH DIRECTIONS



Several directions can extend the proposed framework.

21.1 Personalized Federated GNNs

Instead of forcing all industrial sites to use exactly the same model, personalized models could learn site-specific behavior.

21.2 Continual Learning

Continual learning could enable the system to adapt to new attacks without complete retraining.

21.3 Digital Twins

Industrial digital twins could provide controlled environments for evaluating attack scenarios and AI responses.

21.4 Adversarially Robust GNNs

Future work should investigate graph perturbation attacks and adversarial manipulation of industrial communication structures.

21.5 Privacy-Preserving XAI

Explainability mechanisms themselves should be examined for information leakage.

21.6 Lightweight Edge Models

Model compression, pruning, quantization, and knowledge distillation could make the proposed system more appropriate for resource-constrained industrial edge devices.

21.7 Human-in-the-Loop Security

Future systems should allow analysts to validate AI-generated explanations and use analyst feedback to improve subsequent model updates.

CONCLUSION

This paper proposed an integrated AI-driven cybersecurity framework for the Industrial Internet of Things that combines Graph Neural Networks, Federated Learning, and Explainable Artificial Intelligence. The framework addresses three complementary requirements of modern industrial cybersecurity: relational threat detection, privacy-preserving distributed learning, and interpretable decision-making.

GNNs provide a mechanism for modeling relationships among industrial devices and communication flows. Federated learning enables multiple industrial environments to collaborate without requiring direct exchange of raw telemetry. XAI provides explanations that can help security analysts understand the features, devices, and relationships associated with an alert.

The proposed methodology also incorporates protection against malicious federated clients, non-IID data handling, concept-drift monitoring, and evaluation across detection, privacy, communication, robustness, and explainability dimensions.

The framework should not be considered a replacement for established industrial cybersecurity controls. Instead, it can function as an intelligent analytical layer supporting network monitoring, zero-trust policies, incident investigation, and risk-based response.

Future empirical research should implement the proposed architecture on representative IIoT/ICS datasets and evaluate it under realistic non-IID, adversarial, and resource-constrained conditions. Such experiments



would be necessary to determine whether the theoretical benefits of combining GNNs, federated learning, and XAI translate into measurable improvements in real industrial deployments.

REFERENCES

1. Fagan, M., Megas, K., Cuthill, B., Marron, J., & Hoehn, B. (2026). *Foundational Cybersecurity Activities for IoT Product Manufacturers*. NISTIR 8259 Rev. 1. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8259r1>.
2. NIST. (2019). *Considerations for Managing Internet of Things (IoT) Cybersecurity and Privacy Risks*. NISTIR 8228. National Institute of Standards and Technology.
3. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero Trust Architecture*. NIST Special Publication 800-207. National Institute of Standards and Technology.
4. Nguyen, D. C., Pham, Q.-V., Huynh-The, T., et al. (2022). Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for Internet of Things. *IEEE Internet of Things Journal*, 9(11), 8229–8249. <https://doi.org/10.1109/JIOT.2022.3150363>.
5. Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2022). Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Transactions on Industrial Informatics*, 18(5), 3501–3509. <https://doi.org/10.1109/TII.2021.3119038>.
6. Vogl, G., Cornelius, A., & Jia, X. (2026). *2026 Roadmap on Artificial Intelligence and Machine Learning for Smart Manufacturing*. National Institute of Standards and Technology.
7. *AI-Driven IoT Security: A Comprehensive Review on Current Trends and Future Perspectives*. (2025). 2025 International Conference on Circuit, Systems and Communication. IEEE. <https://doi.org/10.1109/ICCSC66714.2025.11135299>.
8. *AI-Powered IoT: A Survey on Integrating Artificial Intelligence With IoT for Enhanced Security, Efficiency, and Smart Applications*. (2025). *IEEE Access*, 13, 50296–50339. <https://doi.org/10.1109/ACCESS.2025.3551750>.
9. *Paradigm Shift Toward Distributed Learning in IoT Intelligence: A Comprehensive Survey of Opportunities and Challenges*. (2026). *IEEE Internet of Things Journal*, 13(10), 20128–20154. <https://doi.org/10.1109/JIOT.2026.3657733>.
10. *Secure and Explainable Federated Learning for IoT Intrusion Detection: A Comprehensive Survey*. (2026). *IEEE Open Journal of the Communications Society*, 7, 3650–3679. <https://doi.org/10.1109/OJCOMS.2026.3681580>.
11. *AI-Driven Anomaly Detection in IoT Networks: A Comprehensive Survey*. (2026). 2026 International Biennial Conference on Advances in Artificial Intelligence and Data Science. IEEE. <https://doi.org/10.1109/IBAAIDS66771.2026.11567421>.