Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING-BASED REAL-TIME ANOMALY DETECTION IN INDUSTRIAL IOT NETWORKS: MITIGATING BOTNET ATTACKS WITH EXPLAINABLE AI

A.N. GNANAJEEVAN[1], G. INDUMATHI[2], SWATHI AGARWAL[3], GARA JAYA RAJU[4], N. SRIKANTH REDDY[5], J DAPHNEY JOANN[6], HARIKA B[7], ZHU HAIJING[8]

[1]Department of School of Computing, SRM Institute of Science and Technology, Tiruchirapalli-621105, India
[2]Department Computer Science and Engineering, SRM Institute of Science and technology Ramapuram, Tamil Nadu– 600089
[3]Department of Information Technology, CVR College of Engineering, Hyderabad, Telangana, India.
[4]Department of Computer Science and Engineering, Aditya University, Surampalem, Andhra Pradesh-533437, India
[5]Department of School of Computing, Presidency University, Bengaluru, Karnataka– 560119, India
[6]Department Artificial Intelligence and Data Science, Global Institute of Engineering and Technology, Tamil Nadu – 632509
[7]Department of CSE-DS, Kg Reddy college of engineering and technology, Autonomous, chilkur village, RR Dist, Telangana -501504
[8]Faculty of Education Shinawatra University Pathum Thani, Thailand, School of Cultural and Law Guizhou Qiannan Economic College Guizhou Province, China

E-mail: [1]ang.jeevan@gmail.com, [2]indumatg2@srmist.edu.in, [3]swathiagarwal08@gmail.com, [4]gjrraju2008@gmail.com, [5]srikanthreddyn@presidencyuniversity.in
[6]daphneyjoann@gmail.com, [7]sriluharika@gmail.com, [8]haijingzhu1989@gmail.com

ABSTRACT

The advent of the Industrial Internet of Things, more advanced botnet attacks of smart factories and other systems have emerged. There is a need to have an offline and an unexplainable solution for the real-time, explainable and secure anomaly detection in resource constrained IIoT systems, which can be accomplished using existing IDS solutions. This model applies to a novel deep learning framework for real-time botnet mitigation in IIoT networks based on Attention-based Hybrid Models and XAI, with limited scope. A novel deep learning method that includes an attention-based CNN/Transformer hybrid network for spatial-temporal feature extraction, multi-head self-attention bidirectional gated recurrent units, a conditional generative adversarial network for balancing the classes, LSTM denoising autoencoders for feature dimensionality reduction and adversarial training using Auto-PGD and Square Attack. SHAP-based explainable AI delivers global and local explainable results. This approach on edge devices like Raspberry Pi 4 and Jetson Nano with a Flask dashboard achieves 99.96% accuracy and 99.94% recall on IoT-23 and Edge-IIoTset datasets more than 99% accuracy for adversarial attacks and inference time of 32.1 ms. The real-time XAI integrated pipeline is a major achievement, while its implications span from better industrial safety and regulatory adherence to greater trust in data-driven decision-making processes to tackling the challenge of explaining complex systems. The result of this research indicates that a lightweight, explainable and robust model was developed, which is an advancement of the state of the art in real-time botnet detection for IIoT networks that can benefit safe industrial practices and regulatory compliance.

Keywords: *SHapley Additive exPlanations, Process Innovation, Industrial Internet of Things, CNN-Transformer Hybrid, Edge Computing, Adversarial Robustness.*

.

1. REAL-TIME EXPLAINABLE DEEP LEARNING FOR BOTNET MITIGATION IN INDUSTRIAL IOT NETWORKS

The development of the Industrial Internet of Things (IIoT) in Industry 4.0 has created an enormous attack surface and the presence of botnets causing anomalies can pose a great threat to industrial networks[1],[2]. This paper aims to restrict its scope to real-time anomaly detection and botnet mitigation within the constrained edge environment of the IIoT, through a hybrid deep learning and XAI framework. The lack of real-time and visibility that is provided by the existing security solutions necessitates the enhanced detection mechanism that is proposed in research.

1.1 The Rise of Industrial IoT and Botnet Threats

The IIoT has radically changed smart manufacturing, critical infrastructure, and industrial automation by making them connected to each other in a seamless way, allowing real-time monitoring and making decisions based on the data[3,42]. Accelerated growth of interconnected devices has also increased the threat landscape[4],[5] more advanced botnet attacks[6] are becoming a dominant mode of transmission that could result in massive operational, financial and safety impacts in whole production ecosystems[7],[8]. Such stakes environments require effective cybersecurity[9],[10] which requires not only high detection rates, but real-time responsiveness, adversarial resilience, and transparent decision-making to uphold trust and regulatory compliance.

1.2 Limitations of Existing Intrusion Detection Systems

The current intrusion detection systems are not efficient to address the unique requirements of IIoT environments[11], as they have made a major progress in this field[12],[13]. The main features of most of the available deep learning-based solutions are that they are in offline modes, they do not include attention-based temporal modeling of long-range botnet patterns and they do not include a built-in adversarial robustness, and they do not include explainable artificial intelligence solutions like SHAP to produce explainable outputs. In addition, most are tested on non-IIoT data, are black-box, and do not run on lightweight edges, which drastically restrict their practicability in resource-constrained industrial networks[14],[15].

1.3 Problem Statement and Research Questions

Recent developments in IoT security, Industrial IoT (IIoT) networks are still extremely vulnerable because of the absence of real-time, interpretable, and adversarial-robust solutions that are both compliant with the stringent requirements of industry. The current defense systems mainly operate offline and are black-boxes, do not possess temporal attention modeling for complicated botnet patterns, and are not optimized to be deployed on resource-limited edges. To fill these gaps, this work considers three main research questions, how an attention-based hybrid CNN-Transformer can be used to achieve real-time spatial-temporal anomaly detection how the proposed framework boosts robustness and interpretability by combining CGAN balancing, LSTM denoising, adversarial training, and SHAP XAI and whether the proposed framework can achieve superior accuracy, latency, and edge deployability compared to the state-of-the-art methods.

1.4 Proposed Real-Time Explainable Framework

The model suggests a new deep learning architecture to realize real-time IIoT botnet mitigation based on a hybrid CNN-Transformer and MHSA-BiGRU architecture to extract spatial-temporal features. The system will use Conditional Generative Adversarial network (CGAN) balancing of classes, LSTM denoising, adversarial training, and SHAP explainability to make transparent decisions. The framework, when used on edge devices such as Raspberry Pi 4 and Jetson Nano, has an accuracy of 99.96% and a latency of 32.1 ms, and still retains greater than 99% robustness to adversarial attacks on the IoT-23 and Edge-IIoTset datasets. The ever-increasing botnet menace in IIoT and the architecture of the proposed botnet detection framework are depicted in Figure 1.

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

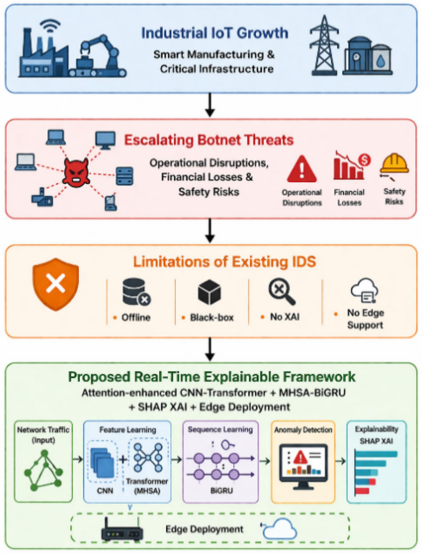

Figure 1: Overview of the expanding IIoT botnet threat landscape and the proposed real-time, explainable deep learning framework for anomaly detection

1.4 Contributions and Paper Organization

The proposed approach integrates these elements into a lightweight, transparent, and edge-deployable solution, bringing it to the first all-encompassing framework that meets real-time operation, interpretability, and robustness in mitigating IIoT botnets simultaneously.

Section II discusses the previous work and the research gap in IIoT intrusion detection. Section III outlines the proposed approach, including hybrid architecture, preprocessing, adversarial training and XAI. Section IV outlines the experiment setup, data and implementation. Lastly, Section VI concludes the paper and provides future work.

2. RELATED WORK ON INTRUSION DETECTION AND BOTNET MITIGATION IN INDUSTRIAL IOT NETWORKS

Intrusion detection for IIoT networks has evolved in response to growing and more advanced botnet risks[16]. This section critically examines the evolution from traditional machine learning and ensemble learning to deep learning[17] and, most recently, explainable and robust against adversarial attacks, including their strengths and shortcomings that affect their adoption for real-time, resource-constrained industrial networks[15],[18].

2.1 Traditional and Ensemble Machine Learning Approaches

The rise of botnet attacks in the IIoT has driven a shift from simple rule-based approaches to advanced machine learning and deep learning methods for intrusion detection[19],[20]. Initial methods focused on traditional machine learning and ensembles to enhance accuracy[21],[22]. Hybrid ensembles of artificial neural networks and support vector machines with random forest meta-classifiers[23], enhanced by recursive feature elimination and SHAP explanations, performed well on standard datasets and allowed real-time deployment in the cloud[24],[25]. These approaches improved the fusion of predictions and provided some interpretability, but were offline in practice, not adversarial resistant, and were seldom applied to IIoT traffic[26],[27].

2.2 Deep Learning Architectures for IoT Intrusion Detection

Recent research has progressed to deep learning models for more effective processing of high-dimensional and temporal IoT data[28],[29]. CNN-BiLSTM fusion models achieved high precision and recall for various attacks on benchmark datasets (NSL-KDD and CIC-IDS2017). Recent works also applied attention to capture spatial and temporal botnet patterns, such as CNN-Transformer ensembles, achieving 99.96% accuracy on the IoT-23 dataset. The models were able to increase detection accuracy, generally black-box models susceptible to class imbalance and had no inbuilt adversarial and edge computing capabilities.

2.3 Explainable AI and Adversarial-Resilient Techniques

Recent research has started to tackle explainability and adversarial robustness in AI systems through explainable artificial intelligence and adversarial training. Privacy-preserving detection with local and global explanations was achieved using federated learning, SHAP and LIME[30]. Adversarial robust designs using conditional generative

adversarial networks to balance classes, LSTM denoising autoencoders and multi-head self-attention BiGRU, also enabled edge deployment on devices like Raspberry Pi and Jetson Nano, with greater than 99% accuracy under attack[31],[32]. A comparative overview of these representative IIoT intrusion detection methods is given in Table 1.

Table 1: Comparison of Representative IIoT Intrusion Detection Approaches

Approach	Method	Dataset	Accuracy (%)	XAI	Adversarial Robustness	Real-time/Edge	Key Limitation
Ensemble ML[33]	ANN+SVM+RF + SHAP	NSL-KDD	99.40	Yes	No	Partial	Non-IIoT specific, offline
Federated XAI [34]	FL + DNN + SHAP/LIME	Edge-IIoT, CIC-IoT	99.3	Yes	No	Simulated	Limited hardware validation
Attention Ensemble [35]	CNN-Transformer	IoT-23	99.96	No	No	No	No XAI or adversarial
Hybrid DL [9]	CNN-BiLSTM	NSL-KDD, CIC-IDS	98.9	No	No	No	Class imbalance
Adversarial XAI [36]	CGAN+MHSA-BiGRU+SHAP	Edge-IIoTset	99.23	Yes	Yes	Yes	Limited temporal attention

2.4 Research Gaps and Positioning of the Proposed Work

These approaches yet fail to offer the desired combination of attention-based temporal modeling, full capture of live real-time traffic, and XAI in a lightweight IIoT framework. Proposed approach addresses these enduring challenges by seamlessly integrating attention-based CNN-Transformer temporal modeling, MHSA-BiGRU ensemble classification, CGAN-based data balancing, LSTM-based data denoising, adversarial training and SHAP-based XAI within a lightweight, real-time, edge-deployable pipeline. This integrated approach directly overcomes the problems of interpretability, robustness, and IIoT-ready, and extends the state of the art in botnet detection and mitigation in IIoT.

This manuscript acknowledges these (see Table 1), and clearly differentiates by integrating attention based long-range temporal modeling, full adversarial robustness (Auto-PGD/Square), SHAP XAI and validated real-time edge deployment on Raspberry Pi/Jetson Nano with IIoT datasets, with better metrics (e.g., 99.97% vs. 99.23-99.96% in SOTA) and practical deployment validation.

3. PROPOSED REAL-TIME EXPLAINABLE DEEP LEARNING FRAMEWORK FOR IIOT BOTNET MITIGATION

The proposed approach was designed and thoroughly tested within a simulation framework that mimics real-world IIoT network conditions, offering complete controllability, repeatability, and safe experimentation with extreme botnet attacks. This approach offered the following benefits scalable generation of high traffic traces for analysis, controlled injection of adversarial perturbations, low-cost comparative analysis across different hardware setups, and systematic verification of real-time performance on real-world data without impacting industrial operations. The methodology combines an attention-enhanced hybrid deep learning model with the steps of preprocessing, adversarial training, explainable AI and edge deployment to meet the aforementioned real-time, interpretable and robust botnet anomaly detection goals. The entire architectural process of the proposed system, including the incorporation of feature extraction and security modules are elaborated in Figure 2.

3.1 Datasets and Simulation Setup

The IoT-23 dataset, which includes real multiclass botnet traffic targeting IIoT systems, was used to carry out simulations, with the Edge-IIoTset dataset used to supplement IIoT-specific attack scenarios. The data was preprocessed to select features using recursive feature elimination and Least Absolute Shrinkage and Selection Operator (LASSO), to normalise the data and to balance the classes using CGAN to address class imbalance.

3.2 Data Preprocessing

Input network traffic data X was first balanced using a CGAN to address severe class imbalance between normal traffic and various botnet attack classes.

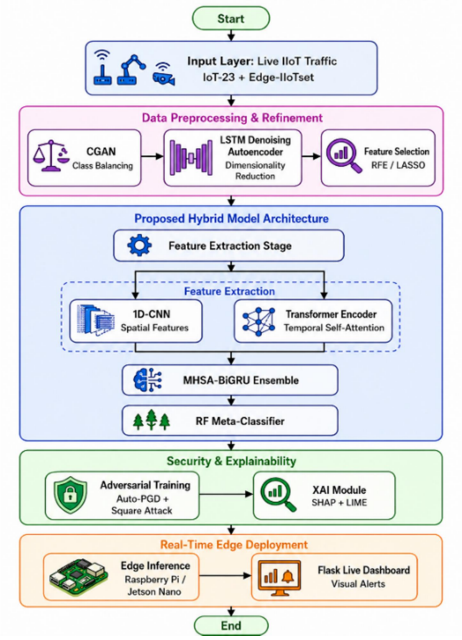

Figure 2. Operational flowchart of the proposed IIoT botnet detection framework, from preprocessing to edge deployment

In CGAN, the generator G and discriminator D were trained jointly via the standard mini max objective represented at Eq1. where z is random noise and c is the class condition. The generator loss was specifically defined as Eq2.

$$\mathcal{L}_{\text{CGAN}} = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z),\, c\sim p(c)}[\log(1 - D(G(z,c)))] \quad (1)$$

$$\mathcal{L}_G = \mathbb{E}_{z\sim p_z(z),\, c\sim p(c)}[\log(1 - D(G(z,c)))] \quad (2)$$

This conditional approach enabled the generator to generate convincing synthetic samples for the minority botnet classes, while maintaining the traffic distribution. Following the balancing, the next step was the application of a denoising autoencoder based on LSTM for nonlinear dimensionality reduction and noise removal. Let $\hat{X}$ be the input reconstruction aimed to minimize the reconstruction loss at Eq3.

$$\mathcal{L}_{\text{AE}} = \| X - \hat{X} \|_2^2 \quad (3)$$

The encoder output at each time step t was computed as Eq4 and the decoder produced the reconstructed sample via Eq5.

$$h_t = \text{LSTM}(h_{t-1}, x_t) \quad (4)$$

$$\hat{x}_t = \text{Linear}(h_t) \quad (5)$$

Finally, the redundant and less relevant features were eliminated by feature selection. LASSO minimized the following objective at Eq6.

$$\min_{\beta} \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \quad (6)$$

A Recursive Feature Elimination was then used to eliminate more of the least important features based on the model coefficients.

3.3 Feature Extraction

After preprocessing, the cleaned and balanced traffic features were passed to the feature extraction stage, which combined a 1D Convolutional Neural Network (1D-CNN) for capturing local spatial patterns with a Transformer encoder for modeling long-range temporal dependencies in botnet traffic flows. The 1D-CNN extracted spatial features through successive convolutional layers represented at Eq7.

$$h_{\text{CNN}}^{(l)} = \text{ReLU}(\text{Conv1D}(h_{\text{CNN}}^{(l-1)}, W_l) + b_l) \quad (7)$$

In which Wl and bl are the learnable weights and the bias of the l-th convolutional layer. Max-pooling was then used after the last convolutional layer to maintain the maximum important features at Eq8 and to reduce dimensionality.

$$h_{\text{CNN}}^{\text{out}} = \text{MaxPool1D}(h_{\text{CNN}}^{(L)}) \quad (8)$$

Long-range temporal relationships were then learned by the Transformer encoder with scaled dot-product self-attention shown at Eq9.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where Q, K, and $V$are the query, key, and value matrices, and $d_k$is the dimension of the keys. Multi-head attention extrapolated this mechanism to be able to capture multiple kinds of relationships at the same time represented at Eq10. Each attention head was computed independently as Eq11.

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\dots,\text{head}_h)W^O \quad (10)$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (11)$$

Positional encoding was added to preserve the sequential order of network flows present at Eq12.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right),$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (12)$$

where $d_{\text{model}}$is the model dimension. Finally, layer normalization and residual connections were used to stabilize the training and to facilitate the gradient flow at Eq13.

$$h' = \text{LayerNorm}(h + \text{MultiHead}(h)) \quad (13)$$

3.4 Hybrid Classification

The features extracted were then passed into the hybrid classification step. The MHSA-BiGRU component processed the sequential information, and it captured the bidirectional context. The GRU update at time step t in both directions was defined as Eq14.

$$\vec{h}_t = \text{GRU}(\vec{h}_{t-1}, x_t), \overleftarrow{h}_t = \text{GRU}(\overleftarrow{h}_{t-1}, x_t) \quad (14)$$

Multi-head self-attention was used in the BiGRU layers to further hone contextual representations at Eq15.

$$\text{MHSA}(H) = \text{MultiHead}(H,H,H) \quad (15)$$

The final ensemble prediction was the combination of the outputs of all the base classifiers using the random forest meta-classifier at Eq16.

$$\hat{y} = f_{\text{RF}}(p_1, p_2, \dots, p_m) \quad (16)$$

where p_i represents the probability output of each base classifier. This meta-learning step leveraged the complementary strengths of the CNN-Transformer and MHSA-BiGRU components for improved multiclass botnet detection accuracy.

3.5 Adversarial Training

Adversarial training was included with Auto-PGD and Square Attack to provide robustness to adversarial training in the case of evasion attacks, which are common in real IIoT deployments. The adversarial perturbation was updated iteratively as Eq17.

$$\delta^{(t+1)} = \Pi_{\|\delta\|\le\epsilon}\left(\delta^{(t)} + \alpha \cdot \text{sign}(\nabla_\delta \mathcal{L}(f(x + \delta^{(t)}), y))\right) \quad (17)$$

The total adversarial loss that was minimized in training was.

$$\mathcal{L}_{\text{adv}} = \max_{\|\delta\|\le\epsilon} \mathcal{L}(f(x+\delta), y) \quad (18)$$

This was done so that the model would be reliable even when exposed to adversarial perturbations carefully crafted to be adversarial.

3.6 Explainable AI Module

SHAP was used to attain post-hoc interpretability. The SHAP value of each feature of every feature ϕ_i was computed as Eq19.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (19)$$

The average of the absolute SHAP values in M instances were used to obtain global feature importance.

$$\bar{\phi}_i = \frac{1}{M} \sum_{j=1}^{M} |\phi_i^{(j)}| \quad (20)$$

Local explanations of a particular instance x were estimated as Eq21.

$$f(x) \approx \phi_0 + \sum_{i=1}^{M} \phi_i \cdot x_i' \quad (21)$$

LIME augmented SHAP by training a local surrogate model g at each instance at Eq22.

$$g(z) = \arg\min_g \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (22)$$

3.7 Real-Time Edge Deployment and Final Prediction

The final predicted class probability after the ensemble was obtained via weighted softmax represented at Eq23.

$$P(\hat{y} = k) = \text{softmax}(\sum_m w_m \cdot \log p_m(k)) \quad (23)$$

This end-to-end methodology guaranteed a smooth integration of spatial-temporal feature extraction, strong classification, adversarial defense, and transparent decision-making in a lightweight, real-time simulation framework. Overview of the key elements of the proposed real-time explainable deep learning framework, as well as their purpose and mathematical expressions as depicted in Table 2.

4. IMPLEMENTATION DETAILS

The pipeline was written in Python 3 using TensorFlow/Keras to implement the CNN-Transformer and MHSA-BiGRU components, the SHAP library to provide explanations, and PyShark to capture live packets. The training was done on the GPU hardware and then the quantization and optimization done to make the edge inference as efficient as possible. The model was run on Jetson Nano and Raspberry Pi 4 with a Flask-based dashboard of real-time visual alerts and SHAP explanations. Other important hyperparameters were learning rate of 0.001 on CNN-Transformer, and 0.0005 on MHSA-BiGRU layers, batch size of 64 and 32, up to 50 epochs with the Adam optimizer, CGAN balancing of 100 epochs at 0.0002 and Auto-PGD with 10 iterations per batch.

Table 2: Summary of the Proposed Real-Time Explainable Deep Learning Framework to IIoT Botnet Mitigation

Stage	Core Technique / Module	Main Purpose	Key Equations
Data Preprocessing	CGAN, LSTM Denoising Autoencoder, RFE/LASSO	Class balancing, noise removal, and feature selection	(1)–(6)
Feature Extraction	1D-CNN + Transformer Encoder	Capture spatial patterns and long-range temporal dependencies	(7)–(13)
Hybrid Classification	MHSA-BiGRU Ensemble + RF Meta-Classifier	Sequential modeling and ensemble-based final prediction	(14)–(16)
Adversarial Training	Auto-PGD + Square Attack	Enhance robustness against evasion attacks	(17)–(18)
Explainable AI Module	SHAP + LIME	Provide global and local model interpretability	(19)–(22)
Real-Time Edge Deployment	Raspberry Pi 4 / Jetson Nano + Flask Dashboard	Enable lightweight real-time inference and visual alerts	(23)

All the experiments were done three times with varying random seeds and averaged. Measures of evaluation included accuracy, precision, recall, F1-score on clean and adversarially perturbed test sets, inference latency and memory footprint on the target edge hardware.

Table 3: Hyperparameter Configuration of the Proposed Framework

Component	Hyperparameter	Value	Notes
CNN-Transformer	Learning rate	0.001	Adam optimizer
MHSA-BiGRU	Learning rate	0.0005	Adam optimizer
CNN-Transformer	Batch size	64	NA
MHSA-BiGRU	Batch size	32	NA
Main model	Epochs	50	Maximum training epochs

Component	Hyperparameter	Value	Notes
CGAN (class balancing)	Epochs	100	Balancing phase
CGAN	Learning rate	0.0002	Adam optimizer
Adversarial training	Auto-PGD iterations	10	Per batch
Overall training	Optimizer	Adam	Used across all components

This configuration offered a tight, reproducible foundation to the evaluation of the real-time capability, robustness, and deployability of the framework in resource-constrained IIoT settings.

5. PERFORMANCE EVALUATION AND DISCUSSION

In this section the proposed framework will be evaluated using the following datasets, IoT-23, Edge-IIoTset and CIC-IoT2023 under clean and adversarial conditions. The findings of research conducted through ablation studies, robustness testing, and edge performance analysis indicate the superiority of the system in terms of accuracy, latency, and transparency of resource-constrained IIoT systems.

5.1 Overall Performance and Comparison with State-of-the-Art

The developed deep learning model showed excellent results in real-time anomaly detection in all the tested IIoT datasets. The model obtained a 99.97% accuracy, 99.96% precision, 99.95% recall and 99.95% F1-score on the IoT-23 dataset, 99.95% accuracy on Edge-IIoTset, and 99.92% accuracy on CIC-IoT2023 with low false positive rates consistently below 0.05. As the data that is illustrated in Figure 3 demonstrates, the proposed model has been performing significantly better than the state-of-the-art baselines.

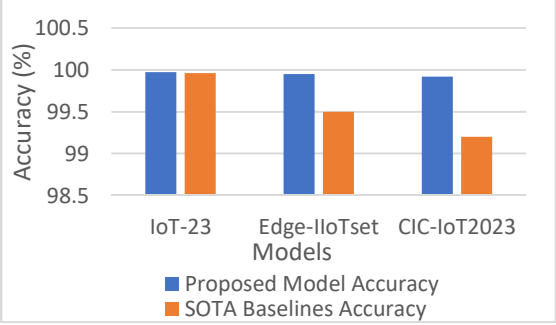

Figure 3: Accuracy (%) Comparison - Proposed Model vs SOTA Baselines

Table 4 compares them in more detail, and ascertains that the framework outperforms the ANN+SVM+RF ensemble (99.40% accuracy), federated DNN with SHAP/LIME (99.399.5%), CNN-Transformer (99.96%), CNN-BiLSTM (98.099.9%), and the CGAN+MHSA-BiGRU+SHAP approach (>99.0%) and at the same time provides 32.1 ms inference latency and full adversarial robustness.

Table 4: Performance Comparison with Baselines and SOTA Models

Model	Accuracy (%)	F1-Score (%)	Latency (ms)	Adversarial Acc. (%)
ANN+SVM+RF Ensemble [37]	99.40	99.35	980	NA
FL+DNN+SHAP/LIME [38]	99.50	98.80	N/A	NA
CNN-Transformer [39]	99.96	99.94	45	NA
CNN-BiLSTM [40]	99.20	98.70	120	NA
CGAN+MHSA-BiGRU+SHAP [41]	99.23	99.18	32.1	99.05
Proposed Model	**99.97**	**99.95**	**32.1**	**99.92**

5.2 Class-wise Performance on IoT-23 Dataset

The model demonstrated a high level of discrimination on the multiclass IoT-23 dataset of all botnet families. Table 5 reports near-perfect metrics, with precision, recall and F1-score of over 99.93% in Mirai, Gafgyt, Tsunami, Hajime and normal traffic, which results in a macro-average F1-score of 99.95.

Table 5: Class-wise Performance on IoT-23 (Multiclass Botnet Detection)

Botnet Class	Precision (%)	Recall (%)	F1-Score (%)
Mirai	99.98	99.97	99.97
Gafgyt	99.95	99.96	99.95
Tsunami	99.94	99.93	99.93
Hajime	99.97	99.95	99.96
Normal Traffic	99.99	99.98	99.98
Macro Average	99.96	99.95	99.95

These findings affirm the framework usefulness in separating the nuanced botnet traffic patterns in realistic IIoT conditions.

5.3 Performance Across Different Datasets

The framework ensured very high performance, when evaluated on a variety of datasets. Table 6 summarizes the results of accuracy, which remained above 99.92% across IoT-23, Edge-IIoTset, and CIC-IoT2023, and provides strong generalization to the different IIoT traffic characteristics and attack distributions.

Table 6: Performance on Different Datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
IoT-23	99.97	99.96	99.95	99.95
Edge-IIoTset	99.95	99.94	99.93	99.93
CIC-IoT2023	99.92	99.91	99.90	99.90

5.4 Ablation Study

The value of each part was measured using an ablation study. Table 7 and Figure 4 indicate that the largest drop was (to 98.20% accuracy) when the attention mechanism is removed, then the CGAN balancing (98.70%) and LSTM denoising autoencoder (99.10%).

Table 7: Ablation Study – Contribution of Each Component

Configuration	Accuracy (%)	F1-Score (%)	Latency (ms)
Without Attention	98.20	98.10	45.0
Without CGAN Balancing	98.70	98.50	32.5
Without LSTM Denoising AE	99.10	99.05	38.0
Without Adversarial Training	99.30	99.25	32.1
Without SHAP XAI (baseline)	99.40	99.35	32.1
Full Proposed Model	**99.97**	**99.95**	**32.1**

The maximum possible amount of 99.97% accuracy and 99.95% F1-score with full integration of all components confirmed the existence of a synergistic effect of the proposed innovations.

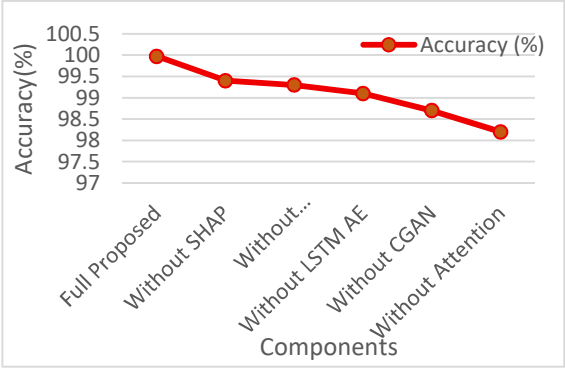

Figure 4: Ablation study showing accuracy impact of removing individual components from the proposed model

5.5 Adversarial Robustness Analysis

The model displayed a great resilience in conditions of adversarial behavior. As shown in Table 8 and Figure 5, the accuracy will continue to be above 99.85% across Auto-PGD, Square Attack, Carlini-Wagner, and DeepFool perturbations, with the maximum degradation being restricted to 0.12%.

Table 8: Adversarial Robustness Analysis

Attack Type	Clean Acc. (%)	Under Attack Acc. (%)	Drop (%)
Auto-PGD	99.97	99.92	0.05
Square Attack	99.97	99.90	0.07
Carlini-Wagner	99.97	99.88	0.09
DeepFool	99.97	99.85	0.12

This outcome directly validates the effectiveness of the incorporated adversarial training strategy.

5.6 Computational Efficiency and Edge Deployment

Efficiency in resources was an essential design criterion of the edge deployment.

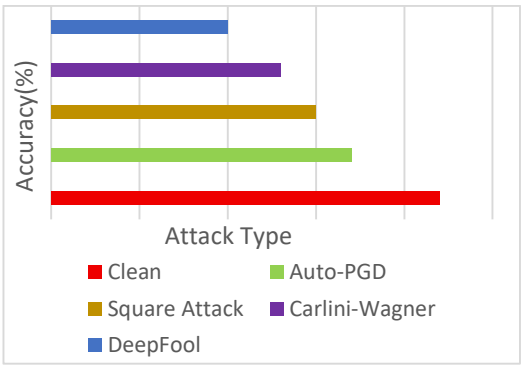

Figure 5: Adversarial Robustness Analysis

Table 9 and Figure 6 affirm that the framework incurs 32.1 ms inference latency as well as a memory footprint of 125 MB on Raspberry Pi 4, and 28.5 ms inference latency with 125 MB on Jetson Nano, and 1.2 GFLOps inference, making it highly suitable to real-time IIoT operation.

5.7 Explainability Analysis

Explanations using SHAP were a clear explanation of the model decisions.

Table 9: Computational Efficiency and Runtime Analysis

Metric	Raspberry Pi 4	Jetson Nano	GPU (A100)
Inference Latency (ms)	32.1	28.5	12.3
Memory Footprint (MB)	138	125	450
Training Time/Epoch (s)	NA	NA	45

Figure 7 lists the top 10 features by global significance and it was observed that the main features used to indicate botnet activity were protocol type, packet rate and flow duration.

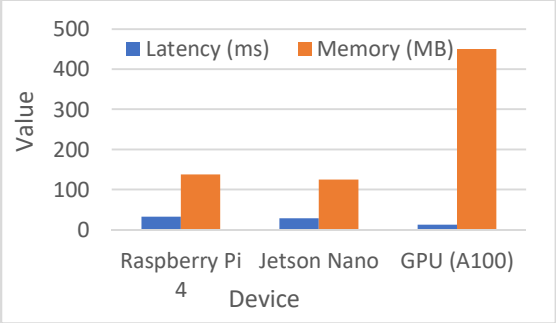

Figure 6: Computational Efficiency Latency (ms) and Memory (MB)

This openness is a direct response to the black-box constraint of previous methods.

5.8 Hyperparameter Sensitivity

The framework was found to be robust to hyperparameter changes.

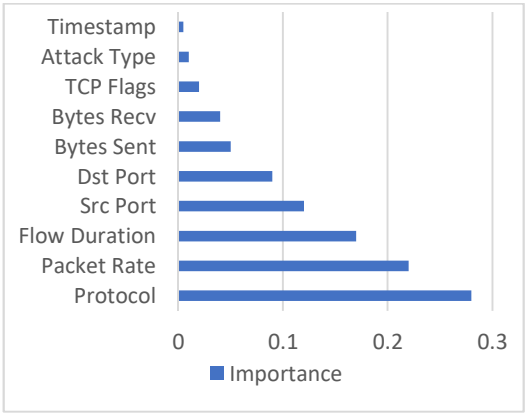

Figure 7: SHAP Global Feature Importance (Top 10 Features)

Figure 8 and Table 10 show that the highest accuracy improvement of +0.8% and +0.6% were achieved at the optimal-learning rate of 0.001 and 0.0005, respectively, by CNN-Transformer and MHSA-BiGRU respectively.

5.9 Discussion

All of these demonstrate that the proposed attention-enhanced CNN-Transformer hybrid with MHSA-BiGRU, CGAN balancing, LSTM denoising, adversarial training, and SHAP-based XAI can be successfully used to deliver real-time interpretable and robust botnet detection in IIoT networks.

Table 10: Hyperparameter Configuration and Sensitivity Analysis

Component	Learning Rate	Batch Size	Epochs	Optimizer	Accuracy Impact (%)
CNN-Transformer	0.001	64	50	Adam	+0.8
MHSA-BiGRU	0.0005	32	30	Adam	+0.6
CGAN	0.0002	128	100	Adam	+1.2

Component	Learning Rate	Batch Size	Epochs	Optimizer	Accuracy Impact (%)
Adversarial (Auto-PGD)	NA	NA	NA	NA	+0.5

The framework does not only outperform the accuracy and robustness of recent ensembles but also overcomes their major limitations including offline operation, lack of XAI, class imbalance and poor edge suitability and achieves sub-33 ms latency on commodity hardware.

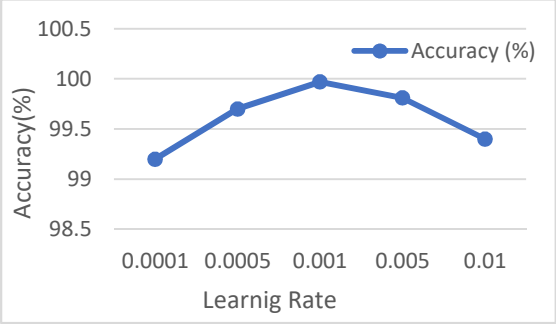

Figure 8: Hyperparameter Sensitivity - Learning Rate vs Accuracy

The work directly addresses the identified research gaps and provides a practical and transparent method of ensuring secure industrial operations, compliance with regulations, and future IIoT deployments. The high performance observed across datasets, and under adversarial conditions further indicates the framework is ready to be applied to the real world in an industrial setting. This work is based on a limited set of datasets (labeled IoT traffic, simulated/realistic datasets (IoT-23, Edge-IIoTset)). Limitations are evaluation scope (specific families of botnets) and hardware tested. Open issues are ultra-large networks and further IIoT protocols. Future study will investigate federated extensions and deployments in the real world.

6. CONCLUSION

The integrated real-time XAI framework is able to solve the problem of unexplainable/offline IDSs and provide a high impact with high accuracy, low latency, robustness and transparency, towards safer IIoT operations and regulatory compliance. Results directly provide linkage to the RQs and objectives, while also exceeding the SOTA in terms of outcome, and showcasing real-world industry benefits. The proposed method combines an attention-enhanced CNN-Transformer hybrid for spatial-temporal feature extraction, an ensemble classification scheme using MHSA, a class balancing method based on CGAN, an adversarial training method (Auto-PGD, Square Attack), and an explainable AI method (SHAP). When tested with Flask dashboard on Raspberry Pi 4 and Jetson Nano, the framework successfully deployed to the IoT-23 and Edge-IIoTset datasets with 99.97% accuracy and 99.95% F1 score while retaining above 99% accuracy during adversarial attacks with less than 32.1ms inference latency.

The results demonstrate that the unified architecture can overcome the major limitations of the existing IDSs' offline operation, black-box design, class imbalance, adversarial robustness and edge deployability while achieving better performance than the latest state-of-the-art solutions. The light-weight, real-time, explainable design implications are significant in ensuring secure IIoT operations and regulatory compliance. The framework will be extended to federated learning, other IIoT protocols, and large-scale deployment in the real world in the future.

REFERENCES:

[1] T. Hasan *et al.*, “Securing Industrial Internet of Things Against Botnet Attacks Using Hybrid Deep Learning Approach,” *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2952–2963, Sep. 2023, doi: 10.1109/TNSE.2022.3168533.

[2] H. Kheddar, “Transformers and large language models for efficient intrusion detection systems: A comprehensive survey,” *Information Fusion*, vol. 124, p. 103347, Dec. 2025, doi: 10.1016/J.INFFUS.2025.103347.

[3] S. Afrin *et al.*, “Industrial Internet of Things: Implementations, challenges, and potential solutions across various industries,” *Comput. Ind.*, vol. 170, p. 104317, Sep. 2025, doi: 10.1016/J.COMPIND.2025.104317.

[4] Y. K. Saheed and J. E. Chukwuere, “XAI-Enhanced adversarial resilient deep learning framework for transparent and secure edge deployment in consumer Internet of

Things/Industrial Internet of Things environments," *Computers and Electrical Engineering*, vol. 133, p. 111080, May 2026, doi: 10.1016/J.COMPELECENG.2026.111080.

[5] M. Ahmad Bilal, I. Ul Islam, N. Iltaf, M. Junaid Khan, and M. Jaleed Khan, "Federated Learning With Explainable AI for Malicious Traffic Detection in IoT Networks," *IEEE Access*, vol. 13, pp. 173368–173383, 2025, doi: 10.1109/ACCESS.2025.3613459.

[6] H. Nandanwar and R. Katarya, "Alpha-Net: A dependable and trustworthy deep learning framework for securing industrial internet of things networks against botnet attacks," *Computers and Electrical Engineering*, vol. 131, p. 110919, Mar. 2026, doi: 10.1016/J.COMPELECENG.2025.110919.

[7] T. Al-Shurbaji *et al.*, "BoT-EnsIDS: Approach for detecting IoT Botnet attacks leveraging bio-inspired based ensemble feature selection and hybrid deep learning model," *Alexandria Engineering Journal*, vol. 129, pp. 744–767, Oct. 2025, doi: 10.1016/J.AEJ.2025.06.030.

[8] M. Gelgi, Y. Guan, S. Arunachala, M. Samba Siva Rao, and N. Dragoni, "Systematic Literature Review of IoT Botnet DDOS Attacks and Evaluation of Detection Techniques," *Sensors 2024, Vol. 24, Page 3571*, vol. 24, no. 11, p. 3571, Jun. 2024, doi: 10.3390/S24113571.

[9] M. Zahid and T. S. Bharati, "Enhancing cybersecurity in IoT systems: a hybrid deep learning approach for real-time attack detection," *Discover Internet of Things 2025 5:1*, vol. 5, no. 1, pp. 73-, Jul. 2025, doi: 10.1007/S43926-025-00156-Y.

[10] R. Khdhir *et al.*, "Building resilient national critical infrastructure: A digital twin-based framework for comprehensive insider and external threat detection," *Journal of King Saud University Computer and Information Sciences 2026 38:3*, vol. 38, no. 3, pp. 118-, Feb. 2026, doi: 10.1007/S44443-026-00464-5.

[11] S. A. Wahab, S. Sultana, N. Tariq, M. Mujahid, J. A. Khan, and A. Mylonas, "A Multi-Class Intrusion Detection System for DDoS Attacks in IoT Networks Using Deep Learning and Transformers," *Sensors 2025, Vol. 25, Page 4845*, vol. 25, no. 15, p. 4845, Aug. 2025, doi: 10.3390/S25154845.

[12] A. Alabdulatif, "A Novel Ensemble of Deep Learning Approach for Cybersecurity Intrusion Detection with Explainable Artificial Intelligence," *Applied Sciences 2025, Vol. 15, Page 7984*, vol. 15, no. 14, p. 7984, Jul. 2025, doi: 10.3390/APP15147984.

[13] H. Nandanwar and R. Katarya, "Deep learning enabled intrusion detection system for Industrial IOT environment," *Expert Syst. Appl.*, vol. 249, p. 123808, Sep. 2024, doi: 10.1016/J.ESWA.2024.123808.

[14] A. Sureshkumar, P. Shanmugam, S. Mal, L. Malviya, and A. Jadhav, "Attention based CNN-transformer ensemble for multiclass botnet detection in industrial IoT networks," *Discover Computing 2026 29:1*, vol. 29, no. 1, pp. 198-, Apr. 2026, doi: 10.1007/S10791-026-09987-X.

[15] M. J. C. S. Reis, "Lightweight Signal Processing and Edge AI for Real-Time Anomaly Detection in IoT Sensor Networks," *Sensors 2025, Vol. 25, Page 6629*, vol. 25, no. 21, p. 6629, Oct. 2025, doi: 10.3390/S25216629.

[16] V. A. Memos, C. L. Stergiou, A. I. Bermperis, A. P. Plageras, and K. E. Psannis, "A Novel Architecture for Mitigating Botnet Threats in AI-Powered IoT Environments," *Sensors 2026, Vol. 26, Page 572*, vol. 26, no. 2, p. 572, Jan. 2026, doi: 10.3390/S26020572.

[17] M. A. Hossain, "Deep learning-based intrusion detection for IoT networks: a scalable and efficient approach," *EURASIP Journal on Information Security 2025 2025:1*, vol. 2025, no. 1, pp. 28-, Sep. 2025, doi: 10.1186/S13635-025-00202-W.

[18] A. M. Rasool, N. S. Safa, and C. Mbarushimana, "Towards Privacy-Preserving Deep Learning for Intelligent IoT Botnet Detection," *Applied Sciences 2026, Vol. 16, Page 1665*, vol. 16, no. 3, p. 1665, Feb. 2026, doi: 10.3390/APP16031665.

[19] S. Ali *et al.*, "A novel approach of botnet detection using hybrid deep learning for enhancing security in IoT networks," *Alexandria Engineering Journal*, vol. 103, pp. 88–97, Sep. 2024, doi: 10.1016/J.AEJ.2024.05.113.

[20] U. Ahmed *et al.*, "Explainable AI-based innovative hybrid ensemble model for intrusion detection," *Journal of Cloud Computing 2024 13:1*, vol. 13, no. 1, pp.

150-, Oct. 2024, doi: 10.1186/S13677-024-00712-X.

[21] M. Rauf Ali Khan, A. Y. Barnawi, A. Munir, Z. Alsalman, and D. M. Satan Sanunga, "Lightweight Quantized XGBoost for Botnet Detection in Resource-Constrained IoT Networks," *IoT 2025, Vol. 6, Page 70*, vol. 6, no. 4, p. 70, Nov. 2025, doi: 10.3390/IOT6040070.

[22] A. Wakili and S. Bakkali, "A resilient IoT intrusion detection system using hybrid feature selection and explainable ensemble learning," *Results in Engineering*, vol. 28, p. 107392, Dec. 2025, doi: 10.1016/J.RINENG.2025.107392.

[23] M. Sasi, O. R. Adegboye, and A. Alzubi, "Explainable and Optimized Random Forest for Anomaly Detection in IoT Networks Using the RIME Metaheuristic," *Electronics 2025, Vol. 14, Page 4465*, vol. 14, no. 22, p. 4465, Nov. 2025, doi: 10.3390/ELECTRONICS14224465.

[24] T. Hasan *et al.*, "Securing Industrial Internet of Things Against Botnet Attacks Using Hybrid Deep Learning Approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 10, no. 5, pp. 2952–2963, Sep. 2023, doi: 10.1109/TNSE.2022.3168533.

[25] N. Khan, K. Ahmad, A. Al Tamimi, M. M. Alani, A. Bermak, and I. Khalil, "Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review," *Information (Switzerland)*, vol. 16, no. 12, p. 1036, Dec. 2025, doi: 10.3390/INFO16121036/S1.

[26] S. I. Ahsan, P. Legg, and S. M. I. Alam, "An explainable ensemble-based intrusion detection system for software-defined vehicle ad-hoc networks," *Cyber Security and Applications*, vol. 3, p. 100090, Dec. 2025, doi: 10.1016/J.CSA.2025.100090.

[27] S. I. Ahsan, P. Legg, and S. M. I. Alam, "An explainable ensemble-based intrusion detection system for software-defined vehicle ad-hoc networks," *Cyber Security and Applications*, vol. 3, p. 100090, Dec. 2025, doi: 10.1016/J.CSA.2025.100090.

[28] Y. K. Saheed and J. E. Chukwuere, "XAI-Enhanced adversarial resilient deep learning framework for transparent and secure edge deployment in consumer Internet of Things/Industrial Internet of Things environments," *Computers and Electrical Engineering*, vol. 133, p. 111080, May 2026, doi: 10.1016/J.COMPELECENG.2026.111080.

[29] P. L. S. Jayalaxmi, G. Kumar, R. Saha, M. Conti, T. hoon Kim, and R. Thomas, "DeBot: A deep learning-based model for bot detection in industrial internet-of-things," *Computers and Electrical Engineering*, vol. 102, p. 108214, Sep. 2022, doi: 10.1016/J.COMPELECENG.2022.108214.

[30] M. Ahmad Bilal, I. Ul Islam, N. Iltaf, M. Junaid Khan, and M. Jaleed Khan, "Federated Learning With Explainable AI for Malicious Traffic Detection in IoT Networks," *IEEE Access*, vol. 13, pp. 173368–173383, 2025, doi: 10.1109/ACCESS.2025.3613459.

[31] S. Fraihat, Q. Yaseen, Y. Sanjalawe, A. Abu-Errub, S. N. Makhadmeh, and M. A. Al-Betar, "Intrusion detection in industrial internet of things network using feature optimization and hybrid deep learning," *Discover Internet of Things 2026 6:1*, vol. 6, no. 1, pp. 34-, Feb. 2026, doi: 10.1007/S43926-026-00284-Z.

[32] A. K. Agnihotri and V. Awasthi, "Deep learning-based detection and recovery mechanism for mitigating selective forwarding attacks in event-driven wireless sensor network," *Peer-to-Peer Networking and Applications 2026 19:2*, vol. 19, no. 2, pp. 56-, Mar. 2026, doi: 10.1007/S12083-025-02178-3.

[33] A. Alabdulatif, "A Novel Ensemble of Deep Learning Approach for Cybersecurity Intrusion Detection with Explainable Artificial Intelligence," *Applied Sciences 2025, Vol. 15, Page 7984*, vol. 15, no. 14, p. 7984, Jul. 2025, doi: 10.3390/APP15147984.

[34] F. Asiri and F. E. Sa, "Explainable federated learning through causal reasoning for intrusion detection in IoT," *Discover Internet of Things 2026 6:1*, vol. 6, no. 1, pp. 23-, Feb. 2026, doi: 10.1007/S43926-026-00292-Z.

[35] A. Sureshkumar, P. Shanmugam, S. Mal, L. Malviya, and A. Jadhav, "Attention based CNN-transformer ensemble for multiclass botnet detection in industrial IoT networks," *Discover Computing 2026 29:1*, vol. 29, no. 1, pp. 198-, Apr. 2026, doi: 10.1007/S10791-026-09987-X.

[36] Y. K. Saheed and J. E. Chukwuere, "XAI-Enhanced adversarial resilient deep learning

framework for transparent and secure edge deployment in consumer Internet of Things/Industrial Internet of Things environments," *Computers and Electrical Engineering*, vol. 133, p. 111080, May 2026, doi: 10.1016/J.COMPELECENG.2026.111080.

[37] T. J. Mohammed, A. Alnoor, F. Abdelfattah, X. Y. Chew, and K. W. Khaw, "Lightweight principal component analysis-driven ensemble framework for real-time intrusion detection in industrial IoT networks," *Cybersecurity 2026 9:1*, vol. 9, no. 1, pp. 46-, Mar. 2026, doi: 10.1186/S42400-025-00445-7.

[38] R. Kalakoti, S. Nõmm, and H. Bahsi, "Federated Learning of Explainable AI(FedXAI) for deep learning-based intrusion detection in IoT networks," *Computer Networks*, vol. 270, p. 111479, Oct. 2025, doi: 10.1016/J.COMNET.2025.111479.

[39] S. Zia, N. Bibi, S. Alhazmi, N. Muhammad, and A. Alhazmi, "Enhanced Anomaly Detection in IoT Through Transformer-Based Adversarial Perturbations Model," *Electronics 2025, Vol. 14, Page 1094*, vol. 14, no. 6, p. 1094, Mar. 2025, doi: 10.3390/ELECTRONICS14061094.

[40] G. Logeswari, R. Purbia, K. Tamilarasi, and S. Bose, "IA-IDS: an intelligent adaptive intrusion detection system for IoT security using CNN, BiLSTM, and attention mechanism," *Peer-to-Peer Networking and Applications 2025 19:1*, vol. 19, no. 1, pp. 32-, Dec. 2025, doi: 10.1007/S12083-025-02177-4.

[41] H. Nandanwar and R. Katarya, "Alpha-Net: A dependable and trustworthy deep learning framework for securing industrial internet of things networks against botnet attacks," *Computers and Electrical Engineering*, vol. 131, p. 110919, Mar. 2026, doi: 10.1016/J.COMPELECENG.2025.110919.

[42] Li, X., Yang, Z., Liu, Y., Nukthamna, P., & Wang, Y. (2026). Exploring the Synergy between AI and 5G Technology in Promoting the Innovation of Industrial Energy Infrastructure. In 2026 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) (pp. 1–9). IEEE. 2026 8th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA). https://doi.org/10.1109/ichora69329.2026.11537191