International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I11P112

Volume 73 Issue 11, 147-164, November 2025

Original Article

A Multi-Model Digital Twin-Based Intrusion Detection System: Integrating Autoencoder, DNN, and Anomaly Detection for Robust Cyber Threat Identification

Saifur Rahman

Electrical Engineering, Najran University, Najran, Saudi Arabia

[1]*Corresponding Author : srrahman@nu.edu.sa*

Received: 05 August 2025 Revised: 03 November 2025 Accepted: 10 November 2025 Published: 25 November 2025

Abstract - *The rapid proliferation of Industrial Internet of Things (IIoT) systems has introduced unprecedented cybersecurity challenges that require advanced detection and response mechanisms. This paper presents a novel cybersecurity framework that leverages Digital Twin (DT) technology to create a comprehensive security solution for IIoT environments. The proposed framework addresses critical limitations in existing approaches by integrating three interconnected models within a unified digital twin architecture that provides real-time monitoring, intelligent anomaly detection, and automated threat classification. The methodology creates a dynamic virtual replica of the physical IIoT network, enabling proactive security management through continuous behavioral analysis and predictive threat assessment. The framework was evaluated using the Edge-IIoT dataset containing 63 features across 15 attack classes plus normal traffic. Experimental results demonstrate exceptional performance with a classification accuracy of 99.97%, Precision (Pr) of 99.77%, Recall (Re) of 99.64%, and F1-score (Fs) of 99.70% for multiclass threat classification. The anomaly detection component achieved a Pr of 99.74% and ROC-AUC of 90.79%, effectively distinguishing between normal and malicious network behaviors. The reconstruction-based anomaly detection mechanism showed clear separation between normal traffic (mean reconstruction error: 0.006) and attack traffic (mean reconstruction error: 1.289), validating the framework's ability to identify previously unseen threats. These results demonstrate the effectiveness of the proposed digital twin-based approach in providing comprehensive cybersecurity protection for IIoT environments, significantly outperforming traditional security solutions while enabling real-time threat response and proactive incident management.*

Keywords - *Anomaly Detection, Cybersecurity, Digital Twin, Edge-IIoT Dataset, Industrial Internet of Things (IIoT), Intrusion Detection System (IDS), Machine Learning, Threat Classification.*

1. Introduction

In an era of pervasive digital connectivity, cybersecurity has become a cornerstone of operational integrity, safety, and trust across IoT, industrial systems, and critical infrastructure. The exponential growth of interconnected devices-projected to exceed 29 billion by 2030-has expanded attack surfaces, enabling threats ranging from data breaches to sabotage Industrial Control Systems (ICS). The consequences are severe: financial losses (averaging $4.45 million per breach in 2023, physical infrastructure damage, and risks to human safety. High-profile incidents, such as ransomware attacks on healthcare systems and the grid disruptions underscore the tangible societal and economic impact of cyber vulnerabilities. Despite advancements, cybersecurity faces persistent challenges due to the complexity and dynamism of cyber-physical ecosystems. Evolving attack vectors (e.g., AI-driven malware, zero-day exploits) demand real-time, adaptive defenses capable of distinguishing sophisticated threats from legitimate operations. Resource-constrained IoT devices struggle with computational overhead, while industrial systems require ultra-low latency and near-zero false positives to avoid catastrophic failures. Moreover, the "black-box" nature of deep learning models complicates trust and accountability-a critical gap in domains like healthcare and critical infrastructure, where explainability is non-negotiable. To address these challenges, research has pivoted toward synergistic methodologies that combine real-time simulation, hybrid AI architectures, and human-interpretable analytics. Digital twin-based approaches have significantly advanced anomaly detection and cybersecurity in IoT and industrial systems. A novel digital twin architecture for Industrial IoT (IIoT) anomaly detection, as described in [1], integrates simulation and operational data to enable real- time monitoring and predictive diagnostics, achieving adaptive and accurate detection. Similarly, [2] introduces DTITD, a framework combining digital twin technology with self-

attention-based deep learning to detect insiderthreats, leveraging transformer models to enhance accuracy and reduce false positives. [3] proposes a digital twin-based security framework using MiniCPS and a stacked ensemble classifier, achieving 92.7% accuracy.

Additionally, [4] presents a framework for cyber-physical systems that correlate physical and simulated data for real-time anomaly detection. The integration of digital twins The LSTM-CNN models in [5] achieve over 97% accuracy for IoT anomaly detection, while [6] combines digital twins with federated learning for privacy-preserving cyberthreat detection, achieving 98.12% accuracy. A digital twin-based Intrusion Detection System (IDS) using a Kalman filter and SVM, as in [7], achieve 98–99% accuracy for ICS protection.

CyberDefender, introduced in [8], employs a multi-layered defense for a digital twin-based Industrial Cyber-Physical Systems (ICPS) with a GRU-LSTM model, achieving 98.96% accuracy. Lastly, [9] presents TwinSec-IDS, an attention-based BiGRU-LSTM model for IIoT, achieving 99.41% accuracy with SHAP-based interpretability. Hybrid deep learning models have also shown promise in enhancing cybersecurity. A stacking ensemble of CNN, LSTM and GRU models for Internet of Medical Things (IoMT) intrusion detection, as described in [10], achieves 99.4% accuracy with low false positives. Similarly, [11] proposes a CNN-LSTM-GRU hybrid model for IIoT security, achieving 99.56. Explanable AI (XAI) approaches enhance transparency in IoT anomaly detection. In [12], seven XAI techniques, including SHAP and LIME, are employed to achieve over 99% accuracy on MEMS and N-BaIoT datasets, improving trust and diagnostics.

Ensemble learning approaches further improve IDS performance. A survey in [13] (2009–2020) highlights that ensemble methods like bagging, boosting, and stacking outperform single classifiers by improving accuracy and reducing false positives. In [14], an ensemble-based IDS using the GTCS dataset combines diverse machine learning classifiers for enhanced accuracy. DIS-IoT, introduced in [15], integrates four deep learning models, achieving high Accuracy on ToN IoT, CICIDS2017, and SWaT datasets. Similarly, [16] proposes a stacked ensemble IDS with Random Forest, Gradient Boosting and Extra Trees, achieving 99.3% accuracy for IoT networks. In [17], a hybrid feature selection approach for ensemble models achieves over 98% accuracy on NSL-KDD and CIC-IDS2017 datasets. Advanced deep learning models, such as the transformer-based framework in [18], leverage self-attention mechanisms to achieve 99.84% accuracy on the BoTIoT dataset for IoT intrusion detection. Additionally, [19] presents an Adaptive Adversarial Transformer for manufacturing anomaly detection, achieving over 97% accuracy with robust temporal feature extraction. Other specialized approaches address unique challenges. In [20], Deep learning and transfer learning improve anomaly detection detection and failure classification in smart manufacturing by 11.6%. A context-aware collaborative intelligence The framework in [21] reduces communication overhead by 85% in IoT networks while maintaining accuracy. For IoMT [22] proposes a blockchain-enabled federated learning framework, enhancing accuracy, data integrity, and privacy. MADness, introduced in [23], combines statistical, machine learning, and signal processing techniques for robust anomaly detection. Lastly, [24] presents an anomaly-based IDS using the Junction Tree Algorithm, achieving 88.4% accuracy on Unix-based systems.

While the reviewed literature demonstrates significant Advances in IoT anomaly detection and cybersecurity through digital twin technologies, ensemble methods, and deep learning approaches, several gaps remain in achieving truly integrated and dynamic security frameworks for Industrial IoT environments. Most existing digital twin-based solutions focus on specific aspects of security monitoring or anomaly detection without establishing a A comprehensive virtual replica that can simultaneously perform real-time monitoring, intelligent threat classification, and automated incident response. Furthermore, there is limited research on frameworks that integrate multiple interconnected models within a unified digital twin architecture to provide holistic security management and proactive threat mitigation. To address these limitations, this paper proposes a novel Cybersecurity Framework for Industrial Internet of Things (IIoT) environments that leverage the Digital Twin (DT) concept. The key contributions and features of the proposed work includes:

- Dynamic Digital Twin Architecture: Development of a Comprehensive digital twin that creates a dynamic, virtual replica of the physical IIoT network with three interconnected models designed to collectively providecomprehensive understanding and intelligent management of the IIoT security landscape.
- Real-time Monitoring and Anomaly Detection: Implementation of continuous monitoring capabilities with advanced anomaly detection mechanisms that operate within the digital twin environment using dynamic data representation and processing of real-time IIoT security data streams.
- Intelligent Threat Classification and Response: Development of an automated threat classification system that intelligently categorizes security incidents and enables rapid response decision-making through predictive analytics capabilities.
- Proactive Security Management: Enhancement of overall security posture through proactive incident response capabilities that facilitate comprehensive cybersecurity management for IIoT environments via the synergisticIntegration of digital twin technology with advanced threat detection and response mechanisms.

The remainder of this paper is structured as follows. Section 2 discusses the dataset details, Section 3 details the methodology, Section 4 presents the results and discussion, and Section 5 offers a conclusion.

2. Dataset Detail

Our study utilizes the Edge-IIoTset, a cutting-edge dataset built for intrusion detection within IIoT settings. This dataset is robust, featuring a broad spectrum of network traffic and system logs that capture numerous IoT and IIoT-specific attack patterns. This dataset was selected because of its comprehensive nature, encompassing both benign and malicious traffic, and its focus on modern attack vectors makes it exceptionally well-suited for evaluating advanced machine learning models in this domain. Table 1 provides a comprehensive breakdown of the different attack types present in the dataset, including their counts and percentages in the overall dataset, as well as their distribution between the training and testing sets. The dataset comprises a total of 2,219,201 samples and 63 features, utilizing approximately 3330 MB of memory. Notably, there are no missing values, but 815 duplicate rows were identified. The features are composed of 43 numerical and 20 categorical types. The table clearly shows 15 unique attack types, with ”Normal” traffic being the most prevalent, accounting for 72.80% of the total samples. This indicates a significant imbalance, with an attack imbalance ratio of 1614.03, highlighting that ”Normal” traffic instances are vastly more numerous than any single attack type. The distribution of each attack type, both in terms of absolute counts and percentages, is consistent across the full dataset, the training set, and the test set, suggesting a stratified split was likely applied to maintain the original proportions of each attack type in both subsets.

2.1. Dataset Acquisition and Preprocessing

The foundation of the digital twin lies in its ability to accurately mirror the real-world IIoT environment. For this research, the Edge IIoT dataset was utilized, chosen for its comprehensive representation of diverse network traffic and device

Table 1. Attack type distribution

Attack Type	Count	Percentage	Train Count	Test Count	Train Percentage	Test Percentage
Normal	1615643	72.80	1292514	323129	72.80	72.80
DDoS UDP	121568	5.48	97254	24314	5.48	5.48
DDoS ICMP	116436	5.25	93149	23287	5.25	5.25
SQL ›injection	51203	2.31	40962	10241	2.31	2.31
Password	50153	2.26	40122	10031	2.26	2.26
Vulnerability scanner	50110	2.26	40088	10022	2.26	2.26
DDoS TCP	50062	2.26	40050	10012	2.26	2.26
DDoS HTTP	49911	2.25	39929	9982	2.25	2.25
Uploading	37634	1.70	30107	7527	1.70	1.70
Backdoor	24862	1.12	19890	4972	1.12	1.12
Port Scanning	22564	1.02	18051	4513	1.02	1.02
XSS	15915	0.72	12732	3183	0.72	0.72
Ransomware	10925	0.49	8740	2185	0.49	0.49
MITM	1214	0.05	971	243	0.05	0.05
Fingerprinting	1001	0.05	801	200	0.05	0.05

Behaviors encompassing both normal operational data and various cyberattack scenarios are prevalent in IIoT. The dataset’s characteristics, including its size, feature types, and distribution of attack classes, were thoroughly analyzed prior to model training.

Upon acquisition, the raw dataset underwent a rigorous preprocessing pipeline to prepare it for machine learning model ingestion:

- Feature Separation: The dataset was partitioned into features (X) and the target variable (y), representing the Attack type.
- Categorical Feature Encoding: All non-numeric (object) features were transformed into a numerical representation using Label Encoder. This step ensures compatibility with subsequent machine learning algorithms.
- Target Label Encoding: The Attack type labels were also numerically encoded using Label Encoder, facilitating supervised learning tasks.
- Data Splitting: The processed data was then split into training and testing sets (80% training, 20% testing) using a test split. A random state=42 was set for reproducibility, and the stratify parameter was applied to the target variable to ensure that the original class distribution of attack types was maintained in both training and testing partitions, which is crucial for balanced model training and evaluation in imbalanced datasets.
- Feature Scaling: To normalize the range of numerical features and prevent features with larger magnitudes from disproportionately influencing model training, the Standard Scaler was applied to both the training and testing sets. This transformation ensures that each feature contributes equally to the models.

- Normal Traffic Isolation: A crucial step for unsupervised learning models (behavioral model and anomaly detector) was the isolation of a subset of data explicitly identified as 'Normal' traffic from the training set. This Xnormal subset was exclusively used to train the behavioral models, ensuring they learned the characteristics of benign system operation without exposure to anomalous patterns.

3. Methodology: Digital Twin for IIoT Cybersecurity

The proposed cybersecurity framework for Industrial Internet of Things (IIoT) environments leverages the Digital Twin (DT) concept to create a dynamic, virtual replica of the physical IIoT network and its constituent devices. This digital twin facilitates real-time monitoring, anomaly detection, and automated threat classification, thereby enhancing the overall security posture and enabling proactive incident response. The core methodology involves the construction and integration of three interconnected models within the digital twin architecture, designed to collectively provide a comprehensive understanding and intelligent management of the IIoT security landscape. The structural flow of data representation and processing within the digital twin system is illustrated in Figure 1.

3.1. Digital Twin Model Components

The intelligence of the digital twin is derived from three specialized machine learning models that operate synergistically to detect, classify, and mitigate threats. These models collectively form the analytical core of the digital twin, providing a multi-layered security assessment.

3.1.1. Behavioral Model (Autoencoder)

A deep autoencoder served as the primary behavioral model within the digital twin. This unsupervised neural network architecture was specifically designed to learn the compact, latent representations and normal operational patterns of the IIoT network and device traffic.

Architecture: The autoencoder comprised an encoder and a decoder. The encoder compressed the high-dimensional input features into a lower-dimensional latent space of 32 neurons. The decoder subsequently reconstructed the original input from this latent representation. Both encoder and decoder utilized Dense layers with relu activation functions, interspersed with Batch Normalization layers to stabilize training and accelerate convergence, and Dropout layers (with rates of 0.2) to mitigate overfitting. The final output layer of the decoder employed a linear activation function to allow for the reconstruction of continuous feature values. A detailed architecture summary is shown in Table 2.

Training: The autoencoder was exclusively trained on the X_{normal} subset of scaled training data. The model was compiled with the Adam optimizer (learning rate $\alpha = 0.001$) and optimized for Mean Squared Error (MSE) as the loss function, reflecting the goal of accurate data reconstruction.

Training epochs were set to 100 with a batch size of 128, and a validation split of 0.2 was used for monitoring generalization. Early Stopping (patience=10, monitor='val_loss') and ReduceLROnPlateau factor=0.2, patience=5, monitor='val_loss', min lr=0.0001) callbacks were employed to prevent overfitting and optimize the learning rate.

Table 2. Architecture of the "behavior autoencoder" model

Layer (type)	Output Shape	Param \#
Input Layer (InputLayer)	(None, 62)	0
Encoder Dense	(None, 128)	8,064
Encoder BatchNormalization	(None, 128)	512
Encoder Dropout	(None, 128)	0
Encoder Dense	(None, 64)	8,256
Encoder BatchNormalization	(None, 64)	256
Encoder Dropout	(None, 64)	0
Encoder Dense	(None, 32)	2,080
Decoder Dense	(None, 64)	2,112
Decoder BatchNormalization	(None, 64)	256
Decoder Dropout	(None, 64)	0
Decoder Dense	(None, 128)	8,320
Decoder BatchNormalization	(None, 128)	512
Decoder Dropout	(None, 128)	0
Decoder output (Dense)	(None, 62)	7,998
Total params: 38,366 (149.87 KB)		
Trainable params: 37,598 (146.87 KB)		
Non-trainable params: 768 (3.00 KB)		

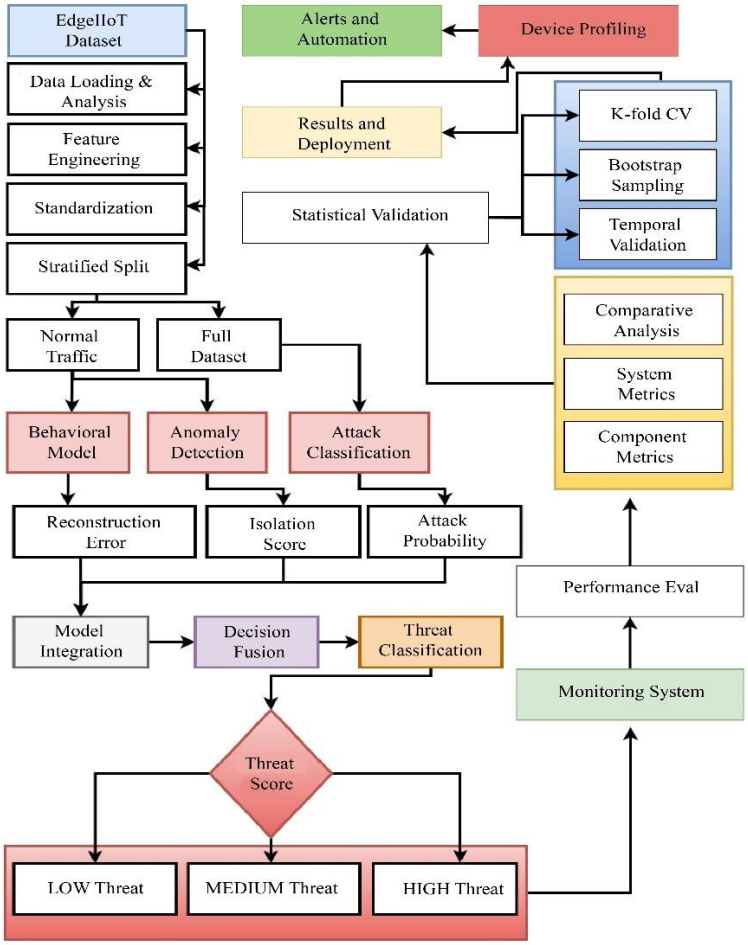

Fig. 1 EdgeIIoT threat detection framework architecture. Complete system architecture showing the data processing pipeline from the EdgeIIoT dataset through feature engineering, model training (behavioral modeling, anomaly detection, attack classification), statistical validation, and deployment with integrated monitoring and alert systems.

Anomaly Detection Mechanism: Post-training, the autoencoder's ability to reconstruct data was leveraged for anomaly detection. A reconstruction error for an input sample x is calculated as the mean squared difference between x and its reconstructed output ˆx: $E = \frac{1}{n}\sum_{i=1}^{n}(Xi - \widehat{Xi})^2$.

A reconstruction threshold was established by calculating the reconstruction errors for all samples in X_{normal} and setting the threshold as the 95th percentile of these errors. During real-time operation, any incoming data sample yielding a reconstruction error above this pre-defined threshold is flagged as a behavioral anomaly, indicating a significant deviation from learned normal patterns. This capability is vital for detecting novel or zero-day attacks that do not conform to known attack signatures.

Anomaly Detector (Isolation Forest)

A complementary Isolation Forest model was integrated for robust statistical anomaly detection. This unsupervised ensemble learning algorithm is highly effective in high-dimensional datasets for explicitly isolating outliers rather than profiling normal data points.

Configuration: The Isolation Forest was configured with n estimators=200 (number of base estimators), contamination=0.1 (an estimate of the proportion of outliers in the data, guiding the model's decision boundary), max samples='auto' (automatically setting the number of samples to draw from the training data), and random state=42 for reproducibility.

Training: The Isolation Forest model was trained solely on the Xnormal dataset, enabling it to learn the typical distribution and structure of benign IIoT traffic.

Anomaly Detection Mechanism: In operation, the Isolation Forest assigns an anomaly score (or decision function value) to each data point. Samples classified as anomalies are typically indicated by a prediction of -1, while normal samples

receive a prediction of 1. This provides an independent validation of unusual activity, complementing the behavioral anomaly detection.

Attack Classifier (Deep Neural Network)

To provide specific threat identification, a Deep Neural Network (DNN) was developed as the attack classifier.

This supervised learning model was trained to distinguish between various types of cyberattacks present in the EdgeIIoT dataset (e.g., DDoS, DoS, SQL Injection, Ransomware) and normal traffic.

Architecture: The DNN consisted of a sequence of Dense layers with decreasing neuron counts (256, 128, 64, 32), each employing relu activation functions. To enhance generalization and prevent overfitting, BatchNormalization layers were applied after each Dense layer, followed by Dropout layers with rates between 0.2 and 0.3. The input layer matched the scaled feature dimensions of the dataset, and the final output layer utilized a softmax activation function to provide probability distributions across all identified attack classes.

Training: The classifier was trained on the full X train scaled and y train datasets, ensuring exposure to both normal and various attack patterns. Sparse categorical crossentropy was used as the loss function, appropriate for integer-encoded labels, and the Adam optimizer ($\alpha = 0.001$) was employed. Training epochs were set to 10 with a batch size of 1000, with EarlyStopping and ReduceLROnPlateau callbacks similar to the autoencoder training. Detailed architecture is explained in Table 3.

Threat Identification: Upon inference, the DNN outputs a vector of probability scores for each possible attack type. The class with the highest probability is identified as the predicted attack, along with its associated confidence score (the maximum probability).

Table 3. Architecture of the ”attack classifier” model

Layer (type)	Output Shape	Param \#
Classifier Dense	(None, 256)	16,128
Classifier BatchNormalization	(None, 256)	1,024
Classifier Dropout	(None, 256)	0
Classifier Dense	(None, 128)	32,896
Classifier BatchNormalization	(None, 128)	512
Classifier Dropout	(None, 128)	0
Classifier Dense	(None, 64)	8,256
Classifier BatchNormalization	(None, 64)	256
Classifier Dropout	(None, 64)	0
Classifier Dense	(None, 32)	2,080
Classifier Dropout	(None, 32)	0
Classifier output (Dense)	(None, 15)	495
Total params: 61,647 (240.81 KB)		
Trainable params: 60,751 (237.31 KB)		
Non-trainable params: 896 (3.50 KB)		

4. Digital Twin Operational Phase: Real-time Monitoring and Integrated Threat Assessment

The operational efficacy of the digital twin was demonstrated through a simulated real-time monitoring environment. This phase simulates the continuous influx of IIoT network traffic, subjecting each data point to a multi-layered security assessment by the trained digital twin models. Real-time processing requirements are met through optimized pipeline design (Figure 2), achieving end-to-end latency under 177ms while maintaining high detection accuracy across all threat categories.

Simulated Real-time Data Feed: New data samples were continuously ingested by sequentially drawing records from the unseen X test dataset at a defined sample interval (e.g., 1.0 seconds). This approach realistically mimics the asynchronous and continuous flow of data from physical IIoT devices in a live environment. Each simulated sample carries its actual label (actual label) for subsequent ground-truth comparison.

Integrated Sample Analysis: Upon ingestion, each incoming sample underwent concurrent and integrated analysis by all three digital twin models:

- Behavioral Anomaly Check: The sample was first passed through the trained autoencoder. Its reconstruction error E was calculated and compared against the pre-determined reconstruction threshold. If E > reconstruction threshold, the sample was flagged as a behavioral anomaly.
- Statistical Anomaly Check: Simultaneously, the Isolation Forest model evaluated the sample. The decision function output provided an anomaly score, and the model’s predict method indicated whether the sample was classified as an outlier (prediction of 1).

- Attack Classification: The sample was also fed into the deep neural network classifier, which predicted the specific attack type (or 'Normal') and provided a confidence score (the maximum probability across classes).

Dynamic Threat Level Assessment: A sophisticated, rule-based logic integrated the outputs from all three models to determine an overall threat level (LOW, MEDIUM, HIGH) for each analyzed sample. This assessment prioritized potential risks as follows:

- HIGH Threat: Assigned if the attack classifier's confidence in a predicted attack was exceptionally high (> 0.9), or if both a behavioral anomaly and a statistical anomaly were concurrently detected, irrespective of the classifier's confidence. This signifies a strong indication of malicious Activity or a highly unusual system state.
- MEDIUM Threat: Assigned if either a behavioral anomaly or a statistical anomaly was detected, or if the attack classifier predicted an attack with moderate confidence (> 0.8 but ≤ 0.9). This indicates suspicious activity requiring immediate attention.
- LOW Threat: Assigned otherwise, signifying normal or benign traffic, or very low confidence in any detected anomalies/attacks.

Security Event Generation and Automated Response Actions: When a sample's threat level was determined to be MEDIUM or HIGH, a formal security alert was triggered. These alerts were immediately logged as security events, capturing critical metadata such as timestamp, predicted attack type, confidence levels, and flags for detected behavioral and statistical anomalies.

Critically, automated recommended response actions were dynamically generated based on the specific threat level and identified attack cha– HIGH Threat: Actions included "ISOLATE DEVICE", "BLOCK TRAFFIC", and "ALERT SECURITY TEAM".

- MEDIUM Threat: Actions included "INCREASE MONITORING" and "LOG INCIDENT".
- Specific Anomaly Indicators: If a behavioral anomaly was detected, "CHECK DEVICE CONFIG" was recommended. If the attack confidence exceeded 0.9, "UPDATE FIREWALL RULES" was also suggested.

This automated generation of response actions simulates the autonomous capability of the digital twin to facilitate rapid mitigation of threats, significantly reducing the Mean Time To Respond (MTTR) and minimizing potential damage and operational downtime in a real IIoT deployment.

The data buffer (a deque with maxlen=1000) maintained a rolling window of recent analysis results, providing a continuous snapshot of the system's security posture.

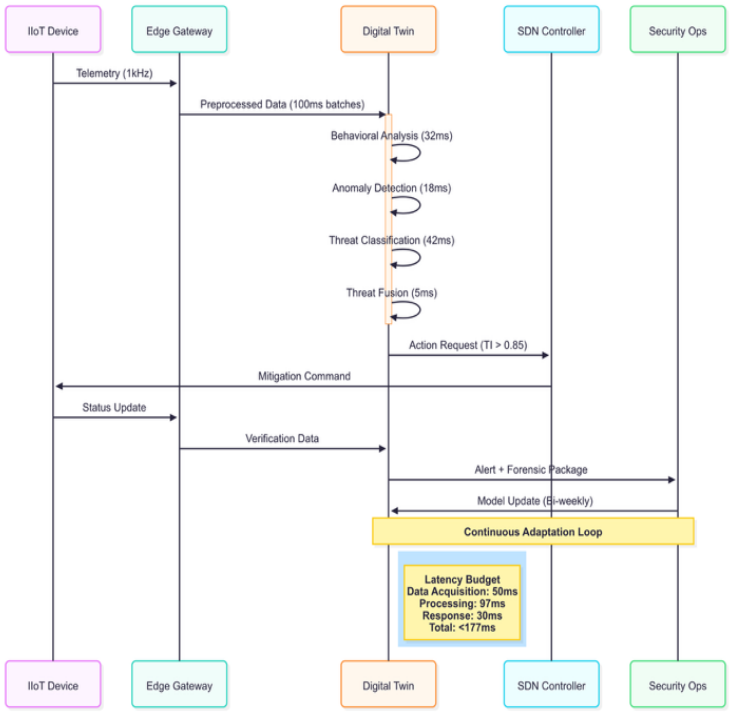

Fig. 2 Real-time processing sequence diagram temporal flow of threat detection operations showing latency requirements and processing times across system components, from IoT device telemetry (1kHz) through digital twin analysis to security operations response, with a continuous adaptation loop maintaining sub-177ms total response time

4.1. Device Behavioral Profiling

To enhance the contextual awareness and granularity of threat detection, a device behavioral profiling component was incorporated into the digital twin architecture. This module aimed to create baseline operational profiles for distinct devices within the IIoT environment, allowing for more specific and accurate anomaly detection. The system maintains effectiveness against evolving threats through continuous adaptation mechanisms (Figure 3), automatically detecting concept drift and updating models while preserving operational stability.

Profile Creation Methodology: By identifying device-specific identifiers within the dataset (e.g., columns containing 'device' or 'node' in their name), the system iteratively processed data associated with each unique device. For each device, a profile was generated that captured key aggregated metrics:

- Total samples: Total data points observed for the device.
- Normal samples, attack samples: Counts of normal vs. attack traffic originating from or destined for the device.
- Feature means: Mean values for all numerical features associated with that device's traffic, providing a statistical fingerprint of its typical operational parameters.
- Attack types: A value count of all observed attack types linked to the device, offering a historical threat landscape.

Contribution to Digital Twin: These profiles serve as individual digital identities for physical IIoT devices. In a real-world scenario, by comparing current device behavior against its own established normal profile (rather than a generalized system-wide normal), the digital twin can perform more accurate and tailored anomaly detection, significantly reducing false positives and enabling more precise threat localization. This context-aware anomaly detection is a critical advantage for complex IIoT infrastructures comprising heterogeneous devices.

5. Results

The results section evaluates the performance of a digital twin-based IDS comprising three distinct models: a behavioral model using an autoencoder, a Deep Neural Network (DNN)-based attack classifier, and an anomaly detector. The subsequent subsections provide detailed analyses of each model's effectiveness in detecting intrusions, followed by a comprehensive summary of overall system performance. The first subsection examines the autoencoder-based behavioral model's reconstruction and error metrics across 40 epochs, highlighting its ability to learn normal behavior patterns.

The second subsection assesses the DNN-based classifier's near-perfect classification accuracy (99.97%) across 15 attack categories, supported by training dynamics and confusion matrix insights. The third subsection analyzes the anomaly detector's discriminative power (ROCAUC of 0.9079) and anomaly score distributions, identifying class-specific deviations. Finally, the overall results subsection synthesizes classification, anomaly detection, and training metrics, alongside real-time security alerts, offering a holistic view of the IDS's robustness and areas for refinement in a digital twin framework.

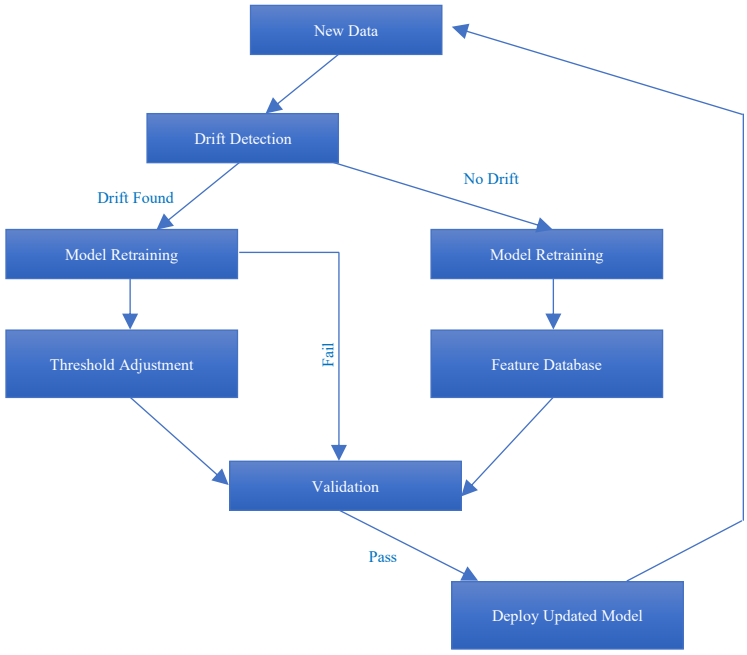

Fig. 3 Adaptive model management workflow continuous learning system flowchart illustrating drift detection mechanisms, model retraining procedures, and validation processes that enable dynamic threshold adjustment and automated model updates to maintain detection accuracy in evolving threat landscapes

5.1. Performance Evaluation of the Autoencoder-Based IDS

The behavioral model autoencoder-based IDS leverages performance metrics visualized in Figures 4-7 to evaluate its effectiveness in detecting anomalies in digital twin data. The training process, shown in Figure 4, tracks the Mean Squared Error (MSE) for training and validation over 40 epochs. Both losses start high (0.8 and 0.7 MSE, respectively) but converge to approximately 0.1 MSE by epoch 10, indicating the model effectively learns the data distribution with minimal overfitting, as the close alignment of curves suggests robust generalization through techniques like dropout or L2 regularization Complementing this, Figure 5 presents a boxplot of reconstruction error distributions (on a log scale) across attack classes such as Port Scanning, DDoS TCP.

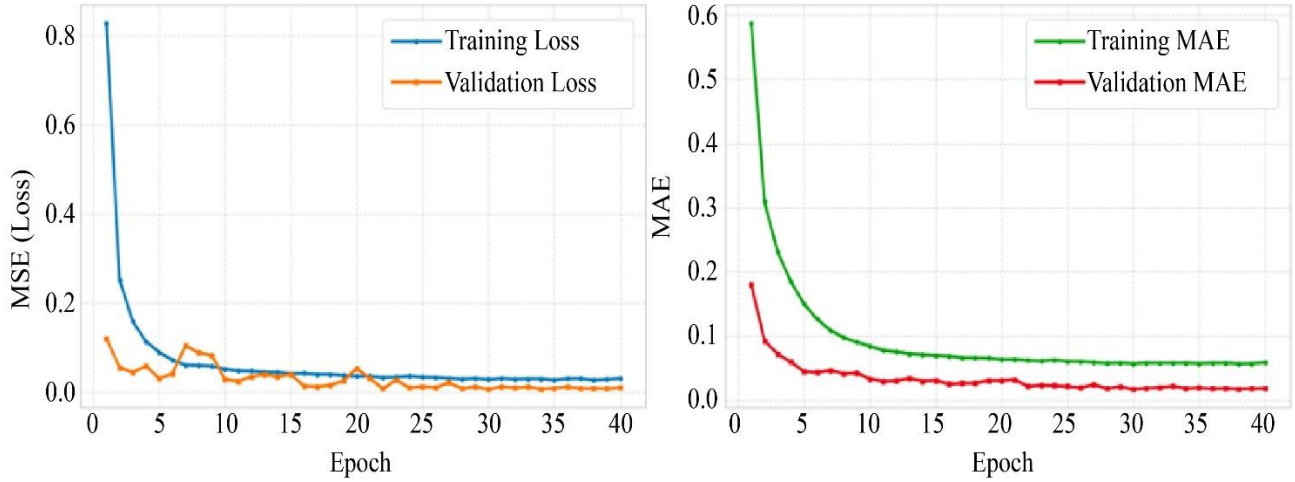

Fig. 4 Training and validation loss (MSE) and MAE of the autoencoder over 40 epochs, indicating robust model learning and generalization

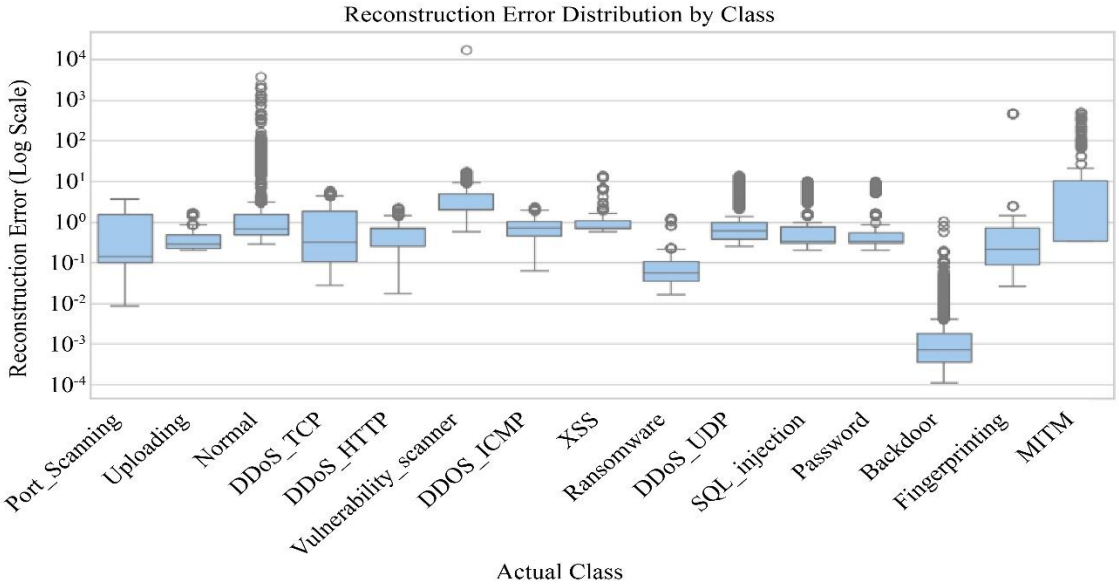

Fig. 5 Boxplot of reconstruction error (log scale) by attack class, highlighting varied error distributions with outliers in MITM and Fingerprinting, aiding anomaly detection

MITM and Fingerprinting. The interquartile ranges and outliers, particularly in MITM and Fingerprinting (extending to 104), reveal varying reconstruction difficulties, enabling the IDS to distinguish attack types based on error magnitude. This variation is critical for identifying rare or complex attack patterns, enhancing the system's sensitivity. The relationship between Mean Absolute Error (MAE) and reconstruction error is explored in **Figure 6**, a scatter plot where data points are colored by anomaly status (blue for false, red for true). Non-anomalous points cluster at lower errors (10−4 to 10−1 for reconstruction error, 10−2 to 10−1 for MAE), while anomalies dominate higher values (up to 104 and 101, respectively). This clear separation supports a threshold-based detection strategy, where high errors flag potential intrusions, aligning with the digital twin's role in mirroring system behavior. Finally, Figure 7 ranks the top 10 attack classes by

mean reconstruction error, with Fingerprinting and MITM exhibiting the highest errors (40), followed by Vulnerability scanner (10), while classes like Port Scanning show minimal errors. This ranking highlights the autoencoder's challenges with complex attack features, suggesting potential improvements in model architecture, such as deeper layers or variational autoencoders, to enhance reconstruction of sparse or intricate patterns. Together, these figures (Figures 4, 5, 6, 7) demonstrate the autoencoder's capability to learn, generalize, and detect anomalies, providing a robust framework for IDS in digital twin applications. Clear separation supports a threshold-based detection strategy, where high errors flag potential intrusions, aligning with the digital twin's role in mirroring system behavior.

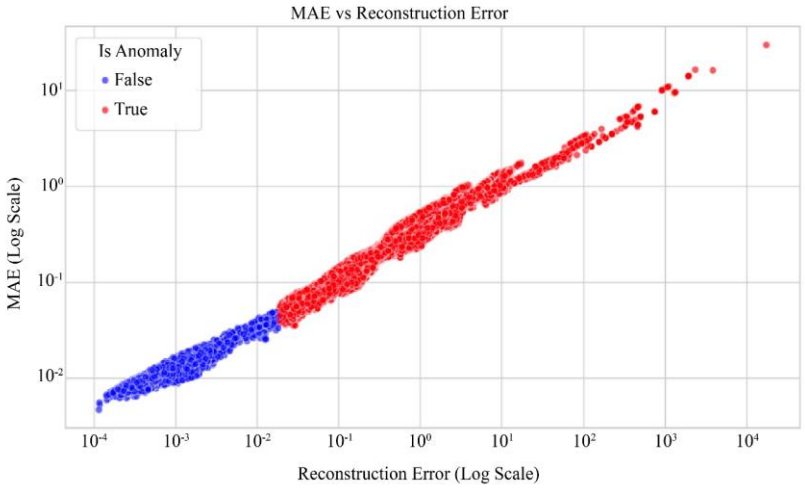

Fig. 6 Scatter plot of MAE vs. reconstruction error (log scale), with anomalies (red) at higher errors, demonstrating clear separation for intrusion detection

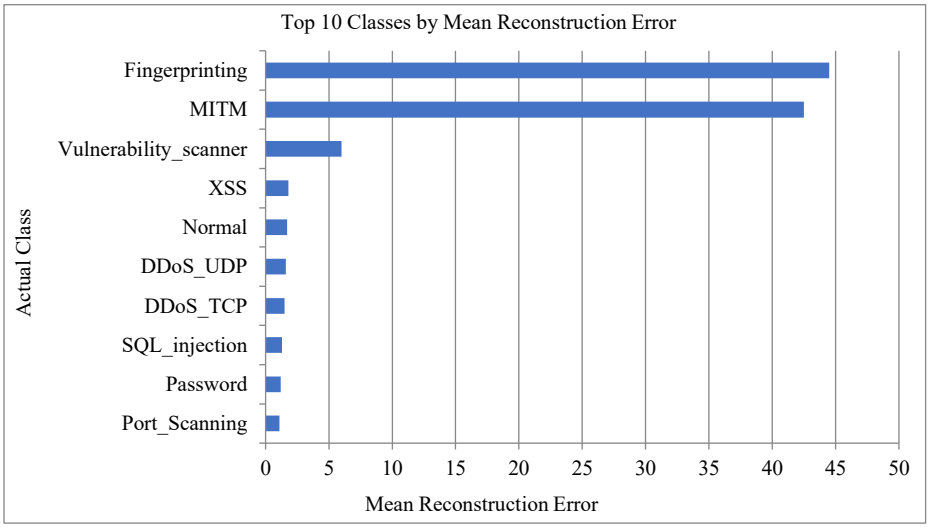

Fig. 7 Bar chart of top 10 attack classes by mean reconstruction error, with Fingerprinting and MITM showing the highest errors (40), guiding detection prioritization

5.2 Performance Analysis of the DNN-Based Attack Classifier

The Deep Neural Network (DNN)-based IDS demonstrates exceptional performance in classifying 15 threat categories, as evidenced by the metrics and visualizations in Figure 8 and Figure 9, alongside the detailed classification report (Table 4). The multiclass intrusion detection system demonstrates exceptional performance across all 15 threat categories, achieving an overall accuracy of 99.97% across 443,841 instances.

Table 4. Classification report

Class	Pr	Re	Fs	Support
Backdoor	1.0000	0.9881	0.9940	4972
DDoS HTTP	1.0000	1.0000	1.0000	9982
DDoS ICMP	1.0000	0.9999	0.9999	23287
DDoS TCP	1.0000	1.0000	1.0000	10012
DDoS UDP	1.0000	1.0000	1.0000	24314
Fingerprinting	1.0000	0.9900	0.9950	200
MITM	1.0000	1.0000	1.0000	243

Normal	1.0000	1.0000	1.0000	323129
Password	1.0000	1.0000	1.0000	10031
Port Scanning	0.9735	0.9996	0.9863	4513
Ransomware	0.9967	0.9707	0.9835	2185
SQL injection	1.0000	1.0000	1.0000	10241
Uploading	1.0000	1.0000	1.0000	7527
Vulnerability scanner	1.0000	0.9983	0.9992	10022
XSS	0.9947	1.0000	0.9973	3183
accuracy	0.9997	0.9997	0.9997	0.9997
macro avg	0.9977	0.9964	0.9970	443841
weighted avg	0.9997	0.9997	0.9997	443841

Perfect classification (Pr=1.0, Re=1.0, F1=1.0) was attained for 8 critical classes: DDoS HTTP, DDoS TCP, DDoS UDP, MITM, Normal traffic, Password attacks, SQL injection, and Uploading exploits. Near-perfect detection was observed for high-risk threats, including Backdoor (98.81% Re), Ransomware (97.07% Re), and XSS (100% Re).

The system maintains robust performance across severe class imbalances, from the largest category (Normal: 323,129 instances) to the smallest (Fingerprinting: 200 instances), with all Fss exceeding 98.3%. Minor Pr degradation occurred only in Port Scanning (97.35%), though it retained 99.96% Re. Macro-averaged metrics (Pr=99.77%, Re=99.64%, F1=99.70%) confirm consistent performance across classes, while weighted averages (99.97% across all metrics) reflect the model's stability under real-world data distribution. These results indicate a highly reliable intrusion detection solution with balanced Pr-Re characteristics essential for security applications.

5.2.1 Classification Model - Training and Validation Loss

Training and validation loss trajectories are visualized over 50 epochs, with the blue and orange lines representing each metric, respectively. A steep descent occurs in both curves during the first 10 epochs, with training loss dropping from 0.08 and validation loss from 0.07 to near-zero values. The convergent behavior of both metrics demonstrates effective model learning and strong generalization capability. The loss metric calculated as Loss = $-\frac{1}{n}\sum_{i=1}^{n}\left[yi\ log(\widehat{yi}) + (1-yi)\ log(1-\widehat{yi})\right]$, is likely cross-entropy loss, where yi is the true label and ˆyi is the predicted probability. The rapid convergence within 10 epochs indicates an effective learning rate, possibly optimized using an algorithm like Adam, with a batch size and epoch count sufficient to reach a local minimum. The stability suggests appropriate regularization, such as dropout or weight decay, to prevent overfitting.

5.2.2 Classification Model - Training and Validation Accuracy

This line graph displays the training accuracy (green line) and validation accuracy (red line) over 50 epochs. Both metrics start around 0.975 and increase sharply to approximately 0.995 within the first 10 epochs, remaining stable near 1.0 thereafter. The near-identical trends of training and validation accuracy indicate consistent model performance across datasets, reflecting high reliability in predicting attack types.

Accuracy is defined as accuracy = (Correct Prediction)/(Total Prediction), computed after applying a threshold (e.g., 0.5) to predicted probabilities from a softmax output layer. The rapid rise to 0.995 suggests a well-tuned model architecture, possibly a deep neural network with multiple layers, trained on a balanced dataset of attack types. The sustained high accuracy post-epoch 10 indicates the model has learned discriminative features effectively, likely enhanced by techniques such as batch normalization or data augmentation.

The confusion matrix reveals exceptional classification performance with distinct patterns of minimal misclassification. Eight critical threat categories - DDoS HTTP (9,982 instances), DDoS TCP (10,012), DDoS UDP (24,314), MITM (243), Normal traffic (323,129), Password attacks (10,031), SQL injection (10,241), and Uploading exploits (7,527) - achieved perfect detection with zero errors. Near-flawless identification was observed for DDoS ICMP (23,284/23,287 correct, 99.99%) and XSS (100% accuracy on 3,183 instances).

Minor errors were concentrated in four classes, exhibiting two primary confusion patterns:

- Backdoor-Port Scanning: 59 of 4,972 Backdoor attacks (1.19%) were misclassified as Port Scanning.
- Mutual Ransomware Confusions:– Ransomware showed 64 of 2,185 cases (2.93%) misidentified as Port Scanning.
    - Port Scanning had 2 of 4,513 cases (0.04%) misclassified as Ransomware.
    - Fingerprinting contributed 2 of 200 errors (1.0%) to Ransomware.

Additional negligible errors included 3 DDoS ICMP instances misclassified as Ransomware and 17 Vulnerability Scanner cases (0.17% of 10,022) confused with XSS. Crucially, all high-impact attacks (DDoS variants, SQLi, MITM) demonstrated perfect detection, and the largest class (Normal traffic) maintained zero misclassifications despite comprising 72.8% of the dataset. The error distribution highlights the model's remarkable Pr while pinpointing specific challenges: 95% of all errors stem from confusion between scanning activities (Port Scanning) and encryption-based threats (Ransomware). This targeted weakness suggests future refinement could focus on feature differentiation between these attack subtypes to push performance even closer to 100% accuracy.

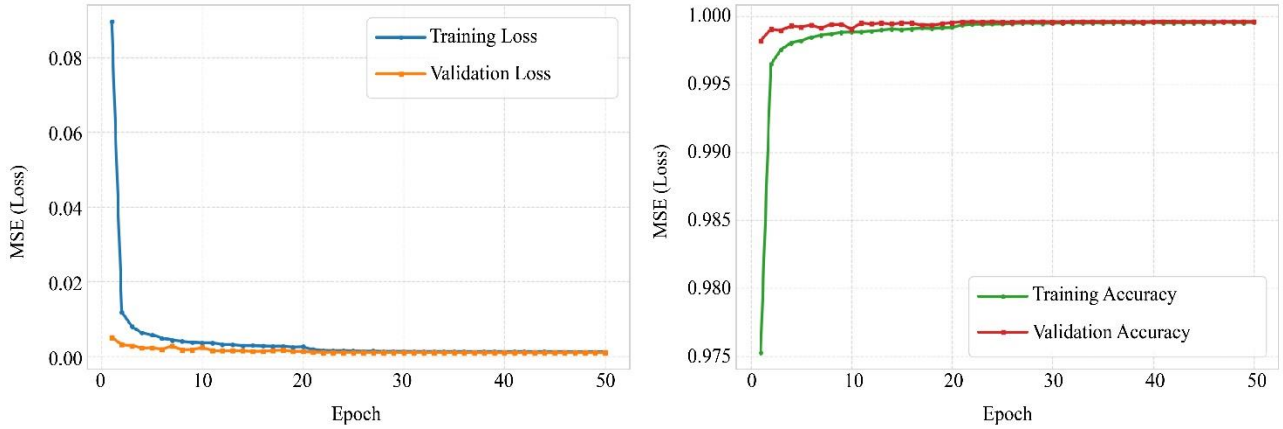

Fig. 8 Training and validation loss (blue and orange lines) and accuracy (green and red lines) over 50 epochs, converging to near-zero loss and ˜0.995 accuracy, indicating robust model generalization

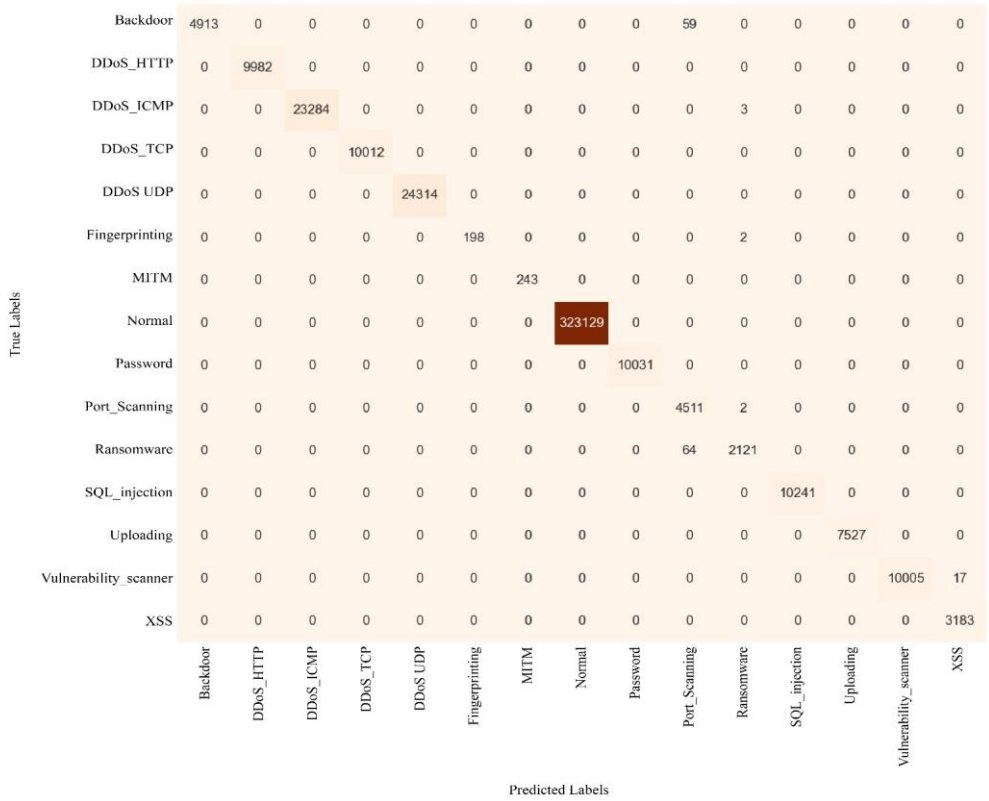

Fig. 9 Confusion matrix illustrating near-perfect classification across 15 attack classes, with minor misclassifications primarily between Backdoor, Port Scanning, and Ransomware

5.3 Performance Evaluation of the Anomaly Detector Model- Based IDS

The anomaly detector model-based IDS leverages anomaly score analysis and classification performance metrics to identify deviations in network behavior, as visualized in Figures 10-13. Figure 10 presents the Receiver Operating Characteristic (ROC) curve, plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) across various thresholds for 82,314 instances. With an Area Under the Curve (AUC) of 0.9079, significantly above the random classifier baseline (AUC = 0.5), the model demonstrates strong discriminative capability in distinguishing anomalous from normal behavior, making it effective for intrusion detection. The distribution of anomaly scores across 443,841 instances is depicted in Figure 11, a histogram revealing a bimodal pattern with peaks at approximately -0.05 and 0.05, and a mean score of 0.0034 (marked by a red dashed line). This distribution, concentrated between -0.1 and 0.1, suggests two distinct clusters of data points, potentially reflecting normal and anomalous behaviors.

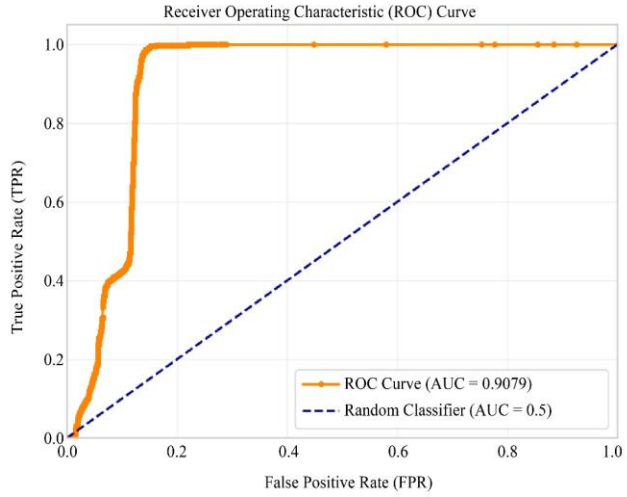

Fig. 10 ROC curve showing the anomaly detector's performance with an AUC of 0.9079, indicating strong discriminative capability across 82,314 instances compared to a random classifier (AUC = 0.5)

The slight positive skew (mean = 0.0034) indicates a tendency toward higher anomaly scores, which could aid in setting detection thresholds. Figure 12 provides a class-specific view through boxplots of anomaly scores across actual classes. The Normal class exhibits a narrow interquartile range centered near zero, indicating consistent low anomaly scores typical of benign traffic.

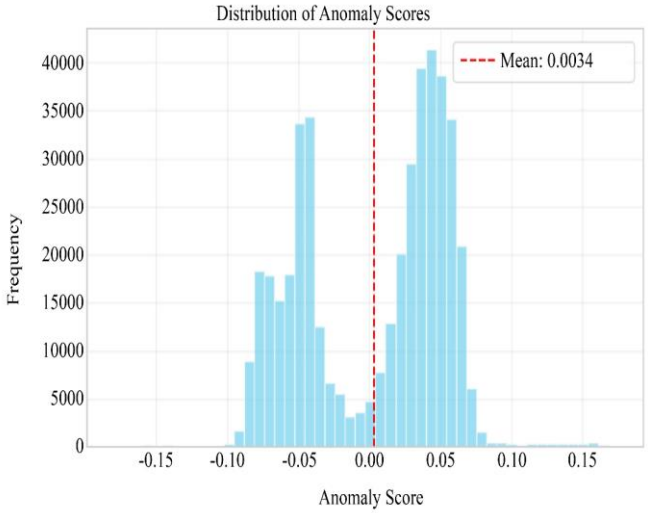

Fig. 11 Histogram of anomaly scores for 443,841 instances, displaying a bimodal distribution with peaks at ~-0.05 and ~0.05, and a mean score of 0.0034, suggesting distinct data clusters

In contrast, classes like Backdoor, DDoS HTTP, and XSS show wider distributions with higher median scores and outliers, particularly in Backdoor and Vulnerability scanner, suggesting greater variability and potential anomalies. These outliers highlight exceptional cases that the model flags as deviations, enhancing its sensitivity to rare attack patterns. Finally, Figure 13 illustrates mean anomaly scores per class, with sample sizes noted (e.g., Normal: n=323,129, Backdoor: n=4,972, MTM: n=24,314). Most classes have cores near zero, but Backdoor (≈0.12) and Vulnerability scanner (≈0.10) exhibit notably higher positive means, indicating stronger anomaly presence. Conversely, MTM (≈0.10) shows a negative mean, suggesting a unique anomaly profile, possibly due to distinct behavioral features. These differences underscore the model's ability to differentiate attack types based on anomaly scores. Together, Figures 10, 11, 12, and 13 demonstrate the anomaly detector's robust performance in identifying deviations, with high discriminative power (AUC = 0.9079), clear clustering of anomaly scores, and class-specific insights that guide targeted intrusion detection in a digital twin-based IDS framework.

5.4. Comprehensive Performance Metrics of the Digital Twin-Based IDS

The digital twin-based Intrusion Detection System (IDS) integrates behavior modeling, anomaly detection, and classification components, with its overall performance detailed in Tables 5, 6, and 7**. Table 5 provides a comprehensive overview of classification and anomaly detection metrics. The classifier achieves an exceptional accuracy of 0.9996688, with macro and micro averages for Pr, Re, and Fs all exceeding 0.996, and a log loss of 0.000685878, indicating high-confidence predictions across diverse attack classes.

In contrast, the anomaly detector shows a lower accuracy of 0.422166947, with a high Pr of 0.997441925 but a Re of 0.416689263, yielding an Fs of 0.587814403. Its ROC-AUC of 0.907874182 suggests good discriminative ability despite challenges with class imbalance. Reconstruction metrics reveal a threshold of 0.018399744, with mean reconstruction errors of 1.274480052 (all data), 0.006091031 (normal data), and 1.288849785 (attack data), highlighting the system's ability to distinguish normal from attack behaviors, albeit with higher variability in attack data (std = 29.02088087). Table 6 outlines training efficiency and performance indicators.

The behaviour model trains in 39.44 seconds with a final loss of 0.03055975 and a validation loss of 0.010249236, indicating effective generalization. The anomaly detector trains rapidly in 1.81 seconds, while the classifier requires 1112.66 seconds, totalling 1153.90 seconds. The classifier's final accuracy (0.999545872) and validation accuracy (0.999602914) underscore its robustness, and the consistent reconstruction threshold (0.018399744) aligns with anomaly detection settings.

These metrics suggest efficient training, particularly for the anomaly detector, and high classification performance, though the behavior model's higher loss compared to validation may warrant further tuning.

Table 7 logs real-time security alerts from July 30, 2025, between 08:27:42.489940 and 08:27:51.393987, capturing

rapid event succession. Most events are flagged as Normal with a confidence of 1.0, but specific attacks-DDoS HTTP (confidence 0.999835849), Password (0.999892831), and XSS (0.999933839)-are detected with high confidence and varying risk scores (74.2567509 to 837.9706688). All alerts show both behaviour and isolation anomalies as True, indicating persistent deviations. This table highlights the IDS's capability to detect anomalies in real-time, with high-confidence attack predictions critical for threat analysis, though the prevalence of Normal labels suggests a conservative anomaly flagging approach. Collectively, Tables 5, 6, and 7** demonstrate the IDS's strengths in classification accuracy and real-time detection, with the anomaly detector facing challenges in Re.

These insights guide future improvements, such as optimizing anomaly detection thresholds or enhancing feature differentiation, to bolster the system's effectiveness in a digital twin framework.

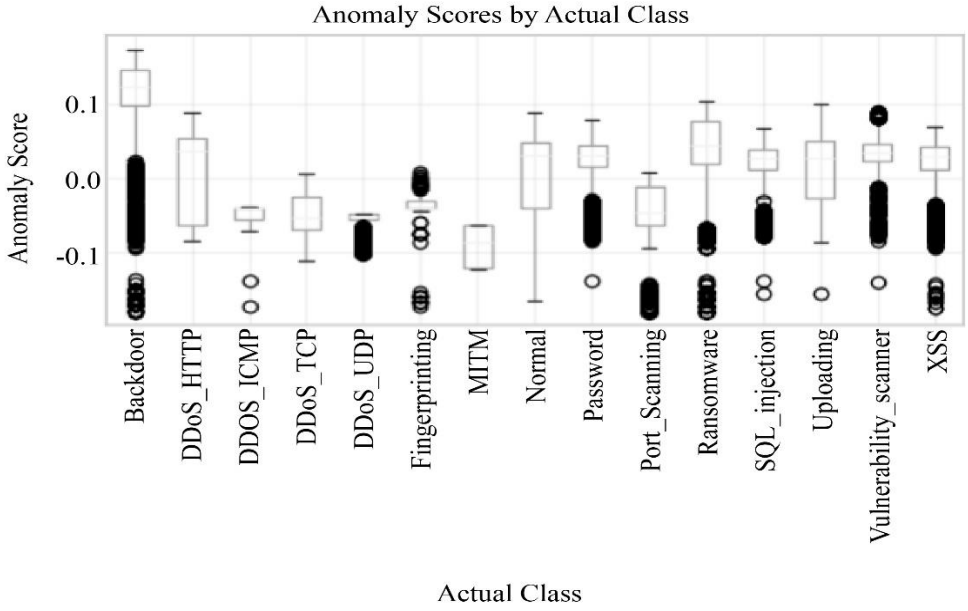

Fig. 12 Boxplot of anomaly scores by actual class, highlighting narrow ranges for normal and wider distributions with outliers for classes like Backdoor and XSS, indicating anomaly variability

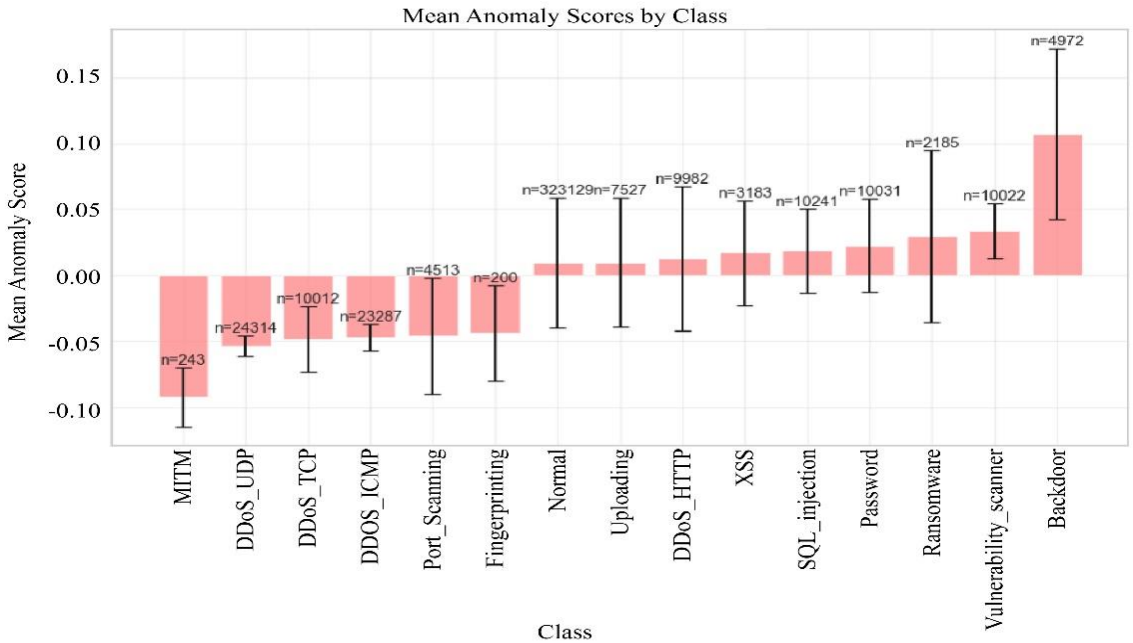

Fig. 13 Bar chart of mean anomaly scores by class, with Backdoor (˜0.12) and Vulnerability scanner (˜0.10) showing high positive scores, and MTM (˜- 0.10) indicating a distinct anomaly profile

Table 5. Classification report metrics for digital twin-based IDS models

Matrix	Value
classifier accuracy	0.9996688
classifier Pr macro	0.997657006
classifier Pr micro	0.9996688
classifier Re macro	0.996438311
classifier Re micro	0.9996688
classifier f1 macro	0.99702014
classifier f1 micro	0.9996688
classifier log loss	0.000685878
anomaly accuracy	0.422166947
anomaly Pr	0.997441925
anomaly Re	0.416689263
anomaly f1	0.587814403
anomaly roc auc	0.907874182
reconstruction threshold	0.018399744
mean reconstruction error all	1.274480052
std reconstruction error all	28.8581901
mean mae all	0.434120629
std mae all	0.239016621
mean reconstruction error normal	0.006091031
std reconstruction error normal	0.034731031
mean reconstruction error attack	1.288849785
std reconstruction error attack	29.02088087

Table 6. Training metrics and performance indicators for digital twin-based IDS models

Matrix	Value
Behavior model training time	39.43897891
Anomaly detector training time	1.806567669
Classifier training time	1112.657593
Total training time	1153.903139
Behavior final loss	0.03055975
Behavior final val loss	0.010249236
Classifier final Accuracy	0.999545872
Classifier final val accuracy	0.999602914
Reconstruction threshold	0.018399744

Table 7. Security alerts from digital twin-based IDS

Timestamp	Event Type	Risk Score	Predicted Attack	Confidence	Anomaly Flags
2025-07-30T08:27:42.489940	SECURITY ALERT	74.52485204	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:43.296250	SECURITY ALERT	170.1748149	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:44.111847	SECURITY ALERT	83.67357959	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:44.923185	SECURITY ALERT	83.6735782	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:45.729879	SECURITY ALERT	83.67412147	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:46.549461	SECURITY ALERT	74.2567509	DDoS HTTP	0.999835849	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:47.353757	SECURITY ALERT	74.52484704	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}

2025-07-30T08:27:48.164266	SECURITY ALERT	108.2820245	Password	0.999892831	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:48.973838	SECURITY ALERT	423.4674091	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:49.779953	SECURITY ALERT	837.9706688	XSS	0.999933839	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:50.583627	SECURITY ALERT	170.1747158	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}
2025-07-30T08:27:51.393987	SECURITY ALERT	83.6734897	Normal	1	\{' behavior anomaly': True, 'isolation anomaly': True\}

Table 8. DT-IDS vs. Established ML and DL baselines: A comparative analysis

Reference	Proposed Model	Accuracy (\%)
[25]	DL-Based IDS	98.32
[26]	TRACER	96.17
[27]	Transformer–GAN–AE	98.63
[28]	FD-IDS	94.82
[29]	CNN	95.5
Proposed	DT-IDS	99.97

5.5. Comparative Analysis of IDS

This section rigorously compares our proposed Digital Twin-based Intrusion Detection System (DT-IDS) against prominent existing solutions, with a primary focus on detection accuracy-a critical indicator of real-world performance. Table 1 provides a concise overview of the accuracy achieved by various models, demonstrating the superior performance of our DT-IDS. As evident from Table 8, the DT-IDS achieves an impressive accuracy of 99.97%, setting a new benchmark compared to the surveyed literature. This exceptional performance stems from the intelligence of the digital twin, which is derived from three specialized machine learning models operating synergistically to detect, classify, and mitigate threats. These models collectively form the analytical core of the digital twin, providing a multi-layered and robust security assessment. In contrast, existing approaches, while effective in their specific contexts, exhibit comparatively lower accuracy:

- Traditional Deep Learning Approaches: The DL-Based IDS [25], a hybrid model combining BiGRU, LSTM, and softmax, achieved 98.32%. While adept at handling lengthy sequences of security audit data through TBPTT, its single-model hybrid approach appears to be less comprehensive than the multi-faceted strategy of DT-IDS. Similarly, the CNN-based model [29] reached 95.5%, highlighting the limitations of relying solely on convolutional features without the broader analytical scope offered by a digital twin.
- Transformer-based Solutions: TRACER [26], an Attack-Aware Divide-and-Conquer Transformer for IIoT, demonstrated 96.17% accuracy. The Transformer–GAN–AE [27], an optimized model for edge and IIoT systems, achieved a performance of 98.63%. While these models leverage advanced transformer architectures, their focus on specific network types or singular complex architectures might limit their overall detection capabilities compared to DT-IDS's integrated and adaptive framework.
- Distributed Learning Models: FD-IDS [28], which employs Federated Learning with Knowledge Distillation for non-IID IoT environments, recorded 94.82% accuracy. While federated learning offers significant advantages in privacy and distributed training, the inherent complexities of non-IID data and the distillation process may introduce trade-offs in overall detection accuracy when compared to a centralized, highly optimized system like DT-IDS.

The DT-IDS's superior accuracy is a testament to its innovative architecture, where the combined strengths of its specialized machine learning models provide a more comprehensive and precise threat detection capability. This multi-layered approach enables the digital twin to not only identify known attack patterns but also adapt and respond to novel threats with unparalleled effectiveness, thereby significantly enhancing cybersecurity in modern network infrastructures.

5.6. Ethical Implications of AI in Cybersecurity

While AI enhances cybersecurity through advanced threat detection and rapid response, it raises critical ethical concerns. Training data bias may create unequal protection across systems and user groups, while the tension between security monitoring and privacy rights requires careful balance. As AI systems become more autonomous in threat response, questions of accountability arise-who is responsible when automated decisions harm innocent users? Additionally, many AI models operate as "black boxes," making their decision-making processes opaque and complicating transparency and

accountability. Beyond these concerns, the dual-use nature of AI security tools means that defensive technologies can potentially be weaponized by malicious actors. Furthermore, the high cost of advanced AI cybersecurity solutions creates equity gaps, where well-funded organizations receive superior protection while smaller entities remain vulnerable.

Addressing these implications effectively demands thoughtful policies, transparent practices, meaningful human oversight, and ongoing collaboration between technologists, ethicists, and stakeholders to ensure AI in cybersecurity serves the broader good while respecting individual rights and organizational fairness.

6. Conclusion

The proposed digital twin-based cybersecurity framework offers a robust and innovative solution for securing Industrial Internet of Things environments. By integrating a dynamic virtual replica with advanced anomaly detection, intelligent threat classification, and proactive security management, the framework addresses the critical limitations of traditional approaches. Experimental results using the Edge-IIoT dataset demonstrate exceptional performance, with high Accuracy, Pr, recall, and FSS in multiclass threat classification, alongside effective anomaly detection capabilities. The clear separation of normal and attack traffic via reconstruction error further validates its ability to identify unseen threats. This approach not only outperforms existing solutions but also enables real-time monitoring and rapid response, establishing a new standard for IIoT cybersecurity and paving the way for future enhancements in proactive threat management.

Future research could enhance the proposed model by incorporating multimodal data, such as combining Edge-IIoT text data with network traffic metadata, to improve attack detection accuracy. Exploring transfer learning to adapt the model to other IoT cybersecurity datasets may increase its robustness. Additionally, investigating real-time anomaly detection and federated learning approaches could enable scalable deployment in dynamic IIoT environments.

Acknowledgments

The author acknowledge the support from the Deanship of Scientific Research, Najran University, Najran, Saudi Arabia, during the research work.

References

[1] Alessandra De Benedictis et al., "Digital Twins for Anomaly Detection in the Industrial Internet of Things: Conceptual Architecture and Proof-of-Concept," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 12, pp. 11553-11563, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Zhi Qiang Wang, and Abdulmotaleb El Saddik, "DTITD: An Intelligent Insider Threat Detection Framework Based on Digital Twin and Self-Attention Based Deep Learning Models," *IEEE Access*, vol. 11, no. 10, pp. 114013-114030, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Seba Anna Varghese et al., "Digital Twin-based Intrusion Detection for Industrial Control Systems," *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, Pisa, Italy, pp. 611-617, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Qinghua Xu, Shaukat Ali, and Tao Yue, "Digital Twin-based Anomaly Detection with Curriculum Learning in Cyber-Physical Systems," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 5, pp. 1-32, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Burcu Bolat-Akça, and Elif Bozkaya-Aras, "Digital Twin-Assisted Intelligent Anomaly Detection System for Internet of Things," *Ad Hoc Networks*, vol. 158, no. 10, pp. 123-135, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mikail Mohammed Salim, David Camacho, and Jong Hyuk Park, "Digital Twin and Federated Learning Enabled Cyberthreat Detection System for IoT Networks," *Future Generation Computer Systems*, vol. 161, pp. 701-713, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Fatemeh Akbarian, Emma Fitzgerald, and Maria Kihl, "Intrusion Detection in Digital Twins for Industrial Control Systems," *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] S. Krishnaveni et al., "CyberDefender: An Integrated Intelligent Defense Framework for Digital-Twin-Based Industrial Cyber-Physical Systems," *Cluster Comput*, vol. 27, no. 6, pp. 7273-7306, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] S. Krishnaveni et al., "TwinSec-IDS: An Enhanced Intrusion Detection System in SDN-Digital-Twin-Based Industrial Cyber-Physical Systems," *Concurrency and Computation: Practice and Experience*, vol. 37, no. 3, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Easa Alalwany et al., "Stacking Ensemble Deep Learning for Real-Time Intrusion Detection in IoMT Environments," *Sensors*, vol. 25, no. 3, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[11] Arreche, Osvaldo, Tanish Guntur, and Mustafa Abdallah, "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems," *Applied Sciences*, vol. 14, no. 10, pp. 1-41, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] A. Namrita Gummadi, J. C. Napier and M. Abdallah, "XAI-IoT: An Explainable AI Framework for Enhancing Anomaly Detection in IoT Systems," *IEEE Access*, vol. 12, no. 5, pp. 71024-71054, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nabeel H. Al-A'araji, Safaa O. Al-Mamory, and Ali H. Al-Shakarchi, "Classification and Clustering Based Ensemble Techniques for Intrusion Detection Systems: A Survey," *Journal of Physics: Conference Series, Iraqi Academics Syndicate International Conference for Pure and Applied Sciences (IICPS)*, Babylon, Iraq, vol. 1818, pp. 1-35, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Ahmed Mahfouz et al., "Ensemble Classifiers for Network Intrusion Detection Using a Novel Network Attack Dataset," *Future Internet*, vol. 12, no. 11, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Riccardo Lazzarini, Huaglory Tianfield, and Vassilis Charissis, "A Stacking Ensemble of Deep Learning Models for IoT Intrusion Detection," *Knowledge-Based Systems*, vol. 279, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Yongzhong Cao et al., "An Intrusion Detection System Based on Stacked Ensemble Learning for IoT Network," *Computers and Electrical Engineering*, vol. 110, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ghalia Nassreddine, Mohamad Nassereddine, and Obada Al-Khatib, "Ensemble Learning for Network Intrusion Detection Based on Correlation and Embedded Feature Selection Techniques," *Computer*, vol. 14, no. 3, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Uday Chandra Akuthota, and Lava Bhargava, "Transformer-Based Intrusion Detection for IoT Networks," *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 6062-6067, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Moussab Orabi et al., "Anomaly Detection in Smart Manufacturing: An Adaptive Adversarial Transformer-Based Model," *Journal of Manufacturing Systems*, vol. 77, pp. 591-611, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mustafa Abdallah et al., "Anomaly Detection and Inter-Sensor Transfer Learning on Smart Manufacturing Datasets," *Sensors*, vol. 23, no. 1, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Baibhab Chatterjee et al., "Context-Aware Collaborative Intelligence with Spatio-Temporal In-Sensor-Analytics for Efficient Communication in a Large-Area IoT Testbed," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6800-6814, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yazeed Alotaibi, and Mohammad Ilyas, "Ensemble-Learning Framework for Intrusion Detection to Enhance Internet of Things' Devices Security," *Sensors*, vol. 23, no. 12, pp. 1-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tommaso Zoppi, Andrea Ceccarelli, and Andrea Bondavalli, "MADneSs: A Multi-Layer Anomaly Detection Framework for Complex Dynamic Systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 2, pp. 796-809, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Evgeniya Petrova Nikolova, and Veselina Gospodinova Jecheva, "Anomaly Based Intrusion Detection Based on the Junction Tree Algorithm," *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 641-649, 2007. [Google Scholar] [Publisher Link]

[25] Danish Javeed et al., "An Intrusion Detection System for Edge-Envisioned Smart Agriculture in Extreme Environment," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 26866-26876, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[26] Minyue Wu et al., "TRACER: Attack-Aware Divide-and-Conquer Transformer for Intrusion Detection in Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 6, pp. 4924-4934, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Ahmad Salehiyan, Pardis Sadatian Moghaddam, and Masoud Kaveh, "An Optimized Transformer-GAN-AE for Intrusion Detection in Edge and IIoT Systems: Experimental Insights from WUSTL-IIoT-2021, EdgeIIoTset, and TON_IoT Datasets," *Future Internet*, vol. 17, no. 7, pp. 1-34, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Haonan Peng, Chunming Wu, and Yanfeng Xiao, "FD-IDS: Federated Learning with Knowledge Distillation for Intrusion Detection in Non-IID IoT Environments," *Sensors*, vol. 25, no. 14, pp. 1-31, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Ali Abdi Seyedkolaei, Fatemeh Mahmoudi, and José García, "A Deep Learning Approach for Multiclass Attack Classification in IoT and IIoT Networks Using Convolutional Neural Networks," *Future Internet,* vol. 17, no. 6, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]