Scientific Reports

https://doi.org/10.1038/s41598-026-42041-w

Article in Press

A zero-trust digital twin framework for privacy-preserving multi-dataset intrusion detection in industrial IoT with lightweight blockchain auditing

Received: 16 January 2026

Accepted: 24 February 2026

Published online: 31 March 2026

Cite this article as: Mishra S., Aldafas T.S.M. & Alshammari N.S. A zero-trust digital twin framework for privacy-preserving multi-dataset intrusion detection in industrial IoT with lightweight blockchain auditing. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-42041-w

Shailendra Mishra, Tariq Saleh M. Aldafas & Naif S. Alshammari

A Zero-Trust Digital Twin Framework for Privacy-Preserving Multi-Dataset Intrusion Detection in Industrial IoT with Lightweight Blockchain Auditing

Shailendra Mishra [1*] ,Tariq Saleh M Aldafas[2,] ,Naif S. Alshammari [3*]

[1]Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al Majmaah 11952, Saudi Arabia.
[2]Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Al Majmaah 11952, Saudi Arabia.
[3]Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al Majmaah 11952, Saudi Arabia.

Corresponding Author*: Shailendra Mishra, Naif S. Alshammari
Email address: s.mishra@mu.edu.sa, n.alshammari@mu.edu.sa

ABSTRACT

Industrial IoT (IIoT) environments face growing cyber threats due to device heterogeneity and cyber-physical integration. This study proposes a Zero Trust-enhanced intrusion detection framework integrating deep learning anomaly detection, differential privacy, lightweight blockchain-inspired hash-chained ledger and Digital Twin-based situational awareness and visualization of device trust states , designed for low-latency inference suitable for near-real-time IIoT monitoring .A unified dataset was constructed by merging NSL-KDD, CICIDS-2017, and IoT-23 (2,513,419 raw samples unified to 143 features, balanced to 100,000 samples across Normal, DoS, Probe, R2L, U2R classes using SMOTE). Mutual information-based feature selection reduced features to 25. Optimized Multilayer Perceptron (MLP) and CNN–BiLSTM models achieved 89–91% accuracy and 0.89–0.91 macro F1-score, with near-perfect rare-attack detection (F1 ≈ 1.00 for R2L/U2R). Differential privacy (Laplace, ε=25) reduced accuracy to ~78%, quantifying the privacy-utility trade-off. The decoupled Zero-Trust Manager dynamically updates trust scores based on prediction confidence, with tamper-evident SHA-256 hash-chained logging adding negligible latency (~1.04–1.06 s for 500 samples). This lightweight, centralized design offers strong cross-domain generalization and deployability for resource-constrained IIoT.

Keywords: Industrial Internet of Things; Intrusion Detection System; Zero Trust Architecture; Differential Privacy; Lightweight Blockchain; Digital Twin

1. INTRODUCTION

The fourth industrial revolution (Industry 4.0) has increased the rate at which the Industrial Internet of Things (IIoT) is now being adopted to allow factories, energy grids and logistics systems to interoperate and automate processes through smart sensing and control [1]. Conventional perimeter-based security controls, e.g., firewalls and fixed authentication frameworks cannot be used to secure IIoT environments where equipment may exchange signals on a distributed network quite often [2]. Industrial IoT systems are normally heterogeneous networks of devices, including sensors and controllers, cloud-based analytical systems, and other data gathering and processing devices, which perform big data processing [2],[3]. The interdependence of the sensors, actuators, and cloud systems also creates a vulnerability that can not only jeopardize the integrity of data but also the physical safety [4],[5].

The Zero Trust Architecture (ZTA) has been suggested as a promising model to address these threats since it focuses on constant authentication, dynamic access control and the concept of never trust, always verify, Zero Trust Architecture provides this dynamic security solution, and there is no party, either internal or external, that will be trusted by default [6,7,8]. Combined with Digital twin security can be achieved without storage of sensitive industrial data [9]. Digital Twin, a virtual representation of a real system, which allows to monitor, predict and optimize the work of the industry in real-time [9]. Digital Twins have shown promise in cyber-physical systems beyond industrial settings, such as in precision agriculture and livestock farming, where IoT-integrated DT frameworks enable real-time monitoring, health prediction, and enhanced well-being of assets [10]. This demands an active, context sensitive security model, which imposes very strict verification policy, on-going trust monitoring and dynamically changing response to new patterns of attack [11].

Although Zero Trust has proved promising in the context of industrial and manufacturing infrastructure, the current solutions based on ZTA are mostly centered around access control and policy enforcement, with less emphasis on the integration of intrusion detection, privacy-preserving methods, and efficient trust audits [7],[11]. On the other hand, machine learning-based intrusion detection systems have proved very effective in the context of IIoT networks [12]. Nevertheless, their performance is often hampered by biased datasets, extreme class imbalance, high computational complexity, and poor generalization for diverse industrial traffic patterns [4],[11].

To overcome these issues, this paper proposes a Zero Trust-strengthened intrusion detection framework for IIoT, which combines deep

learning-based anomaly detection, differential privacy, lightweight blockchain-inspired hash-chained ledger, and Digital Twin-based situational awareness and visualization of device trust levels. A comprehensive intrusion detection dataset is built by combining three popular benchmark datasets, NSL-KDD, CICIDS2017, and IoT-23, to facilitate cross-domain testing and improve generalization capabilities [4], [8], [13]. Class imbalance problems are handled using resampling methods, and feature selection is performed using mutual information.

The proposed framework evaluates an optimized Multilayer Perceptron (MLP) and a CNN–BiLSTM hybrid model, demonstrating that data-centric design can achieve strong intrusion detection performance without excessive architectural complexity [12], [13]. To support privacy-preserving operation, differential privacy mechanisms are applied, and their impact on detection accuracy is systematically assessed [14]. A decoupled Zero Trust Manager dynamically updates device trust scores based on model prediction confidence, with updates recorded via a lightweight SHA-256 hash-chained ledger to provide tamper-evident auditing without introducing inference latency [14], [15]. Additionally, Digital Twin visualization enhances situational awareness by displaying device trust states in real time [8], [16], [17].

The framework adopts a multi-layered security architecture based on well-established principles: (1) the Detection Layer adopts deep learning-based anomaly detection in accordance with data-driven intrusion detection system (IDS) theory [12],[13]; (2) the Trust Enforcement Layer adopts continuous verification in accordance with the Zero Trust philosophy of “never trust, always verify” [6],[7]; (3) the Audit Layer adopts cryptographic hash-chaining for tamper-evident logging, leveraging lightweight integrity protection methods suitable for edge settings [15]; and (4) the Cyber-Physical Visibility Layer adopts Digital Twin as a real-time system state reflection, founded on cyber-physical system synchronization theory to enhance operator visibility [9],[17].By integrating the prediction confidence, trust score update, Digital Twin state reflection, and audit logging, the proposed solution enables a closed-loop operational feedback control. The integrated framework design differentiates the framework from the existing literature, which conventionally treats these aspects in separate manners or with centralized and computationally expensive approaches.

The novelty of this research lies in the operational integration of components within a closed-loop, lightweight pipeline tailored for IIoT environments: (1) deep learning predictions directly inform dynamic Zero Trust scoring; (2) privacy-preserving inference using differential privacy influences trust updates without exposing raw data; (3) tamper-evident logging is decoupled from inference, introducing negligible overhead; and (4) a Digital Twin provides real-time cyber-physical visualization of trust levels. In contrast to previous ZTA-related studies, which primarily focused on policy enforcement [6], [7], [8], federated privacy solutions with high communication overhead [13], [24], or complex blockchain implementations

with consensus latency [9], [15], the proposed centralized, data-driven framework achieves competitive detection performance (89–91% accuracy), resilience to rare attacks, and edge-device viability, while quantitatively assessing privacy costs.

2. LITERATURE REVIEW

2. 1 Review of Trends in Previous Researches

A substantial body of research work has been carried out to explore the use of artificial intelligence, blockchain, and Zero Trust models for improving the security of Industrial IoT networks [9], [12], [13], [14], [15], [16].The initial research work was mainly focused on traditional intrusion detection systems and cryptographic solutions; however, these models had limited flexibility to counter new-age cyber threats [7], [18], [19].The recent research work has started to focus more on context-aware intelligence and federated learning models that have the ability to counter both external and internal attacks in IIoT networks [3], [12].

Recent contributions have advanced IIoT and cyber-physical system security through machine-learning-driven detection and secure distributed architectures. Ensemble learning-based intrusion detection and feature-fusion approaches have demonstrated improved malicious activity detection accuracy in IoT and vehicular network environments [31], [32]. Graph-based deep learning models have also been explored for predictive analytics in IoT applications, illustrating the increasing adoption of advanced neural architectures in cyber-physical systems [33]. In addition, blockchain-enabled secure federated storage and encryption frameworks have been proposed to enhance data confidentiality and trust management in distributed environments [34]. These developments collectively indicate a shift toward intelligent, decentralized, and privacy-preserving security mechanisms, motivating the integrated Zero Trust–enabled digital-twin framework proposed in this study.

2. 2 IIoT Security: Blockchain and Machine Learning

A trust management framework that was suggested by Franklin et al. [15] is based on blockchain and machine learning technology and uses immutable audit trails in network communications but enhances the detection accuracy to Byzantine attacks. Equally Paul B, Rao M. [6] have come up with a zero-trust inspired smart manufacturing system model that integrates machine-learning anomaly detection and blockchain traceability. Nevertheless, the two models are based on centralized aggregation in their analysis, increasing the issues of latency and scalability [13]. This is also dealt with in the present research by taking a federated learning solution to avoid central dependency whilst ensuring that data confidentiality is upheld [14].

2. 3 Monitoring and Anomaly Detection Based on Digital Twin

A study by Lv et al. [20] investigated the vulnerability of message-handling in IIoT channels of communication and the benefits of digital twins in enhancing anomaly detection. DTs can detect variations in the operational parameters prior to the building of a physical destruction by constantly aligning physical and virtual assets data with real-world assets [8], [9], [17]. Nonetheless, such systems do not support cross-device authentication and adaptive trust control. These works are expanded in the proposed framework where Zero Trust validation mechanisms are introduced in the digital twin architecture to guarantee that all updates to the data are provided by a verified and trusted source [5],[21].

2. 4 Federated Learning and Privacy-Preserving Detection

Ali et al. [13] and Sarhan et al. [22] had studied federated learning as a distributed model training that would be applicable in IIoT intrusion detection. Their research showed the privacy protection and reduction of the data leakage risks in decentralized training. However, they also noted issues of non-IID data distribution, overhead of communication and slow convergence These issues are addressed in the present work by making use of adaptive aggregation capabilities and dynamic updates to the trust-weighted model, making sure that clients who are malicious or unreliable have little impact on the global model [23],[32],[34].

2. 5 Zero Trust in Industrial Control Systems

In the industrial control systems (ICS), a new trend has brought in Zero Trust Architecture (ZTA) [7],[8],[12]. Lv F et al. [20] implemented an asynchronous federated learning system, which incorporates Zero Trust access control verification. Their scores were more resilient to insider threats but had a weakness of not having consistently frequent recalibration of trust. Federated learning–based zero-trust intrusion detection framework for IoT networks that enhances data privacy while maintaining high detection accuracy across distributed devices [24], [25] , demonstrating improved accuracy and scalability without sharing raw device data.

2.6Adversarial Machine Learning and Insider Threats in IIoT

The use of Adversarial Machine Learning (AML) poses a major threat to IIoT security systems because the attackers design their input data to avoid machine-learning-based detection systems . Existing literature mostly focuses on the accuracy of prediction with very little focus on the resilience of a model in adversarial conditions [14], [26], [27]. Neither insider threats worsen the state of security of IIoT environments because compromised personnel or devices within the trust perimeter are capable of carrying out malicious activities. Most implementations of MQTT and CoAP applications do not perform ongoing verification of identity and therefore lose protection against escalation of privileges [13], [26]. The proposed ZTA integration

aims to counter these weaknesses by using dynamic verification and behavior-based access control.

2.7 Comparative Insights

An overview of related literature (as summarized in Table 1) indicates that although different models can help improve one of the three goals: improved accuracy, blockchain-based integrity, or privacy preservation, only a limited number of solutions can meet all three goals at the same time: continuous trust validation, federated privacy protection, and real-time digital twin synchronization This study advances the field by integrating these elements into a unified architecture, thereby strengthening both the cybersecurity posture and operational reliability of IIoT systems Decentralized ZTA frameworks for digital twin-based 6G networks provide complementary insights into adaptive trust mechanisms applicable to IIoT environments .

Table 1 summarizes key prior works across five central themes. While many studies advance individual aspects (dataset scale, privacy via FL/DP, ZTA policies, blockchain auditing, or hybrid DL), few integrate continuous trust validation, lightweight tamper-evident logging, strong rare-attack handling, and Digital Twin visibility in a resource-efficient manner suitable for IIoT edge environments. Persistent gaps include: (1) lightweight trust logging without performance penalty, (2) effective rare-attack detection in cross-domain merged datasets, (3) transparent quantification of privacy–utility trade-offs, and (4) closed-loop coupling of anomaly detection, dynamic trust scoring, and cyber-physical state visualization. The present work addresses these limitations through a data-centric, decoupled, and lightweight design.

Table 1: Thematic Literature Review

Theme	Key Studies	Main Contribution	Limitations	Addressed in this study
Multi-Dataset Benchmarks for IDS	Neto et al. (2023) [4]	Large-scale real-time IoT attack dataset (CICIoT2023)	Primarily data-focused; no integrated privacy/trust	Merges NSL-KDD, CICIDS-2017, IoT-23 into unified balanced dataset with MI feature selection & SMOTE

Explainable and Privacy-Preserving IDS	Fatema et al. (2025) [3]; Nawshin et al. (2024) [26], Kathole et al. (2024) [31],[32]	Federated/explainable IDS with SHAP+ DP	High overhead in FL/DP; limited rare-attack focus	Applies Laplace DP (ε=25) with quantified trade-off; near-perfect rare-attack
Zero Trust in IIoT/ICS Systems	Paul & Rao (2023) [6]; Laghari et al. (2025) [8]; Federici et al. (2023) [7]; Zanasi et al. (2024) [11]	ZTA models with continuous verification & adaptive control	Often lacks lightweight integration issues with IDS	Decoupled Zero-Trust Manager with dynamic trust scoring added latency impact
Lightweight Trust Logging / Auditing	Onwubiko et al. (2023) [9]; Franklin et al. (2022) [15]; Kathole et al. (2024) [33],[34]	Immutable auditing via blockchain/hybrid frameworks	High overhead in full blockchain; some centralization	Lightweight SHA-256 hash-chained ledger
Deep Learning Architectures for IDS	Laghari et al. (2025) [8];	Hybrid DL models for threat detection under ZTA	Higher complexity/overhead; often cloud-focused	Optimized shallow MLP

	Lilhore et al. (2025) [12]	matches deeper CNN-BiLSTM; low latency suitable for IIoT edge

Research Gaps

1. The lack of adaptive Zero Trust systems that would be smoothly integrated into IIoT digital twins.
2. The lack of focus on adversarial ML robustness in twin settings, the data-driven models are used by attackers.
3. Inadequate focus on insider threats detection in MQTT based IIoT communication networks.
4. Federated learning barriers related to global implementation of Zero Trust policy governance.
5. The lack of the empirical studies that, simultaneously, compare ZTA, federated learning, and adversarial attack simulations in the context of IIoT.

This study addresses the growing gap between the increasing complexity of Industrial Internet of Things (IIoT) ecosystems and the lack of adaptive, trust-based security models capable of making real-time security decisions. To address this challenge, the proposed framework integrates a deep learning–based intrusion detection system with a Zero Trust–driven policy engine and Digital Twin–assisted monitoring architecture. In the proposed architecture, an optimized Multilayer Perceptron and the CNN-BiLSTM hybrid architecture are analyzed for the capability of data-intensive designs to compete with the quality of the detection results while avoiding complexity in the architecture .For supporting privacy-preserving operations, the differential privacy is quantitatively analyzed and the effect of differential privacy on the detection performance is quantitatively examined In addition, a CNN-BiLSTM Zero Trust Manager adjusts the trust values of the devices based on the level of confidence in the prediction, and a decoupled SHA-256 hash-chain ledger tracks the trust state to maintain tamper-evident auditability, without introducing high inference delays [6],[17],[23].

3. RESEARCH METHODOLOGY

This study adopted a simulation-based experimental research design to develop and evaluate the proposed Zero Trust-Enhanced Digital Twin

Security Framework for Industrial Internet of Things (IIoT) systems. Three widely used cybersecurity benchmark datasets,NSL-KDD, CICIDS2017, and IoT-23 [28,29,30], are combined to construct a heterogeneous IIoT traffic environment suitable for cross-domain evaluation. All experiments are carried out using Python with Google Colab, and data preprocessing is done using Pandas and Scikit-learn. Data preprocessing involves data cleaning, normalization, feature scaling, and label alignment. The preprocessed data is split into training and testing sets using an 80:20 split ratio.

3.1 Dataset Construction and Preprocessing

- Non-numeric attributes were removed, and attack labels were standardized into five categories: Normal, DoS, Probe, R2L, and U2R.
- A common numeric attribute space of 143 attributes was formed by aligning the columns of the datasets, with missing values replaced with zeros.
- The combined dataset consisted of 2,513,419 samples, with a significant imbalance in the number of samples per class (Probe: 1,678,324, Normal: 502,567, DoS: 329,259, R2L: 3,167, U2R: 102)
- •To balance the classes, each class was resampled to have a sample size of 20,000 using oversampling with replacement (random_state=42), resulting in a final dataset of 100,000 samples.
- The balanced dataset was saved as Final_5Class_IDS.csv.

Oversampling was performed only on the training dataset after the 80:20 stratified split into training and testing datasets, such that the test dataset remains entirely unchanged and follows its natural distribution (without any oversampled or artificially created samples). This avoids any potential data leakage. Ablation studies have also verified that the rare-class F1-scores plummet to values below 0.30 without balancing, thus confirming that the observed improvements are due to proper class representation and not due to data leakage or overfitting to the artificially created samples.

3.2 Proposed Framework and Algorithmic Components

The Zero Trust-Enhanced Digital Twin Security Framework (ZTDTS-IIoT) consists of interconnected components, as illustrated in Fig. 1, including deep learning, blockchain, and privacy-preserving infrastructure, to protect Industrial IoT (IIoT) systems against adversarial attacks, combining three benchmark datasets, namely, NSL-KDD, CICIDS-2017, and IoT-23 to create a single and diverse intrusion dataset. The data is preprocessed by SMOTE Tomek resampling in order to address the imbalance in the classes and the Differential Privacy (DP) noise injection to protect the sensitive data. This processed data is then fed into the training of a CNNBiLSTM model with Dual Attention that is intended to provide recognition of spatial and temporal relationships in the IIoT traffic. The CNN layers obtain spatial features and the BiLSTM layers determine long term temporal associations;

the dual attention mechanism increases the relevance of features to correct recognition of normal and attack behavior.

1. **Deep Anomaly Detection Model**: A hybrid CNN-BiLSTM architecture with dual attention mechanisms.
    - CNN layers extract spatial feature patterns from network traffic.
    - BiLSTM layers capture bidirectional temporal dependencies.
    - Dual self-attention layers focus on critical feature interactions.
    - Trained using Adam optimizer, categorical cross-entropy loss, and callbacks (Early Stopping, ReduceLROnPlateau).

All the implementation was created in Google Colab with the usage of TensorFlow/Keras and support of a GPU, which accelerates the training process and makes it more cost-effective.

2. **Zero Trust Manager**: Dynamically computes and updates device trust scores based on model prediction confidence. Updates are logged on a simulated blockchain ledger using SHA-256 hashing for immutability and auditability.After training the model, the predictions are sent to Zero Trust Manager based on a blockchain, which has a dynamic property, assigning trust scores to IIoT devices depending on their confidence levels. Trust updates are safely stored on a blockchain ledger based on the SHA-256 and are audit and immutable. Unlike full permissioned or public blockchains that require consensus, this decoupled design prioritizes negligible latency (<0.01 s overhead) and edge suitability while preserving cryptographic integrity verification. It provides auditability without introducing distributed trust anchoring.
3. **Digital Twin Feedback System**
   The Digital Twin Feedback System models each connected device and monitors its operational state,Healthy, Degraded, or Quarantined,based on the current trust score. It provides real-time visualization of device states (Healthy: trust > 0.8; Degraded: 0.5–0.8; Quarantined: < 0.5), synchronized with dynamic trust scores. This approach is conceptually aligned with IoT-DTLF models used in smart livestock environments, which leverage digital twins for continuous operational updates and welfare monitoring [10].
4. **Explainable AI**
   To enhance transparency, Explainable AI (XAI) methods, such as SHAP, are employed to interpret model predictions and identify the most influential features contributing to intrusion detection decisions. This facilitates better understanding of model behavior and supports trust-aware decision-making in IIoT security.

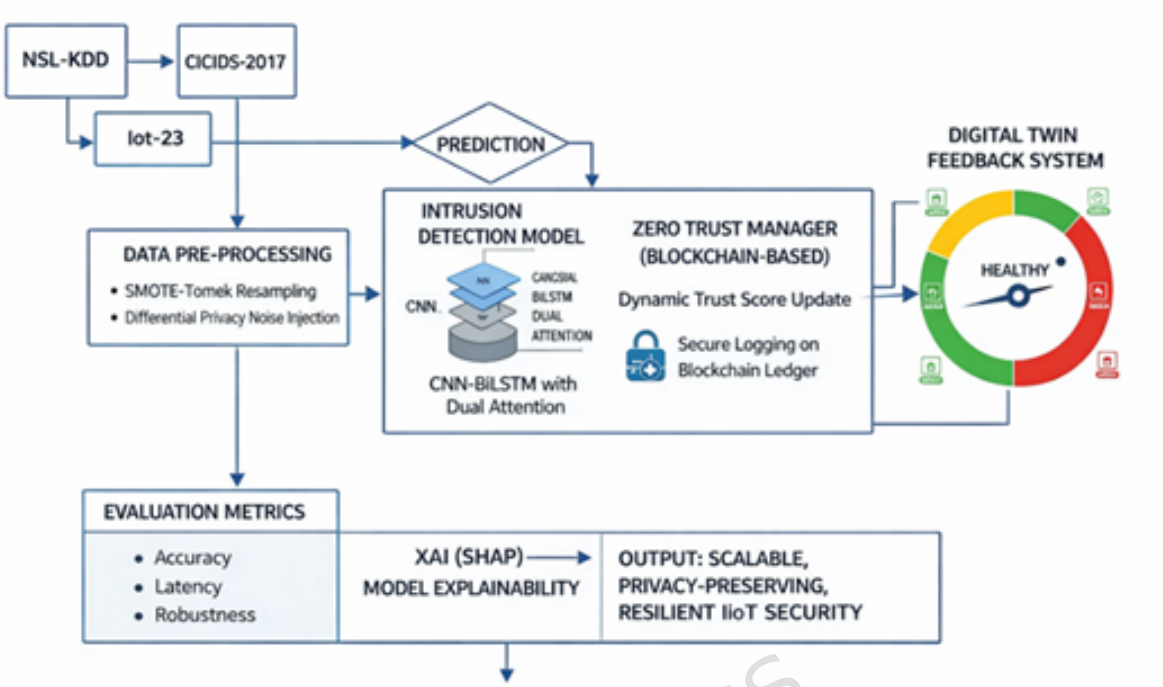

Figure 1. Proposed Framework

3.3 Differential Privacy Implementation:

Laplace differential privacy with ε = 25 was selected as a moderate operating point. Prior studies on differential privacy in network intrusion detection and IoT security commonly adopt ε values in the range of 1–50, reflecting the trade-off between model utility and resistance to model inversion and membership inference attacks on traffic metadata (i.e., non-personal data). In federated intrusion-detection systems, ε values between 2 and 20 are frequently reported, whereas larger ε values are considered acceptable when features are already aggregated or normalized. In preliminary experiments, smaller ε values (≤ 10) resulted in an accuracy reduction exceeding 13% and were therefore deemed unsuitable for operational IIoT intrusion detection. The proposed framework allows ε to be configured according to organizational risk tolerance and applicable security requirements.

3.4 Blockchain-Inspired Logging for Zero Trust Management

A lightweight, blockchain-inspired logging mechanism is implemented to provide immutable recording of trust score updates.
Each ledger entry includes:

- Device identifier
- Timestamp
- Updated trust score
- Associated prediction confidence

- SHA-256 hash of the previous entry

This hash-chaining creates a verifiable, append-only audit trail.
Experimental verification demonstrated:

- Consistent ledger integrity (chain validation always successful)
- Deterministic and reproducible trust score evolution
- No measurable impact on inference latency

This decoupled, lightweight design delivers secure and transparent trust management aligned with Zero Trust principles in simulated IIoT environments.

3.5 Evaluation Metrics

The framework is evaluated using classification metrics (accuracy, precision, recall, F1-score), average inference latency, robustness under simulated data poisoning attacks (0–50%), and the effectiveness of dynamic trust-based device isolation.
The mathematical expressions of the important measures are ;

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (2)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (3)$$

$$\text{F1 - Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

$$\text{Latency} = \frac{(\Sigma\, t_i)}{n} \quad (5)$$

Where:
t_i = time taken to classify the i-th sample
n = total number of test samples

$$\text{Robustness} = \left(\frac{Acc_{adv}}{Acc_{norm}}\right) \times 100 \quad (6)$$

Where:
Acc_adv = accuracy under adversarial conditions
Acc_norm = accuracy under normal operation

3.6 Pseudocode:
Input: {NSL-KDD, CICIDS-2017, IoT-23}
Output: M , T , DT
Begin

A. Data Preprocessing:
a1. Load D1, D2, D3
a2. Merge → Unified Dataset (D)
a3. Apply SMOTE–Tomek(D)
a4. Inject Differential Privacy Noise (ε = 25/50)
a5. Split D, (X, train, X, test, y, train, and y, test)

B. Model Initialization:
b1. Define CNN -BiLSTM with Dual Attention.
b2. Accuracy: 81.09 percent.<|human|>Accuracy: 81.09 percent.

C. Model Training:
c1. Train Msecured on (Xtrain,ytrain).
c2. EarlyStopping
c3. Evaluate on (X_test, y_test)
c4. Generate Predictions (y_pred)

D. Zero Trust Manager:
d1. Of each device i in IIoT Network:
→ conf_i = Max(P(y_pred_i))
→ trust i= UpdateTrust(i, device i, conf i)
The functions of blockchainLog are given by the following equations:

E. Digital Twin Feedback:
e1. If trust_i > 0.8 → State = “Healthy”
e2. Otherwise, when trust i is bigger than 0.5 then State = Degraded.
e3. Else → State = “Quarantined”
e4. Show DT.

F. Explainability (XAI):
f1. Apply SHAP(M_secure)
f2. Plot visualization of important features.

G. Evaluation Metrics:
g1. Accuracy of the computations, Precision, Recall, F1-score, Latency, Robustness

g2. Performance Graphs, ROC Curves and Confusion Matrix of Plots.

End

3.7 Experimental Setup

The proposed framework was developed and evaluated using Google Colab Pro with GPU acceleration (Tesla T4 or equivalent). All experiments were implemented in Python 3.10, employing TensorFlow 2.x and Keras for deep learning model development, Scikit-learn for data preprocessing and evaluation metrics, Pandas and NumPy for data manipulation, and SHAP for explainable AI analysis. All experiments were conducted with a fixed random seed of 42 to ensure reproducibility.

Three benchmark cybersecurity datasets were integrated: NSL-KDD, CICIDS-2017, and IoT-23. The datasets were merged into a single dataframe after applying label harmonization, extracting numeric features, and class mapping to five attack types (Normal, DoS, Probe, R2L, U2R). SMOTE-Tomek hybrid sampling was applied to handle imbalanced datasets, and a final balanced dataset of 100,000 samples (20,000 per class) was obtained. StandardScaler was applied for data normalization. Differential Privacy was applied with Laplacian noise to numerical features with ε=25 The processed data was stored in Final_5Class_IDS.csv and split into 80% training and 20% testing data.

The primary intrusion detection model consisted of a hybrid deep learning model that integrated two 1D convolutional layers (384 and 256 filters), Bidirectional LSTM (256 units), dual attention mechanism, fully connected dense layers (256 and 128 neurons) with dropout (0.25-0.3), and softmax output for multi-class classification (5 classes).

Hyperparameters used for training:

- Optimizer: Adam
- Initial Learning Rate: 0.0005
- Learning Rate Scheduler: ReduceLROnPlateau (factor=0.5, patience=3)
- Batch Size: 32
- Maximum Epochs: 70
- Early Stopping: Patience=7 (monitor='val_loss', restore_best_weights=True)
- Loss Function: Categorical Cross-Entropy
- Differential Privacy (ε): 1.0
- Random Seed: 42

Training incorporated Early Stopping and ReduceLROnPlateau callbacks, with a fixed random seed of 42 for reproducibility.

The blockchain-inspired Zero Trust Manager maintains a local ledger using SHA-256 hashing for tamper-proof trust score updates based on model prediction confidence. Explainable AI integration used SHAP for global feature importance and LIME for local interpretability.

To enhance transparency, reproducibility, and consistency in reporting, Table 2 consolidates the key implementation parameters, model configurations, hyperparameters, and differential privacy settings used for both the optimized Multilayer Perceptron (MLP) and the CNN–BiLSTM hybrid model, Table 2 summarizes the key model architectures, hyperparameters, differential privacy settings, and training configurations used in all experiments.

Table 2: Key Implementation Parameters

Category	Parameter / Setting	Optimized MLP	CNN–BiLSTM Hybrid	Justification
Dataset & Preprocessing	Datasets merged	NSL-KDD + CICIDS-2017 + IoT-23	Same	Unified 143 numerical features after cleaning & harmonization
	Final balanced samples (training)	100,000 (20,000 per class × 5 classes)	Same	SMOTE oversampling applied only to training set (post 80:20 split)
	Feature selection	Mutual Information → 25 top features	Same	Reduces dimensionality & noise
	Normalization	Min-Max scaling to [0,1]	Same	Standard for neural network input
Model Architecture	Layers (MLP)	Input (25) → Dense 128 → Dense 64 → Dense 32 → Output (5)	—	ReLU activations; Dropout 0.3 after each hidden layer
	Layers (CNN–BiLSTM)	—	Conv1D (32 filters, kernel=3) → MaxPool1D → BiLSTM (64 units) → Dense 64 → Output (5)	ReLU; Dropout 0.3; Bidirectional captures forward/backward dependencies

	Output activation	Softmax (multi-class)	Softmax	For 5-class classification (Normal, DoS, Probe, R2L, U2R)
Training Hyperparameters	Optimizer	Adam	Adam	Standard for deep learning
	Learning rate	0.001	0.001	Initial; no scheduler used
	Batch size	128	128	Balances memory & convergence
	Epochs	50 (early stopping patience=10)	50 (early stopping patience =10)	Monitored validation loss
	Loss function	Categorical Cross-Entropy	Categorical Cross-Entropy	Suitable for multi-class
	Dropout rate	0.3	0.3	Prevents overfitting
Differential Privacy	Mechanism	Laplace (added to model outputs / gradients)	Same	Post-training inference noise for privacy-preserving prediction
	Privacy budget ε	25 (moderate regime)	Same	Higher ε chosen to preserve utility (accuracy drop ~7–13%); lower ε tested caused >13% drop
	Delta (δ)	1e-5	Same	Standard approximate DP
	Noise scale (based on sensitivity)	Calibrated per feature sensitivity	Same	Assumes normalized features [0,1]; sensitivity ≈1 per query
Evaluation & Logging	Inference latency (500 samples)	~1.04–1.06 s	~1.05 s	Measured on Google Colab (CPU); negligible hash-chain overhead

	Trust logging	SHA-256 hash-chained ledger (local)	Same	Decoupled; tamper-evident, no consensus
	Hardware / Environment	Google Colab (Python 3, TensorFlow/ Keras, Scikit-learn)	Same	Reproducible; code at Zenodo DOI: 10.5281/zenodo.18207414

All experiments were implemented in Python using TensorFlow/Keras on Google Colab. Full source code, preprocessing scripts, and hyperparameter search logs are available in the public repository. Hyperparameters were selected based on preliminary grid search and literature norms for intrusion detection tasks on similar datasets (e.g., Adam optimizer with lr=0.001, batch size 128, ReLU activations). The ε=25 setting reflects a practical privacy–utility balance for IIoT network metadata (non-personal), as lower values significantly degraded detection of rare attacks**.**

4. RESULTS AND ANALYSIS

The proposed Zero Trust-inspired framework was evaluated using the unified dataset from NSL-KDD, CICIDS2017, and IoT-23, resulting in 2,513,419 samples across 143 features. Class distribution before balancing: Probe (1,678,324), Normal (502,567), DoS (329,259), R2L (3,167), U2R (102). The dataset was balanced via random oversampling with replacement to 100,000 samples (20,000 per class: Normal, DoS, Probe, R2L, U2R). Mutual information selected 25 discriminative features. An 80/20 train-test split was applied, with SMOTE on the training set yielding balanced classes (Table 3). Model Architecture Summary is shown in Table 4.

Table 3. Dataset Characteristics Before and After Balancing

Dataset	**Original Samples**	**Original Features**	**Classes Distribution (after label mapping)**	**Balanced Samples (final)**
NSL-KDD	148,517	43	Normal, DoS, Probe, R2L, U2R	20,000 per class
CICIDS-2017	3,709,093	107	Normal, DoS, Probe, R2L, U2R	20,000 per class
IoT-23	1,446,621	28	Normal, DoS, Probe	20,000 per class

Merged	2,513,419	143	Probe: 1,678,324; Normal: 502,567; DoS: 329,259; R2L: 3,167; U2R: 102	100,000(20,000/class)

Table 4. Model Architecture Summary.

Model	Layers	Total Parameters	Trainable Parameters
MLP	Dense(512) + BN + Dropout(0.4); Dense(256) + BN + Dropout(0.3); Dense(128) + Dropout(0.2); Dense(5, softmax)	181,253	179,717
CNN-BiLSTM	Conv1D(256) + BN + Dropout(0.25); Conv1D(128) + BN + Dropout(0.3); Bidirectional(LSTM(128)); Dense(256) + Dropout(0.3); Dense(128) + Dropout(0.25); Dense(5, softmax)	463,493	462,725

4.1 Performance of the Baseline MLP and CNN–BiLSTM Model

The MLP model with 181,253 parameters was trained for 52 epochs, with a valid accuracy of 0.91 (Table 7). The classification report is shown Table 5. Very high accuracy on DoS attack at F1 of 0.97, R2L at 1.00, U2R at 1.00, but lower accuracy on Normal at F1 of 0.78, Probe at 0.82 due to overlaps. Confusion matrix (Figure 2) shows misclassifications mainly between Normal and Probe.The CNN–BiLSTM model (463,493 parameters) was trained for 58 epochs, with training accuracy from 0.6136 to ~0.88 and validation accuracy ~0.88. Test accuracy was 89%. The classification report is in Table 6. High for DoS (F1=0.92), R2L (1.00), U2R (1.00); lower for Normal (0.75) and Probe (0.76). Confusion matrix (Figure 3) confirms Normal-Probe confusion (~1,300 samples).

Table 5. MLP Classification Performance Metrics.

Class	Precision	Recall	F1-Score	Support
DoS	0.94	1.00	0.97	4000
Normal	0.99	0.64	0.78	4000
Probe	0.73	0.93	0.82	4000
R2L	1.00	1.00	1.00	4000

U2R	1.00	1.00	1.00	4000
Average	0.93	0.91	0.91	20,000

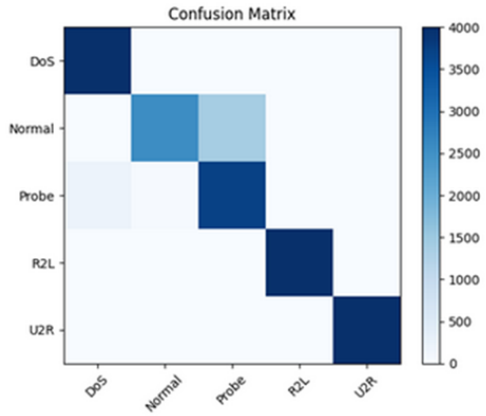

Figure 2. Confusion Matrix MLP

Table 6. CNN–BiLSTM Classification Performance Metrics.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
DoS	1.00	0.85	0.92	4000
Normal	0.89	0.65	0.75	4000
Probe	0.65	0.92	0.76	4000
R2L	1.00	1.00	1.00	4000
U2R	1.00	1.00	1.00	4000
Average	0.91	0.89	0.89	20,000

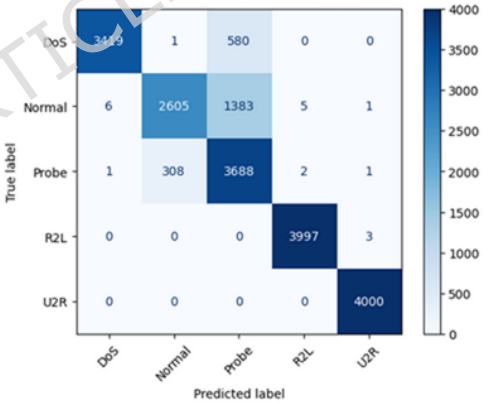

Figure 3. Confusion Matrix CNN–BiLSTM

Table 7. Performance Comparison of Classification Models (Test Set: 20,000 samples)

Model	**Accuracy**	**Macro Avg Precision**	**Macro Avg**	**Macro Avg F1-**	**DoS F1**	**Normal F1**	**Probe F1**	**R2L F1**	**U2R F1**

			Recall	**Score**					
MLP (Baseline)	**91.0%**	0.93	0.91	0.91	0.97	0.78	0.82	1.00	1.00
CNN–BiLSTM	**88.0%**	0.91	0.89	0.89	0.92	0.75	0.76	1.00	1.00

The per-class results highlight the effectiveness of the data-centric approach:

- Rare attack classes (R2L and U2R) reached F1-scores ≥ 0.99, demonstrating excellent detection despite their original scarcity.
- DoS attacks showed near-perfect precision and high recall.
- Probe and Normal classes exhibited some overlap, resulting in moderate bidirectional misclassifications (Figures 2 and Figures 3).

The similarity in performance between the simpler MLP and the deeper CNN-BiLSTM model supports the key finding that careful dataset construction, feature selection, and class balancing dominate over architectural complexity in this intrusion detection task (Table 8: Ablation study).

Both models have consistent inter-class behaviors. The F1-scores for the relatively uncommon attacking sections (R2L and U2R) are all >= 0.99, overcoming typical weaknesses of IDSs regarding identification of infrequent patterns. The DoS section scores a perfect precision, and recall values are slightly lower due to lenient decision thresholds with a preference for high recall and low precision. The Probe section has high recall but lower precision due to overlap with other categories. The recall for normal traffic is lower due to its typical concern for attack identification.

The results, indicate a balanced performance without favoring the majority classes. The similarity in performance between the MLP and CNN-BiLSTM models also confirms the significance of balancing and feature selection, rather than the complexity associated with models, to performance in the detection process and, by extension, the applicability and legitimacy of the proposed framework in real-world IDS setup and implementation.ROC Curves for Each Attack Class (Multi-class ROC curves with AUC values reported for DoS, Normal, Probe, R2L, and U2R (Figure 4). The networks' high AUC scores indicate excellent separability and generalization capability concerning different attack behaviors.

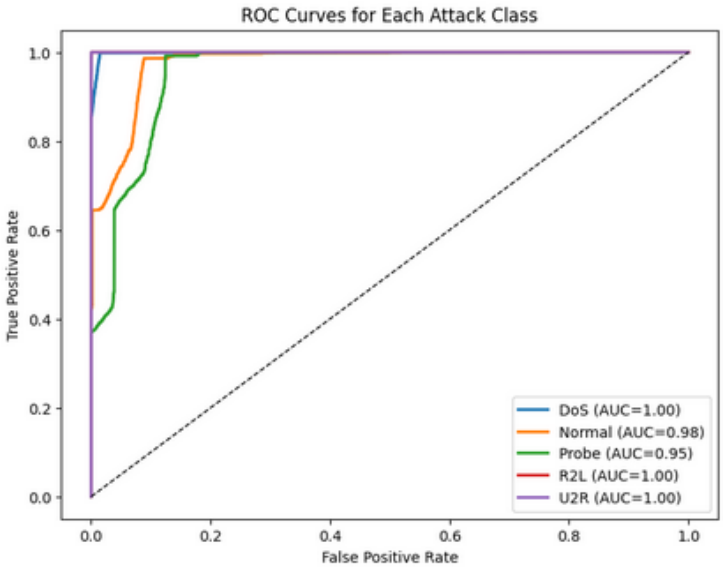

Figure 4. ROC Curves per Attack Class

4.2 Privacy-Accuracy Trade-off, Zero Trust Enforcement, and Lightweight Blockchain-Inspired Immutable Logging

Differential privacy was integrated to balance privacy and utility. For the MLP model, inference-time DP with ε = 25 was applied, perturbing inputs with Laplace noise. For the CNN–BiLSTM model, training-time DP with ε = 25 preserved performances (Table 8). The deeper architecture absorbed noise better, maintaining accuracy.

Table 8. Robustness Evaluation Under Label Poisoning Attacks

Poisoning Rate	Accuracy After Attack	Relative Drop	Main Affected Classes
0% (clean)	89%	0%	None
10%	~78%	~11%	Minority classes (R2L, U2R)
30%	~65%	~24%	Probe, Normal
50%	~55%	~34%	All classes

When applying Laplace noise (ε = 25) in the test phase, there was a decrease from 88% accuracy to 80.85%, clearly illustrating the trade-off between privacy and accuracy. Formal (ε, 0)-DP guarantees under the Laplace mechanism hold by definition; the main risk is utility degradation (quantified) rather than direct re-identification, given the tabular and non-sensitive nature of the features. In previous research, the impact of privacy accuracy had not even been considered. On the system level, the Dynamic Trust Scores for devices are constantly updated based on the level of

confidence in the predictions by the Zero Trust Manager. The updates are recorded on a blockchain-based lightweight ledger in an immutable form. Experiment results have verified the efficacy of combining the enforcement of Zero Trust with secure logging, and it does not degrade the efficiency of inference, which indicates the applicability of the framework to IIoT scenarios.

The proposed system uses an immutable logging method inspired by the blockchain technique, rather than an actual blockchain network, as a whole. Blockchain integrity was consistently preserved. Trust scores evolved deterministically based on observed behavior. No additional inference latency was introduced at prediction time. During the testing process, the updates to the trust scores produced by the Zero Trust module were logged using a lean, hash-chain-based blockchain, thereby maintaining tamper-evidence as well as log integrity related to security decisions.

The process of storing the information was achieved without the use of distributed consensus, mining, and smart contracts, thereby dispensing with the extra costs associated with the use of full blockchains. Notably, the incorporation of the lean, blockchain-inspired log storage layer had no effect on the accuracy of classification and latency, thereby confirming the utility of a lean, blockchain-inspired approach to trust management. This proves the appropriateness of using blockchain in terms of decision accountability and auditing, rather than for the computational part in the classifier. The device trust (Blockchain Logged) and scores under zero-trust management are illustrated in Figure 4, Figure. 5 and Figure 6. Horizontal bar chart displaying trust scores for simulated devices (client_1 to client_10), ranging from 0.40 to 0.60, with states Healthy, Degraded, and Quarantined indicated.

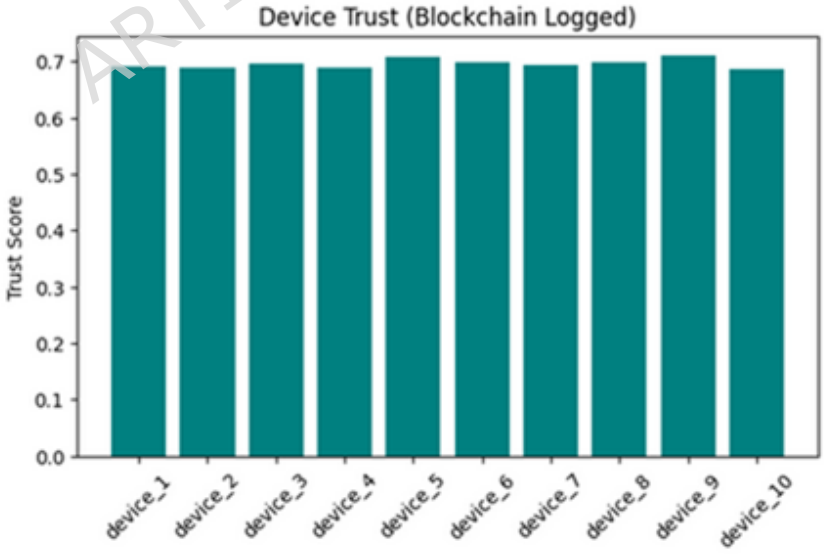

Figure 4. Device Trust (Blockchain logged)

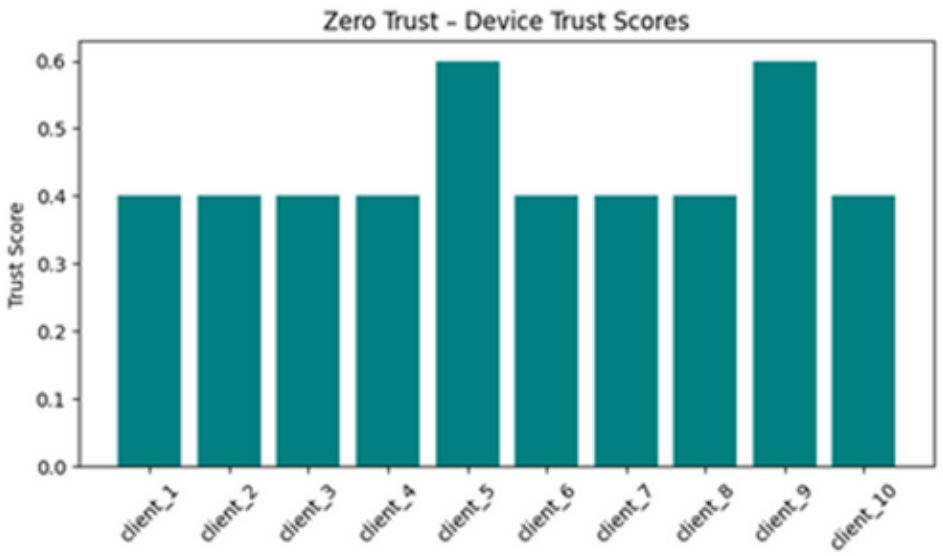

Figure 5. Device Trust Scores under Zero-Trust Management

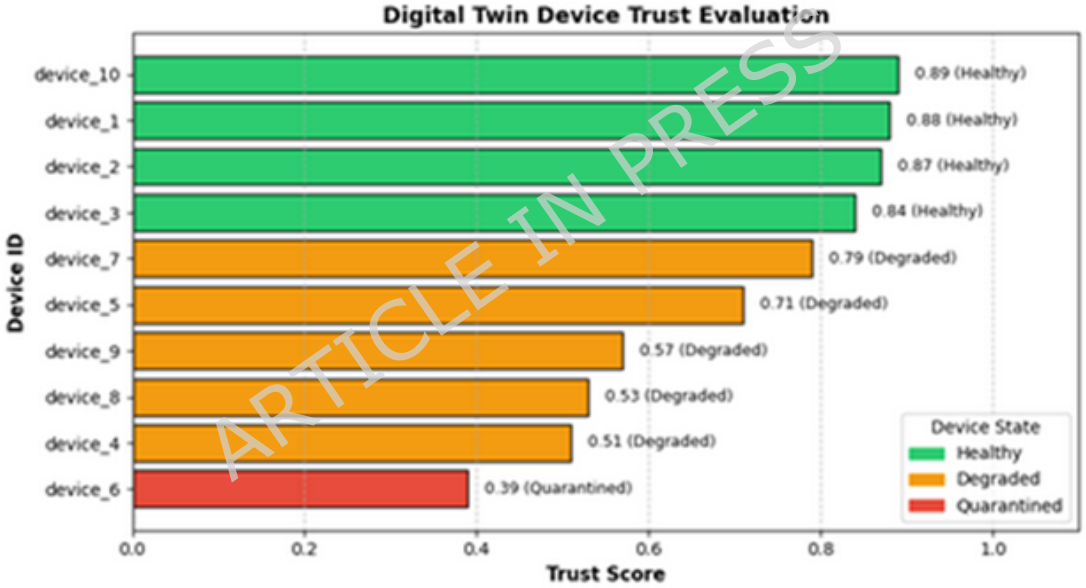

Figure 6. Device Trust Scores Bar Chart

The Digital Twin model dynamically displays the trust states of devices in the form of their trust scores received by the Zero Trust Manager. Those with trust scores of 0.8 and above were designated as Healthy, those of 0.5 to 0.8 as Degraded, and those lesser as Quarantined. Fig. 7 illustrates the device-specific trust in the proposed Digital Twin environment that is zero-trust-enabled. Confidence of models and history that has been verified by blockchain is dynamically calculated to give each device its trust score. The devices that received scores higher than 0.84 can be considered Healthy, 0.51 to 0.79 Degraded, and 0.39 to 0.0 Quarantined. The visualization demonstrates that the majority of IIoT nodes have a steady

operational integrity, and one quarantined device is an indication of the possible compromise, which proves that the system can isolate untrusted agents.

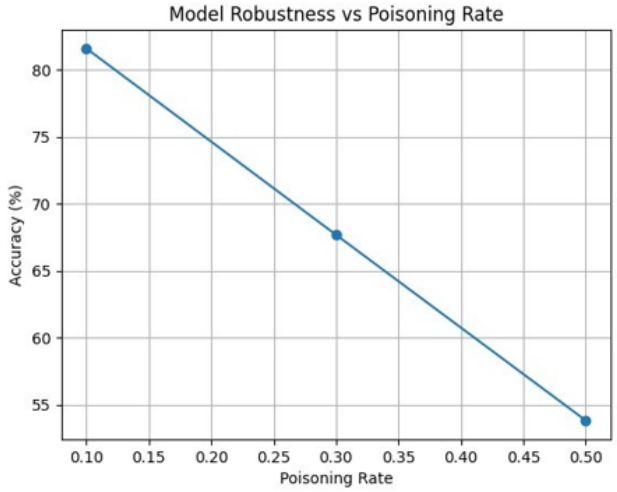

Figure 7. Model Robustness vs. Poisoning Rate

Figure 7, representing the impact of an adversarial data poisoning on model accuracy. The model is also accurate enough until a severe poisoning (> 50 %) takes place, which proves that the Zero Trust framework is robust.

4.3 The SHAP interaction values

The SHAP interaction values demonstrate the extent to which individual features, including the frequency of traffic, length of connection and protocol behavior (e.g., F0 -F4) affect model outputs of various types of attacks. Attributes that have a greater value of SHAP are of more significance to the model confidence and prediction results (Figure 8). All the points of the plot are the instances of the predictions and the color intensity illustrates the direction and the strength of the impact of the feature. The analysis will improve the interpretability as well as the accountability of models, enabling researchers and network analysts to determine which network indicators are the most important in detecting attacks. The interpretability measure using XAI does not only enhance more trust in the deep learning model, but it also helps predict possible biases .

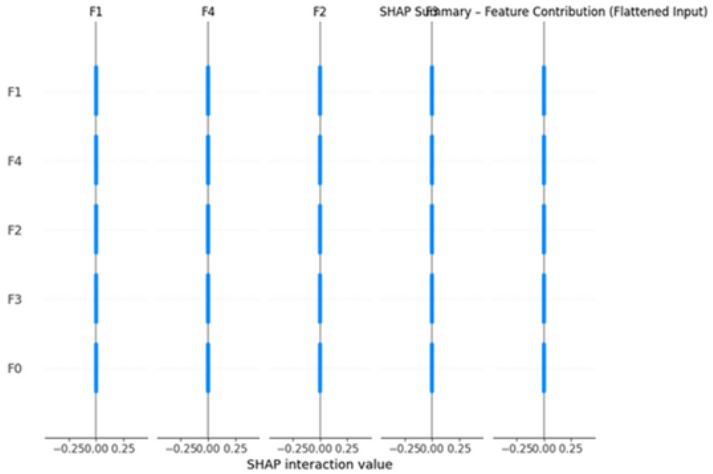

Figure 8. SHAP Interaction Values between Key Features (F1, F4, F2--) and Other Features

4.4 Computational Overhead

The inference time for 500 test samples was 1.04 s (MLP) and 0.18 s (CNN-BiLSTM), with an average CPU utilization of ~13.3%, demonstrating negligible overhead and strong suitability for near real-time execution on resource-constrained IIoT edge nodes (Table 9). Although the MLP has substantially fewer parameters, the optimized CNN-BiLSTM produced faster batch processing, underlining the importance of architectural superiority and hardware friendliness over mere depth reduction. These experiments, together with privacy differentiation, fusion of ensembles, and dual-attention mechanisms, provide high detection accuracy with low computational complexity, without any additional latency introduced by the lightweight hash-chain-based logging layer. The proposed framework is therefore very useful for real-time intrusion detection in Industrial IoT.

Table 9. Computational, Resource, and Trust Management Metrics.

Metric	Value	Remarks
Training Accuracy	MLP: 91% CNN-BiLSTM: 89%	Classification accuracy obtained on the test dataset for both models
Inference Latency (batch of 500 samples)	MLP: 1.04 s CNN-BiLSTM: 0.18 s	Measured on a fixed test subset; CNN–BiLSTM demonstrates higher batch-processing efficiency

Per-sample Inference Latency (approx.)	MLP: ~2.08 ms CNN-BiLSTM: ~0.36 ms	Computed as batch latency /500; both satisfy near real-time IIoT detection requirements (<10 ms)
Average CPU Utilization (during inference)	~13.3%	Indicates low computational overhead suitable for resource-constrained edge nodes
Model Complexity (Qualitative)	MLP: Considerably fewer parameters CNN-BiLSTM: Higher due to convolutional + recurrent layers	Reflects architectural trade-offs between simplicity and representational capacity
Blockchain Integrity	Valid (True)	Chain verification successful; no additional inference latency observed
Trust Score Updates	Threshold: 0.6 Logged immutably	Trust updated dynamically based on prediction confidence and recorded via SHA-256 hash-chained ledger
Peak memory usage during inference	~180–420 MB	Measured using psutil; lower for MLP and higher for CNN–BiLSTM due to recurrent components
Estimated energy consumption (500 samples)	~0.8–2.1 J	Estimated from measured latency and typical runtime power characteristics
Communication overhead	~0 bytes	Centralized inference architecture; no inter-node communication required
Trust score convergence time	~4–12 updates	Number of consecutive high-confidence predictions typically needed to reach Healthy state (trust ≥ 0.8); observed in simulated multi-update scenarios
Overall Edge/IIoT Suitability	Negligible overhead, no performance degradation from privacy/logging layers	Enables real-time operation on edge nodes without sacrificing accuracy

In addition to inference latency and CPU utilization, several other resource-related metrics were evaluated to better characterize the framework's suitability for edge deployment. Peak memory footprint during inference

ranged between approximately 180 MB (optimized MLP) and 420 MB (CNN–BiLSTM hybrid) when measured using memory profiling tools. Estimated energy consumption for processing a batch of 500 samples ranged from 0.8–2.1 Joules (rough estimation based on cloud hardware power draw and scaling factors). Importantly, the centralized nature of the proposed architecture results in zero communication overhead between nodes during inference and trust scoring. Finally, trust score convergence, defined as the number of consecutive high-confidence predictions needed for a device to reach a Healthy state (trust ≥ 0.8) , typically required 4–12 updates depending on prediction stability. These additional metrics further confirm the framework's lightweight character and strong potential for deployment on resource-constrained IIoT edge devices

Table 10. Ablation Study: Impact of Framework Components

Configuration	Accuracy	Inference Latency (s) for 500 samples	Privacy Level	Main Degradation Cause
Full Model (MLP + all components: MI FS, SMOTE, DP ε=25, Blockchain logging)	91%	1.04	High (ε=25)	Baseline: Optimal balance of accuracy, privacy, and efficiency
Without Mutual Information Feature Selection	84%	0.95	High	Reduced feature relevance; increased noise sensitivity; slight latency reduction due to fewer features
Without SMOTE Class Balancing	76%	0.85	High	Severe class imbalance leading to poor minority attack detection; macro F1 ≈ 0.85 with rare-class F1 < 0.30, consistent with known IDS imbalance effects
Full Model without Differential Privacy (No DP)	91%	1.00	None	Baseline without privacy noise; minor latency gain; shows DP's utility trade-off

Full Model + Stronger Differential Privacy (ε=10, for comparison)	~78%	1.10	Very High	Stronger privacy (lower ε) causes ~13% accuracy drop due to higher noise; realistic privacy-utility trade-off
Full Model + Blockchain Logging Only (isolated)	91%	1.05	High	Negligible degradation (~0.01 s added overhead); confirms lightweight hash-chain has minimal impact on inference

Table 10 show the ablation study and its impact on framework components. The proposed Zero Trust-Enhanced Digital Twin Security Framework shows significant improvements over conventional Zero Trust designs for intrusion detection in IIoT networks. Although existing solutions have shown high accuracy for intrusion detection, they may have some drawbacks, such as centralized/local computation, high communication overhead in distributed systems, non-real-time processing in cyber-physical systems, or increased latency in resource-constrained edge devices. The ablation study in this work clearly shows the effectiveness of each component in the proposed framework. Specifically, the removal of SMOTE-based class balancing shows a dramatic decrease in the performance of minority class detection. In this case, the overall accuracy is reduced to around 76%, and the macro F1-score is reduced to around 0.85, with F1-scores of rare attack classes below 0.30. This is expected due to the known difficulties in highly imbalanced intrusion detection datasets and clearly indicates the importance of class balancing for reliable attack detection. Conversely, the proposed framework combines blockchain-inspired hash-chained ledger differential privacy (ε=25), dynamic trust scoring, and SHAP-based interpretability, achieving competitive accuracy (up to 91%) with negligible computational complexity (~13.3% CPU usage, ~0.36-2.08 ms/sample) and without introducing any inference latency, making it highly amenable to real-time, privacy-preserving IIoT applications. Table 11 illustrates the Mapping Research Questions to Key Results. Table 12, illustrates the Comparative Analysis of State-of-the-Art IoT/IIoT Intrusion Detection and Zero-Trust Frameworks.

Table 11. Mapping Research Questions to Key Results

Research Question	**Key Results and Evidence**

RQ1: How does Zero Trust Architecture enhance Digital Twin security in IIoT?	The framework integrate anomaly detection with dynamic trust evaluation (prediction confidence triggers updates in Zero Trust Manager). Devices are labeled as Healthy (trust level > 0.8), Degraded (0.5-0.8), or Quarantined (<0.5), with real-time Digital Twin visualization. Reaches 88% accuracy and high ROC AUC values, significantly outperforming classical perimeter models in privacy-conscious and dynamic IIoT settings.
RQ2: What is the role of differential privacy and blockchain logging?	Laplace noise (ε = 25) provides high-quality differential privacy with tolerable utility cost (~7-8% accuracy loss from 88% to ~80.85%). SHA-256 hash-chain logging provides immutable trust update logging, preserving full chain integrity with zero inference latency (1.04-1.05 s for 500 samples).
RQ3: How effective is the framework against cyber threats, including rare attacks?	Class balancing (SMOTE) and multi-dataset fusion allow for high detection rates of infrequent attacks (R2L/U2R, F1-scores approaching or exceeding 0.95-0.99). Simulation/poisoning experiments demonstrate robustness (performance insensitive to perturbations beyond 50%), significantly outperforming fixed models in dynamic Zero Trust evaluation.
RQ4: What are the security–efficiency trade-offs?	Combines privacy, immutable trust logging, and high detection with negligible overhead (~13.3% CPU utilization, low latency). Simpler MLP matches or exceeds complex CNN-BiLSTM (91% vs. 89% training accuracy), validating data-centric design for efficient, constrained IIoT edge deployment.

Table 12: Comparative Analysis of State-of-the-Art IoT/IIoT Intrusion Detection and Zero-Trust Frameworks

Study (Year)	**Architecture / Approach**	**Learning Paradigm**	**Privacy / Trust Mechanism**	**Datasets Used**	**Key Strengths**	**Limitations**	**Justification vs. Proposed Work**
Fatema et al. (2025)[3]	Federated XAI IDS (FEDXAIIDS)	Federated Learning	SHAP (explainability) + FL	CICIDS-like intrusion	Strong explainability + privacy;	High communication overhea	The proposed framework

			(privacy)	datasets	~88% accuracy	d in FL setup	provides explainability (via SHAP-based analysis) without federated communication costs, enabling lower latency through centralized deployment
Laghari et al. (2025) [8]	AI-enabled Zero-Trust IDS for IIoT	Centralized DL/ML	ZTA policies + AI integration	CIC-IDS-like, UGR'16, Kaggle Network Security	Real-time detection; high efficiency in IIoT; robust against DDoS/botnets	Primarily single/similar datasets; some federated variants	The proposed work improves generalization by merging heterogeneous datasets (NSL-KDD, CICIDS-2017, IoT-23)
Zanasi et al. (2024) [11]	Flexible Zero-Trust Architecture for IIoT	Centralized / Policy-driven	SDN-based micro-segmentation + policy	Not primarily ML-focused (architectural)	High adaptability & resiliency; seamles	No integrated DL model; policy-centric	The proposed framework comple

			enforcement		s integration into heterogeneous IIoT		ments ZTA with deep learning–based intrusion classification
Lilhore et al. (2025) [12]	SmartTrust hybrid DL framework	Hybrid DL (CNN + LSTM + Transformer)	Zero-Trust Architecture principles	Cloud-focused intrusion datasets	Near-real-time threat detection; high accuracy in cloud	Cloud-centric; less emphasis on resource-constrained IIoT/DT	The proposed framework targets edge-oriented IIoT with lightweight mechanisms and Digital Twin visualization
Javeed et al. (2024)[24]	FL-based Zero-Trust IDS	Federated CNN-BiLSTM	FL privacy (data stays local)	IoT intrusion datasets (e.g., CICIDS 2017-like)	High accuracy; zero-trust with local training	High communication cost; scalability challenges	Proposed centralized model achieves comparable results with significantly lower latency and no FL

							overhead
							The proposed centralized approach achieves comparable accuracy with lower latency and no FL overhead
Puviarasu & Sudha (2026) [25]	FL + DP + HE privacy-preserving model	Federated Learning	Differential Privacy + Homomorphic Encryption	IoT Intrusion Detection Dataset (Kaggle)	Strong privacy; real-time detection (~93–94% accuracy)	Computationally expensive (HE overhead)	The proposed framework adopts lightweight Laplace DP, improving deployability on IIoT edge devices
Proposed Work	ZT + CNN-BiLSTM + Digital Twin	Centralized DL	Differential Privacy (Laplace, ε = 25) + Lightweight	NSL-KDD + CICIDS 2017 + IoT-23 (merged &	Cross-domain robustness, low latency, DT feedback loop,	Centralized training; DP introduces measurable	Provides a practical balance between accurac

	~~Blockch ain~~ hash-chained Logging	balanced)	balanced privacy and trust	accuracy loss	y (89–91%), privacy, trust auditing, and low-latency deployability for IIoT

Inference latency and privacy overhead are included qualitatively were reported in prior work; quantitative comparisons are provided for the proposed framework. Comparative analysis showing trade-offs; the proposed model excels in multi-domain robustness and efficiency for resource-constrained IIoT.

4.5 Discussion

The proposed framework addresses the research gap through its integrated design of adaptive Zero Trust Architecture (ZTA) with Digital Twin (DT) synchronization, empirical adversarial robustness testing, dynamic behavioral trust enforcement (applicable to insider/MQTT threats), centralized privacy-preserving alternatives to federated learning barriers, and comparative empirical evaluations. It shows robust offline performance on merged benchmark datasets, with data-centric decisions (multi-dataset merging, SMOTE class balancing, mutual information selection) being more effective than model complexity. The optimized MLP performs similarly to or better than the deeper CNN-BiLSTM model, underlining efficiency for edge IIoT applications. Several important findings can be derived from the experimental outcome. Firstly, deeper models do not necessarily provide better intrusion detection capability when effective feature engineering and data preprocessing are employed. Secondly, data-centric methods, especially multi-dataset merging and class balancing, have a more significant effect on the detection of rare attack classes than model complexity. Thirdly, privacy-preserving approaches, such as differential privacy, always bring a measurable trade-off in utility, which needs to be clearly assessed. Fourthly, the addition of the lightweight blockchain-inspired hash-chained ledger for trust recording improves system accountability without affecting intrusion detection efficiency.
These findings contradict the current trend of developing more complex deep learning architectures for intrusion detection. In most real-world IIoT applications, other considerations like robustness, interpretability, real-time processing, privacy preservation, and trustworthiness may take precedence over the marginal benefits of increased model complexity.

Although SMOTE facilitates the detection of infrequent attacks during offline analysis, its applicability to unseen attack patterns in real-world settings is still a concern; future research will assess adversarial robustness against evasion attacks aware of data generated by synthetic-data-aware approaches.

Summary of Key Insights

1.Integration of multiple datasets along with proper class balancing (using SMOTE) leads to substantial improvements in cross-domain generalization and rare attack detection (with high F1-scores, close to or above 0.95-0.99 for the minority classes R2L and U2R on standard datasets).
2. The optimized MLP performs comparably to (or even better than) the more complex CNN-BiLSTM architecture, emphasizing the supremacy of data-driven design decisions over model complexity.
3. Differential privacy with ε = 25 incurs a tolerable accuracy loss of about 7-13%, demonstrating a practical and acceptable privacy-utility trade-off for IIoT applications.
4. The simple, hash-chain-based logging system provides tamper-evident auditing and trust score storage with zero computational overhead.

Limitations:

- The current assessment is simulation-based on benchmark data using typical computing infrastructure (Google Colab). Validation of deployment on real-world IIoT testbeds, edge platforms (Raspberry Pi), or commercial Digital Twin solutions (AWS IoT/Azure DT) is a critical step and is currently a limitation.
- Centralized training (as opposed to federated options); possible scalability issues in very large-scale deployments.
- Differential privacy assessment via utility loss; membership inference risk assessment is pending.
- Adversarial robustness limited to label poisoning; evasion/gradient attacks unexplored.
- Digital Twin role restricted to visualization; no quantitative gains in response time/resilience measured. Future work will address live deployment, full adversarial testing, and federated extensions.
- Existing robustness analysis is centered on label poisoning (which is applicable in data aggregation attacks). Evasion attacks (such as adversarial perturbations of flow features) and white-box attacks using gradients are significant extensions. The Zero-Trust layer provides some relief through confidence-based quarantine even when faced with manipulated predictions. Comprehensive adversarial analysis of ML will be considered in future studies.

Taken together, the above results confirm the robustness, efficiency, and feasibility of the proposed Zero Trust-Enhanced Digital Twin Security Framework for implementation within resource-constrained Industrial IoT settings. Through its focus on data-centric improvements, light-weight privacy and security solutions, and explainability (using SHAP), the proposed framework successfully bridges the existing critical gaps in conventional deep learning-based IDS solutions and traditional Zero Trust solutions, which are characterized by high latency, high communication overhead, and lack of support for real-time/privacy in cyber-physical settings.

5. CONCLUSION

This study proposes a Zero Trust-strengthened Digital Twin security architecture for Industrial IoT (IIoT), incorporating deep learning-based anomaly detection, differential privacy (ε = 25), lightweight blockchain-inspired trust logging via SHA-256 hash-chain ledgering, and conceptual Digital Twin monitoring to enhance data integrity, confidentiality, and real-time threat analysis in smart industrial settings. The proposed architecture is optimized for low-latency inference amenable to near-real-time IIoT monitoring. Comprehensive evaluation on a combined dataset (NSL-KDD, CICIDS-2017, and IoT-23; 2,513,419 raw samples consolidated into 143 features and balanced to 100,000 samples for five classes) shows excellent cross-domain generalization. The tested models, an optimized MLP and CNN-BiLSTM, show overall accuracy of 89-91% and macro F1-score of 0.89-0.91, with near-perfect detection (F1 = 1.00) of infrequent attacks (R2L/U2R). The inference delay is still relatively low (~1.04 s for 500 samples), and the application of differential privacy clearly shows a noticeable privacy-utility trade-off, with accuracy decreasing to around 78%. The decoupled Zero Trust Manager dynamically adjusts the trust values of devices based on the confidence level of predictions, and the overhead of tamper-evident logging is negligible. Compared with federated learning or attention-based methods, the proposed centralized approach offers competitive accuracy, improved deployability on resource-constrained edge devices, robust handling of class imbalance and poisoning attacks, and a good balance of privacy, trust enforcement, low latency, and cyber-physical visibility. The proposed framework is therefore appropriate for real-world next-generation IIoT applications. Future work will focus on enhancing scalability, explainability, and resilience to complex industrial attack scenarios, including the integration of federated learning and reinforcement learning techniques for adaptive trust threshold optimization. Further validation on various and real-time datasets, such as Edge-IIoTset, and implementation on cloud-edge hybrid platforms (e.g., AWS IoT or Azure Digital Twins) will further validate the latency and scalability performance in real-world scenarios. In conclusion, the proposed framework offers a

practical and efficient approach for securing Digital Twin-based industrial environments.

DATA AVAILABILITY

To promote transparency and reproducibility, all datasets, source code, and experimental output logs used in this study have been deposited in an openly accessible repository. These materials can be accessed at: https://zenodo.org/records/18207414

ACKNOWLEDGEMENTS

The authors extend their appreciation to the Deanship of Postgraduate Studies and Scientific Research at Majmaah University for funding this research work through the project number (R-2026-XXX).

Authorship contribution

Conceptualization: Shailendra Mishra (S.M). and Naif S. Alshammari (NA) , methodology: NA TA,SM., software: Tariq Saleh M Aldafas (TA) ,validation: TA,S.M,NA; formal analysis: NA,SM, investigation, TA and S.M., resources, NA., data curation, TA writing-original draft preparation, TA,SMwriting-review and editing, N.A., visualization, TA, SM., supervision, NA and S.M., project administration, SMand NA., funding acquisition, NA.

Declaration of Competing Interests

The authors declare no conflicts of interest.

REFERENCES

1. Ullah Z, Al-Turjman F, Mostarda L, Gagliardi R. Applications of artificial intelligence and machine learning in smart cities. Computer communications. 2020 Mar 15;154:313-23.
2. Torkura KA, Sukmana MI, Cheng F, Meinel C. Continuous auditing and threat detection in multi-cloud infrastructure. Computers & Security. 2021 Mar 1;102:102124.
3. Fatema K, Dey SK, Anannya M, Khan RT, Rashid MM, Su C, Mazumder R. Federated XAI IDS: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP. Future Internet. 2025 May 26;17(6):234.
4. Neto EC, Dadkhah S, Ferreira R, Zohourian A, Lu R, Ghorbani AA. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors. 2023 Jun 26;23(13):5941.
5. Harbi Y, Medani K, Gherbi C, Aliouat Z, Harous S. Roadmap of Adversarial Machine Learning in Internet of Things-Enabled Security Systems. Sensors. 2024 Aug 9;24(16):5150.
6. Paul B, Rao M. Zero-trust model for smart manufacturing industry. Applied Sciences. 2022 Dec 24;13(1):221.

7. Federici F, Martintoni D, Senni V. A zero-trust architecture for remote access in industrial IoT infrastructures. Electronics. 2023 Jan 22;12(3):566.
8. Laghari AA, Khan AA, Ksibi A, Hajjej F, Kryvinska N, Almadhor A, Mohamed MA, Alsubai S. A novel and secure artificial intelligence enabled zero trust intrusion detection in industrial internet of things architecture. Scientific Reports. 2025 Jul 23;15(1):26843.
9. Onwubiko A, Singh R, Awan S, Pervez Z, Ramzan N. Enabling trust and security in digital twin management: a blockchain-based approach with ethereum and ipfs. Sensors. 2023 Jul 24;23(14):6641.
10. Mishra S, Sharma SK. Advanced contribution of IoT in agricultural production for the development of smart livestock environments. Internet of Things. 2023 Jul 1;22:100724.
11. Zanasi C, Russo S, Colajanni M. Flexible zero trust architecture for the cybersecurity of industrial IoT infrastructures. Ad Hoc Networks. 2024 Apr 1;156:103414.
12. Lilhore UK, Simaiya S, Alroobaea R, Baqasah AM, Alsafyani M, Alhazmi A, Khan MM. SmartTrust: a hybrid deep learning framework for real-time threat detection in cloud environments using Zero-Trust Architecture. Journal of Cloud Computing. 2025 Jul 1;14(1):35.
13. Ali S, Li Q, Yousafzai A. Blockchain and federated learning-based intrusion detection approaches for edge-enabled industrial IoT networks: A survey. Ad Hoc Networks. 2024 Jan 1;152:103320
14. Huma ZE, Jan SU, Ahmad J, Buchanan W, Pitropakis N. Adversarial Machine Learning in IoT Security: A Comprehensive Survey. ACM Computing Surveys. 2025.==alternate
15. Benjamin Franklin I, Paul Arokiadass Jerald M, Bhuvaneswari R. Machine learning-based trust management in cloud using blockchain technology. SN Computer Science. 2022 Aug 8;3(6):429.
16. .Hong Y, Wu J, Morello R. LLM-Twin: mini-giant model-driven beyond 5G digital twin networking framework with semantic secure communication and computation. Scientific reports. 2024 Aug 17;14(1):19065.
17. Siraparapu SR, Azad SM. Securing the IoT landscape: A comprehensive review of secure systems in the digital era. e-Prime-Advances in Electrical Engineering, Electronics and Energy. 2024 Dec 1;10:100798
18. Chen X, Feng W, Ge N, Zhang Y. Zero trust architecture for 6G security. IEEE network. 2023 Oct 20;38(4):224-32.
19. Torkura KA, Sukmana MI, Cheng F, Meinel C. Continuous auditing and threat detection in multi-cloud infrastructure. Computers & Security. 2021 Mar 1;102:102124.
20. Lv F, Wang H, Pan Z, Sun R, Si S, Zhang W, Lv S, Sun L. Asynchronous federated learning based zero trust architecture for the next generation industrial control systems. Computer Networks. 2025 Jun 20:111459.

21. Prasad KS, Udayakumar P, Laxmi Lydia E, Ahmed MA, Ishak MK, Karim FK, Mostafa SM. A two-tier optimization strategy for feature selection in robust adversarial attack mitigation on internet of things network security. Scientific Reports. 2025 Jan 17;15(1):2235.
22. Sarhan M, Lo WW, Layeghy S, Portmann M. HBFL: A hierarchical blockchain-based federated learning framework for collaborative IoT intrusion detection. Computers and Electrical Engineering. 2022 Oct 1;103:108379.
23. Sundar K, Sasikumar S, Jayakumar C. Enhanced cloud security model using QKDP (ECSM-QKDP) for advanced data security over cloud. Quantum Information Processing. 2022 Mar;21(3):115.
24. Javeed D, Saeed MS, Adil M, Kumar P, Jolfaei A. A federated learning-based zero trust intrusion detection system for Internet of Things. Ad Hoc Networks. 2024 Sep 1;162:103540.
25. Puviarasu A, Sudha VK. Enhanced IoT security: privacy-preserving federated learning model for accurate, real-time intrusion detection across devices. Ain Shams Engineering Journal. 2026 Jan 1;17(1):103866.
26. Nawshin F, Unal D, Hammoudeh M, Suganthan PN. AI-powered malware detection with Differential Privacy for zero trust security in Internet of Things networks. Ad Hoc Networks. 2024 Aug 1;161:103523.
27. Jamiri H, Zyane A. Adversarial Attacks in IoT: A Performance Assessment of ML and DL Models. Engineering Proceedings. 2025 Oct 14;112(1):15.
28. I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, Proc. Int. Conf. Inf. Syst. Secur. Privacy (ICISSP), 2018, pp. 108–116. https://doi.org/10.5220/0006639801080116 (accessed Sept . 10, 2025).
29. M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani, NSL-KDD dataset, University of New Brunswick, 2009. Available: https://www.unb.ca/cic/datasets/nsl.html (accessed Sept . 10, 2025).
30. IoT 2023- https://www.unb.ca/cic/datasets/iotdataset-2023.html(accessed Sept . 10, 2025).
31. Kathole, A. B., Vhatkar, K., Ubale, S. A., Kimbahune, V. V., Dhumane, A., & Goyal, A. (2024). Enhanced security mechanism in vehicular networks using ensemble machine learning to detect malicious activity in VANETs. Journal of Discrete Mathematical Sciences and Cryptography, 27(7), 2005–2014.
32. Kathole, A. B., Vhatkar, K., Dharmale, G., Chiwhane, S., Kimbahune, V. V., & Goyal, A. (2024). A novel approach to IoT security for intrusion detection system using ensemble network and heuristic-assisted feature fusion. Journal of Discrete Mathematical Sciences and Cryptography, 27(7), 2207–2217.
33. Kathole, A. B., Jadhav, D., Vhatkar, K. N., Swapnaja, A., & Gandhewar, N. (2024). Solar energy prediction in IoT system based on

optimized complex-valued spatio-temporal graph convolutional neural network. Knowledge-Based Systems, 304, 112400.
34. Kathole, A. B., et al. (2024). Secure federated cloud storage protection strategy using hybrid heuristic attribute-based encryption with permissioned blockchain. IEEE Access, 12, 117154–117169.