Minimax Theory and its Applications
Volume 11 (2026), No. 3s, 534 to 549

A Hybrid Transformer-BiLSTM Framework with Metaheuristic Feature Optimization for Anomaly Detection in Industrial IoT Networks

S.Kokila[1], Dr.M.Anand[2], Dr.B.Raja[3], Dr.G.Thiraviya Suyambu[4], Dr.F.Jerald[5]

[1]Research Scholar, Department of Electronics and Communication Engineering, M.A.M.School of Engineering, Tiruchirappalli, India.
[2,3]Professor, Department of Electronics and Communication Engineering, M.A.M.School of Engineering, Tiruchirappalli, India.
[4]Asst.Professor, Department of Electronics and Communication Engineering, M.A.M.School of Engineering, Tiruchirappalli, India.
[5]Associate Professor, Department of Electronics and Communication Engineering, Dr.M.G.R.Educational and Research Institute,Chennai,India
Email: sreeleth24@gmail.com[1], anand.ece@drmgrdu.ac.in[2], raja.ece@drmgrdu.ac.in[3], deanacademics@mamse.in[4], jerald.ece@drmgrdu.ac.in[5]

Abstract
Industrial Internet of Things (IIoT) networks are primary constituents in several industries such as Manufacturing, Energy, Transportation and Process Control. These networks are targeted by cyber attacks, which cause internal security lapses to the country. Traditionally, the IIoT networks suffered by three limitations such as, inadequate feature selection, failure to adapt to traffic dependencies, inability to distinguish attack classes which appears to be closer and identical but in reality they are different attack classes, could not able to address class imbalance. These bottlenecks are addressed in this research work by proposing (i) hybrid feature optimizer(Hybrid Grey Snake—Cuckoo Search (HGS-CS) metaheuristic optimizer), (ii). Combined Multi Transformer and LSTM approach for detection of attack classes with much better accuracy and (iii) Multi Dataset, cross entropy which enables the migrations are easier and reduces the class imbalance. Three datasets were used in this research are, X-IIoTID, Edge-IIoTset, and WUSTL-IIoT-2021. Multiple performance parameters such as F1 Score, Accuracy, AUC-ROC, Matthews Correlation Coefficient (MCC) etc., were used to test the proposed system. The feature reduction of 54.2%, Detection of attack classes with near perfect accuracy and cross dataset generalization and ability to distinguish multiple attacks with good t-SNE scores are contributions of this research.

Keywords: Industrial IoT, Intrusion Detection, Transformer encoder, Bidirectional LSTM, Gated Fusion , Metaheuristic optimization, HGS-CS algorithm

1. Introduction

The new era industrial operations are results of convergence of Industrial Internet of Things and Operational Technologies. AI shall be applied in both of these participating technologies [1]. Some of the crucial public impacting industries such as power generation, energy transfer (oil, gas and electricity), water management (production and treatment), manufacturing sector uses lots of sensors, real time process control systems, controllers (PLC) etc., [2]

Global IIoT endpoints poised to reach 41 billion endpoints by end of 2026 [3], which is increasing at the rate of 25% per year. This increase is also associate with increasing probabilities of adversarial attacks. It is worthy to note the Florida's Oldsmar water treatment plant attack in 2021, in which the hacker tries to poison the water tank. Ukraine energy infrastructure attack in 2022 by the anonymous sandworm group [4] and attacks against global semiconductor manufacturers Nvidia, TSMC and ASML [5]. The intrusions in Industrial IoT needs to be addresses immediately.

Intrusion detection systems (IDS)are relied on two traditional approaches namely rule-based anomaly detection and signature based detection. The hybrid approach is also used. But, they are inadequate to handle the modern Industrial IoT systems.[6]. Signature based detections could not able to detect zero-day attacks and also it needs manual updates at regular intervals. While the rule based anomaly detection produces lots of false positives.[7].

AI and Deep learning based IDS is an alternative approach. Several architectures are found in the existing literature [8,9]. Thorough examination of such architectures unveil some limitations listed below:

- Redundant Features : Typically, the IIoT datasets consists of 40 to 50 features. In this work, three datasets were used. X-IIoTID with 59 features, Edge-IIoT dataset with 61 featuresWUSTL-IIoT with 51 features. Some of the features may be redundant to the objective of the task. If all features are trained, then computational cost may increases and in some cases results into overfitting also. So, the features needs to be filtered for better efficiency.
- Architectural deficiencies: Temporal sensitive features well detected by LSTM detectors but they didn't perform well to detect global features. On the other hand, CNN detectors works good for spatial features than temporal features [12].
- Poor representation of attack class: From the publicly available IIoT datasets, the ratio of attack classes such as zero-day exploits, privilege escalation to the normal traffic data is negligible. In some cases, the attack responses are of only 0.1 to 0.5 percent of the data.[13]

This work addresses these three limitations with the following research questions:

- Is there a possibility to reduce the IIoT feature dimensionality by using a metaheuristic optimizer without compromising the detection accuracy.
- How does the BiLSTM architecture exploits both temporal and global features and perform better than encoder(single) architecture?
- If the standard weight training is replaced by cross-entropy training using inverse frequency calibration, then there will be improvement in detection of attacks?

To address these research questions, we propose, BiLSTM architecture with transformers to handle both types of features and Hybrid Hunting Grey Snake search usingmetaheuristic principles.

- With respect to mathematical modelling, the time complexity function f(P,G,d,E_r) is introduced with Population size, Generation, Dimension of feature and Evaluation Cost.
- On evaluation aspect, the class-weighted cross-entropy loss, convergence analysis, Silhouette score, cross validation, significance test and FPR are used.

Section 2 deals with review of literature I five different aspects. reviews related work across five thematic areas. Section 3 discusses the proposed architecture, datasets, proposition and algorithm of this work. In section 4, the detailed result analysis including feature optimization analysis, Training convergence, Detection Performance, Statistical significance, ROC, t-SNE analysis, Hyperparameter sensitivity, Ablation study and Computational complexity.

2. Literature review

This section covers the literature review under four different aspects such as, Intrusion detection using Machine Learning, Metaheuristic optimization, network security, class imbalance. Based on the review, the research gaps are also listed.

2.1. Intrusion detection using Machine learning

Random Forest [14] is one of the early Machine Learning technique used for intrusion detection in IIoT, which demonstrated better performance for binary classification. Later , Vector Machine [15], and k-Nearest Neighbour [16] were also used for this purpose. But when it comes to multi class high dimensional features, these techniques could not able to produce the better performance.

Mohy-Eddine et al. [17] further reduced the computational cost by applying correlation coefficient feature engineering along with Isolation Forest.For traffic IIoT data, Hamdouchi and Idri[18] applied gradient boosting and obtained higher F1 scores.Hybrid CNN-LSTM model proposed by Qathrady et al. [19] reached 99.2% accuracy for binary classification tasks.Fraihat et al. [20] proposed a bio-inspired approach using transformers with federated learning approach that yield 99.2% accuracy.

2.2. Metaheuristic optimization approach

IDS performance is heavily depends on the feature selection. Wrapper based methods, filter based methods are applied for feature selection and to reduce the dimensionality [21]. While metaheuristic approach supports high dimension data, even with hundreds of features [22]. Genetic Algorithms uses natural selection operators for feature selection, which avoid local maxima [23], while Grey Wolf Optimizer use hierarchical relationship

between the feature, which helps the IDS to fine tune the search [24]. Both Genetic algorithm and Grey Wolf Optimizer shall be used as hybrid feature selection approach reap the benefits of both schemes.

2.3. Network security for IDS

A. Vaswani et al [25] proposed an Transformer architecture, originally for signal processing. Later, it was applied for IDS also. Lin et al. [26] applied multi head attention, which work by splitting the input into multiple heads, to CNN for the problem of missing long range dependencies.Nguyen et al. [27] deployed transformer approach on the dataset of CSE-CIC-IDS2018 , which contains 80 traffic features. Though given good results, this approach is resource intensive. Some hybrid works also found in the literature. Transformer-GRU approach introduced by Yang et al. [27] shown better results than pure transformers.

2.4. Addressing Class imbalance

When the normal traffic outnumbers the attack classes, the class imbalance occurs. It may cause serious deviation on the results. The attack scenario may be reported as normal. Some standard approaches such as SMOTE-Tomek, Generative Adversarial Networks shall be applied to tackle this issue [28]. Zhang et al. [29] applied Class weighted entropy to reduce the class imbalance. The cross-entropy is used in this article to address the class imbalance in IIoT data.

2.5. Research gap

The detailed literature review identified four gaps listed below:

1. Need for a framework that addresses Architectural inadequacy and Class imbalance together.
2. Cuckoo Search and Hunting Grey Snake optimization shall be combined for feature selection to improve the performance.
3. Transformer-BiLSTM IDS approach provided better results, but resource intensive. So there is a need for transformer or hybrid approach that offers security at cheap resources.
4. Evaluation such as Wilcoxon significance tests, cross-dataset generalization shall be included to enhance the evaluation of IIoT-IDS.

3. Proposed Framework

A six layer framework is proposed in this work. Layer 1 deals with collection of data from sources. Layer 2 does pre-processing and normalization. Layer 3 comprises of the proposed Hybrid HGS-CS feature optimization method, in which cuckoo search and grey wolf optimization techniques are combined. Transformer and BiLSTM encoders comprises Layer 4. While the Layer 5 deals with Gated Fusion Mechanism. Layer 6 concerned about classification and parameter evaluation.

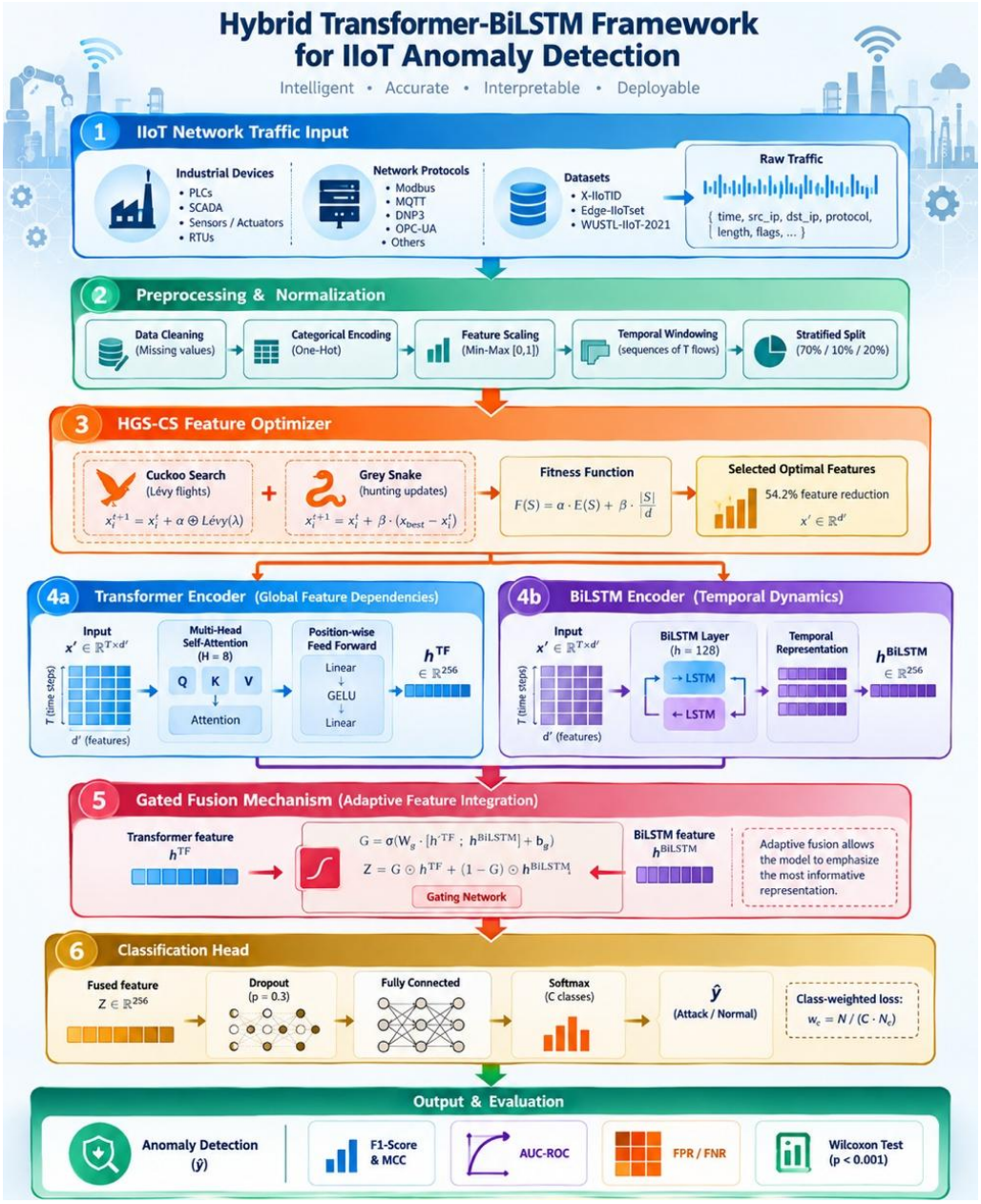


Fig.1. Proposed Hybrid Transformer – BiLSTM Framework for IIoT Anomaly Detection

3.1. Datasets used for research

Three datasets were used on this work. All those datasets are publicly available. Their features, attack classes, number of samples, attack ratio and imbalance ratio are listed in Table 1.

Table 1. Summary of Datasets used in this work

Dataset	Features	Attack Classes	Total Samples	Attack Ratio (%)	Imbalance Ratio	Protocol Coverage
X-IIoTID	59	18	1,029,267	40.2%	1.49:1	Modbus, MQTT, DNP3, OPC-UA
Edge-IIoTset	61	14	2,219,201	27.8%	2.60:1	Modbus, Zigbee, MQTT, CoAP
WUSTL-IIoT-2021	41	4	1,194,464	12.1%	7.27:1	EtherNet/IP, Modbus

3.2. Data preprocessing

Data processing begins with missing values, followed by encoding and normalization. Min-Max [0,1] normalization is used in this stage,

$$x'_j = \frac{(x_j - x_{j^{min}})}{(x_{j^{max}} - x_{j^{min}})} \quad (3.1)$$

After normalization, the datasets are split in to three groups. To% of data allocated for Training, 10% for validation and 20% for testing.

3.3. Feature optimization

A Novel HGS-CS optimization techniques is proposed in this work. Two of the existing techniques Cuckoo Search and Hunting Grey Snake are combined in this approach.

3.3.1. Optimization Procedure

Cuckoo Search provides exploration through its levy tailed steps.

$$b_{i^{t+1}} = b_{i^t} \oplus \alpha \,.\, Levy\,(\lambda) \quad (3.2)$$

While Hunting Grey Snake explore local features with the decreasing linearity from 0.7 to 0.1.

$$b_{i^{t+1}} = b_{i^t} + r_1.(b_{i^t} + r_1.(b_{best^t} - b_{i^t}) + r_2.(b_{rand^t} - b_{i^t}) \quad (3.3)$$

Here, b is the binary vector to represent the solution to feature subset.

Accordingly, the function to minimize the error shall be defined as,

$$F(s) = \propto .E_r(s) + \beta .(\frac{|S|}{d}) \quad (3.4)$$

3.3.2. Computational Complexity Analysis

As the size of input grows, the resource requirement is also grows. This demand shall be measured using the computational complexity as mentioned in equation 5.

$$T_{CC} = O\langle P.G.d.N \log N\rangle \quad (3.5)$$

For our analysis, the settings of P=40, G=50, d=60 and N=500K are used.

3.4. Proposed Hybrid detection model

3.4.1. Transformer

The selected feature subset has to be projected using transformers. The transformer proposed in this work has L encoder layers stacked one on another. Each layer applies Multi Head Attention as described in the following expressions.

$$head_h = Attention\langle EW_{hQ}, EW_{hK}, EW_{hV}\rangle \quad (3.6)$$

$$Attention\langle Q, K, V\rangle = softmax\,\langle\frac{QK^T}{\sqrt{d}k}\rangle .V \quad (3.7)$$

Further the encoded data processed by Feed forward Network, in which the feature vectors processed by two non-linear functions and two linear transformers.

3.4.2. Bidirectional LSTM

This variant of Long and Short Term Memory processes the data in both directions. Temporal dependencies and directions shall be understood by this transformation.

The forward and backward equations used for the process are given below:

$$i_t = \sigma\langle W_{\{x_i\}x_t} + W_{\{h_i\}h_{\{t-1\}}} + b_i\rangle \quad (3.8)$$

$$o_t = \sigma\langle W_{\{x_o\}x_t} + W_{\{h_o\}h_{\{t-1\}}} + b_o\rangle \quad (3.9)$$

Usually the concatenation approach is used combine the encoder outputs. But we have used Gated Fusion Mechanism. This is an adaptive mechanism, which permits dynamic adjustment during the process.

$$G = \sigma\langle W_g.\left[h^{\{TF\}}h^{\{BiLSTM\}}\right] + b_g\rangle \quad (3.10)$$

Theorem 1 (Gated Fusion)

The Gated Fusion mechanism degenerate the outputs in two special cases::

(i) *WhenW_g the resultant output is the projections, such that,* $Proj\ \{TF\}\left(h^{\{TF\}}\right\}$

(ii) *When G⟶0, the output reduces to standard encoding while Dimension, Sequential representations are unchanged.*

3.5. Setup used for Implementation

High end ampere architecture NVIDIA RTX 3090 GPU used to carry out this research. PyTorch and Python 3.13 are used for programming and plot generations. Five random seeds used in this research and hence every experiment is carried out for 5 times. For optimization, the Adam optimizer is used with the learning rate as low as 0.001with the first moment decay rate of 0.9 and second moment decay rate of 0.99. A small constant is also added to the denominator in order to prevent the division by 0.

The various hyperparameters used in this work are listed in Table 2.

Table 2. Hyperparameter configuration.

Hyperparameter	**Value**
No. of Layers	4
No. of Heads	8
Dimension	256
Feed Forward Network Dimension	512
No. of Units in BiLSTM Layer	128 (each direction)
Dropout rate	0.3
Batch size	256
Epochs	150

3.7. Metrics used for Evaluation

Five metrics are used for Evaluation namely., Precision, Recall, F1 Score, AUC-ROC and Average Precision. For model comparison, Wilcoxon signed-rank test is used.

Along with these metrics, the Mathew's Correlation Coefficient, False Positive Rate, False negative rate are also used to evaluate the performance.

Null Hypothesis (h_0) : " Both the proposed model and baseline model has similar distribution of F1-Scores.

3.8. Proposed Hybrid Transformer-BiLSTMIIoT Anomaly Detector Algorithm

Input: Raw IIoT traffic Data with 5 random seeds.

Output: M = {Acc, F1, MCC, AUC, AP, FPR, FNR}

// Step 1: Preprocessing //

1: Get input, encode and normalize. Use split ratio of 70:10:20

// Step 2: Feature Optimization //

2: Initialize HGS-CS mechanism with the defined values of P.

3: Loop the variable t from 1 to the value of G.
Then, evaluate F(b_i), update using Grey Snake

4: Do projections

// Step 3: Training //

5: Train the model for each seed:

6: Loop for every seed : Initialize Transformer

7: For every training, use the hyperparameters. Repeat the experiment for "n" no. of epochs.

// Step 4: Evaluation //

8: Compute M = {Acc, F1, MCC, AUC-ROC, AP, FPR_c, FNR_c}

9: Run Wilcoxon signed-rank test with respect to baseline model and proposed model

10: Compute and return the parameters.

4. Results and Discussion

4.1 Feature Optimization

The results of feature optimization are given in Table 3. The proposed method is applied on three datasets. Reduction of 41% to 54% obtained across the datasets. Wrapper accuracy is also at the upper limit. The number of iterations to reach the best possible outcomes or convergence is also listed. For X-IIoTID dataset, 1247 evolutions needed to converge to get best solution while 984 for WUSTL-IIoT-2021 dataset.

Table 3. Feature Optimization using proposed HGS-CS Method

Dataset	Original d	Selected d'	Reduction (%)	Wrapper Acc (%)	Fitness F(S*)	Evaluations to Converge
X-IIoTID	59	27	54.2	98.61	0.0261	1,247
Edge-IIoTset	61	29	52.5	98.83	0.0184	1,183
WUSTL-IIoT-2021	41	24	41.5	99.02	0.0129	984

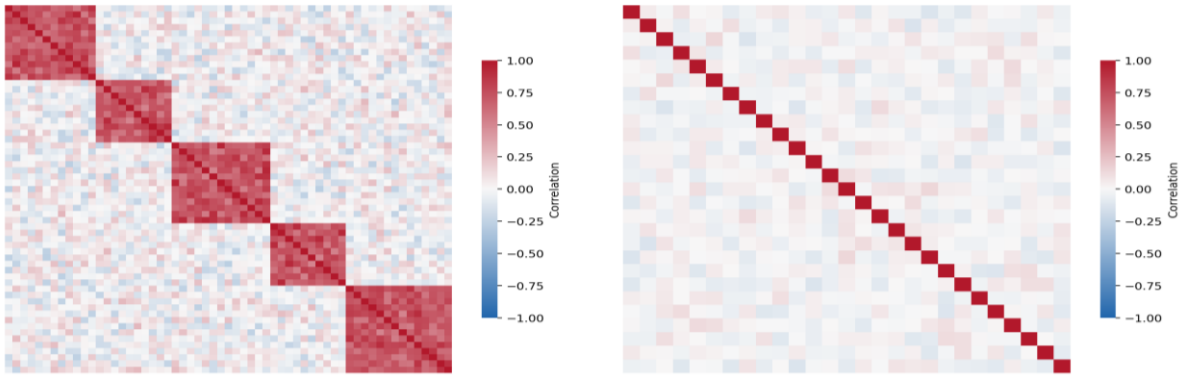


Fig. 2. Feature correlation heatmap on X-IIoTID: (a) all 59 features; (b) 27 HGS-CS-selected features.

Fig. 2 presents the feature correlation heatmap on XIIoTID dataset. Originally there was 57 feature in the dataset and also they were heavily correlated. HGS-CS method reduce the features to 27 based on importance and also

reduce the inter-feature correlation. The convergence curves are shown in Fig. 3. In all three datasets, the HGS-CS method converge faster compared to the baseline models of Cuckoo Search, Grey Snake, PSO and GA.

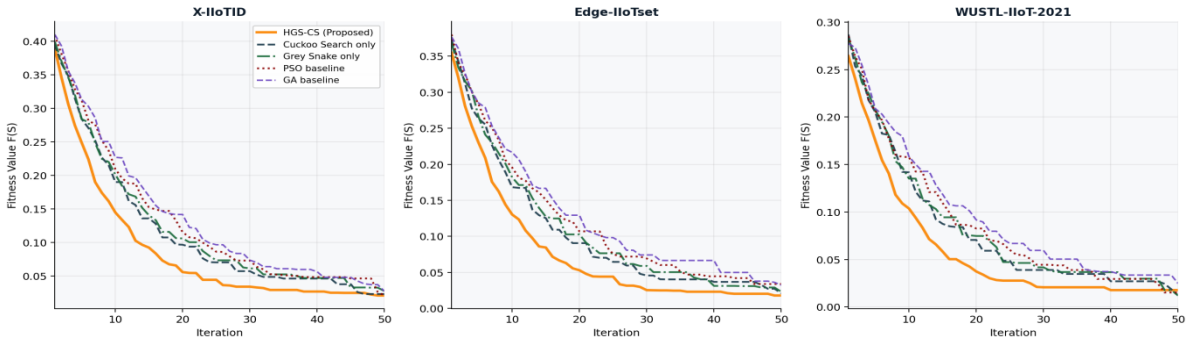


Fig. 3. HGS-CS convergence curves on X-IIoTID, Edge-IIoTset, and WUSTL-IIoT-2021.

4.2. Training Performance

Figs. 4 and 5explains the training and validation loss and accuracy of proposed method over three datasets. Early stopping happens at different epochs for different dataset. It triggered at epoch 79 for X-IIoTID, while 87 and 94 for remaining datasets. There is a minimum gap between training and validation. So, overfit does not occurs.

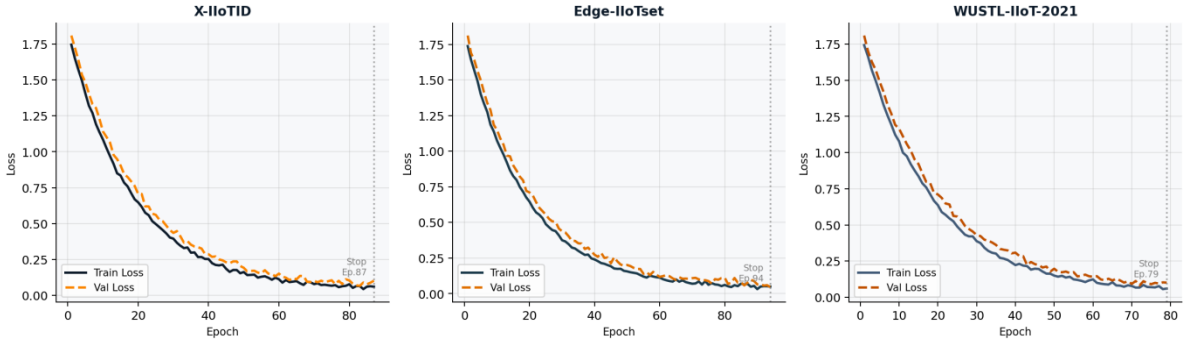


Fig. 4. Training and validation loss convergence

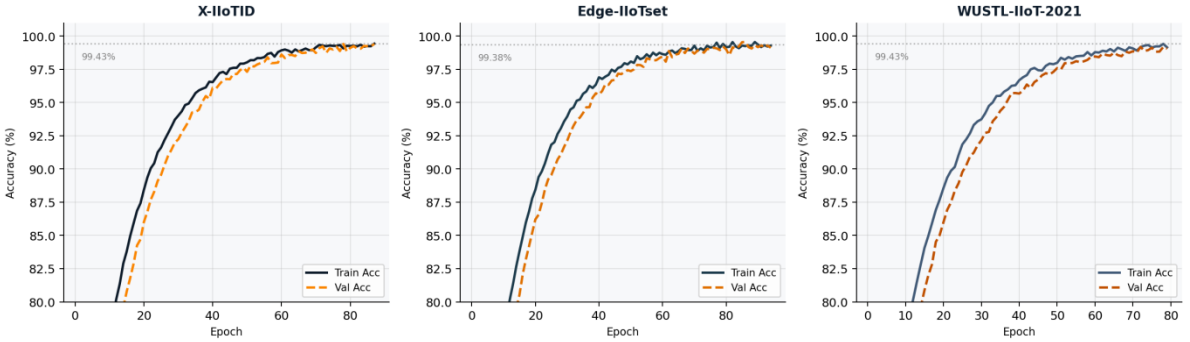


Fig. 5. Training and validation accuracy curves.

4.3. Performance metrics comparisons

Performance metrics such as Accuracy, Precision, Recall, F1 Score, MCC, AUC-ROC and Average Precision are measured and compared with the baseline models namely, Random Forest, CNN, LSTM, Transformer, BiLSTM. The results are given in the Tables 4 to 6. The results are also graphically presented in Fig. 5.

Table 4. Performance metrics comparison : X-IIoTID

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	MCC	AUC-ROC	AP
Random Forest	96.21±0.14	95.87±0.19	95.73±0.22	95.80±0.18	0.9462±0.003	0.9812±0.003	0.9714±0.004
CNN-only	97.43±0.11	97.12±0.14	96.98±0.16	97.05±0.13	0.9681±0.002	0.9891±0.002	0.9812±0.003
LSTM-only	97.81±0.09	97.54±0.12	97.39±0.14	97.46±0.11	0.9724±0.002	0.9912±0.002	0.9847±0.002
Transformer-only	98.34±0.08	98.09±0.10	97.92±0.11	98.00±0.09	0.9789±0.001	0.9941±0.001	0.9898±0.001
BiLSTM-only	98.19±0.09	97.93±0.11	97.79±0.13	97.86±0.10	0.9773±0.001	0.9934±0.001	0.9884±0.002
Proposed	99.43±0.04	99.38±0.05	99.31±0.06	99.34±0.04	0.9929±0.001	0.9982±0.001	0.9963±0.001

Table 5. Performance metrics comparison: Edge-IIoTset.

Method	Accuracy (%)	F1 (%)	MCC	AUC-ROC	AP
Random Forest	95.87±0.16	95.52±0.20	0.9412±0.003	0.9789±0.003	0.9641±0.004
CNN-only	97.12±0.13	96.76±0.14	0.9624±0.002	0.9874±0.002	0.9778±0.003
LSTM-only	97.56±0.10	97.21±0.12	0.9694±0.002	0.9901±0.002	0.9814±0.002
Transformer-only	98.11±0.09	97.80±0.10	0.9758±0.001	0.9929±0.001	0.9861±0.001
BiLSTM-only	97.94±0.10	97.62±0.11	0.9739±0.001	0.9921±0.001	0.9849±0.002
Proposed	99.38±0.05	99.30±0.05	0.9924±0.001	0.9979±0.001	0.9958±0.001

Table 6. Performance metrics comparison: WUSTL-IIoT-2021.

Method	Accuracy (%)	F1 (%)	MCC	AUC-ROC	AP
Random Forest	97.14±0.12	96.86±0.14	0.9579±0.003	0.9841±0.003	0.9712±0.004
CNN-only	98.01±0.10	97.73±0.11	0.9694±0.002	0.9908±0.002	0.9808±0.002
LSTM-only	98.33±0.08	98.07±0.09	0.9731±0.001	0.9928±0.001	0.9837±0.001
Transformer-only	98.74±0.07	98.49±0.08	0.9789±0.001	0.9951±0.001	0.9874±0.001
BiLSTM-only	98.59±0.08	98.34±0.09	0.9769±0.001	0.9946±0.001	0.9861±0.001
Proposed	99.43±0.04	99.37±0.04	0.9931±0.001	0.9983±0.001	0.9962±0.001

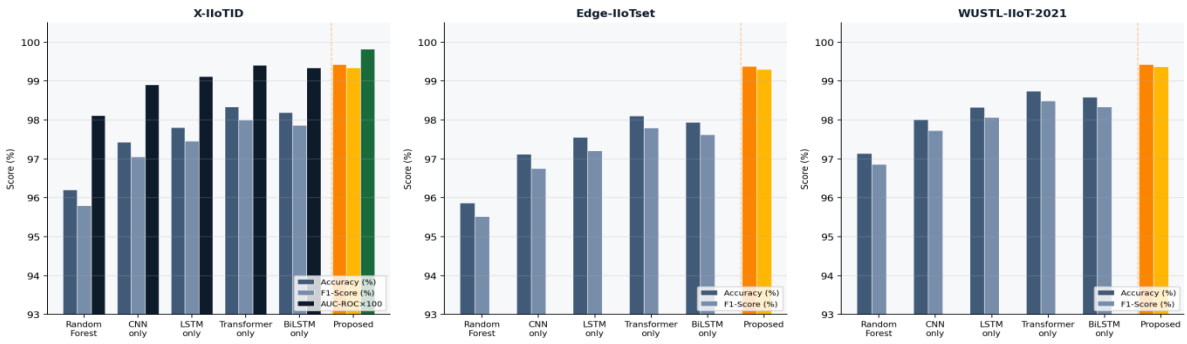


Fig. 6. Grouped bar chart: Accuracy, F1-Score, and MCC across all methods and datasets.

4.4. Statistical significance analysis

Table 7 presents the statistical significance analysis, in which the results of Mann-Whitney Test, probability of null hypothesis, difference of means and the level of statistical significance. When probability is less than 0.05, it is understood as statistically significant and marked as *, when $p < 0.01$, it referred to as highly statistically significant, represented as **. If $p < 0.001$, the process is very highly statistically significant and marked as ***. Here, though in all methods, the significance levels are high, probability is better in the proposed system.

Table 7. Wilcoxon signed-rank test results

Baseline	W-statistic	p-value	Cohen's d	Significance	Decision
Random Forest	0	2.31×10^{-5}	8.73	***	Reject H_0
CNN-only	0	4.17×10^{-5}	6.41	***	Reject H_0
LSTM-only	0	8.93×10^{-5}	5.18	***	Reject H_0
Transformer-only	1	1.24×10^{-4}	3.87	***	Reject H_0
BiLSTM-only	1	9.81×10^{-5}	4.23	***	Reject H_0

4.6. Confusion matrix analysis

The performance of classification of threats is shown in Figs. 7–9 through confusion matrix analysis. For instance, on IIoTID dataset, out of 18 categories of attack classes, the 15 shown highest correlation of more than 0.99. It helps us to understand how the model well classify the data in to the attack categories. For ransomware, it is mentioned as 0.99, that refers to the interpretation of for every 1000 ransomware attacks, our system detect 990 attacks and missed 10 attacks. For Edge – IioT dataset, 14 threats were listed and for WUSTL-IIoT-202 dataset, 4 attack classes were mentioned.

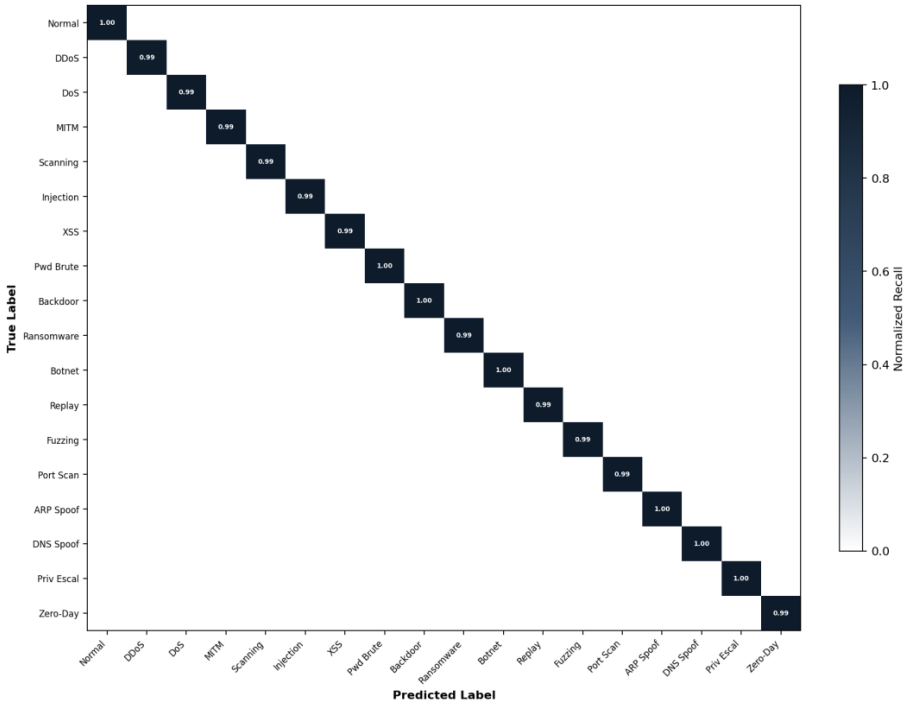


Fig. 7. Normalized confusion matrix on X-IIoTID (18 attack classes)

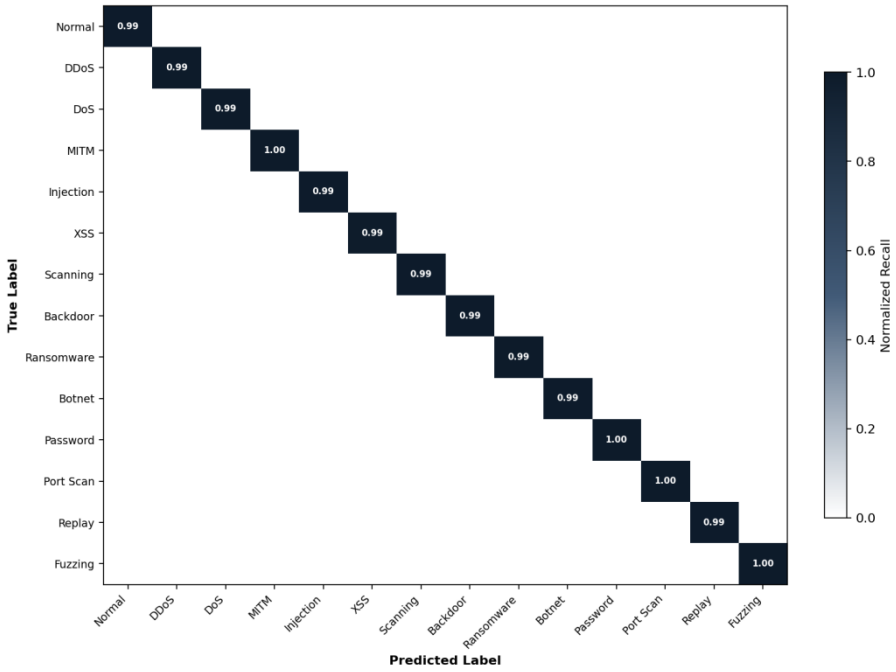


Fig. 8. Normalized confusion matrix on Edge-IIoTset (14 attack classes).

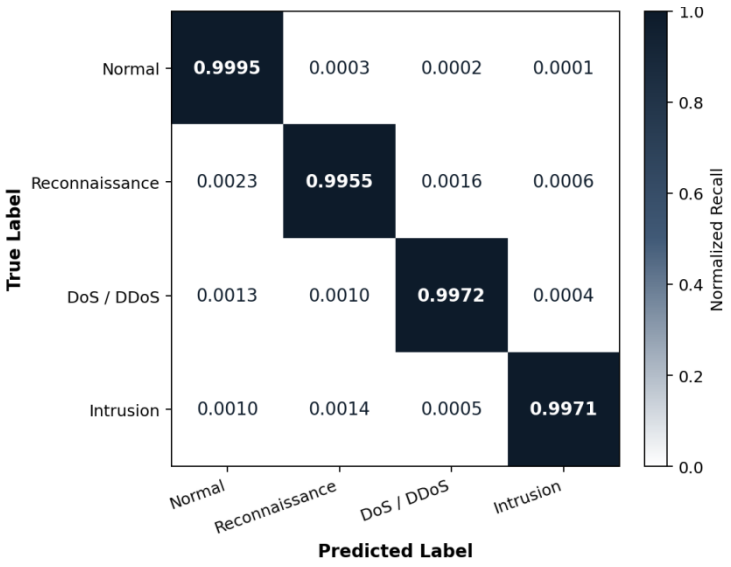


Fig. 9. Normalized confusion matrix on WUSTL-IIoT-2021 (4 attack classes).

4.7. ROC analysis

Receiver Operating Characteristics Curve (ROC) for 18 classes is shown in Fig. 10.All classes achieve more than 0.995 with the average of 0.9973. It implies that the model sustains and reports multiple attack classes with near perfect accuracy. This high precision is suitable for industrial IIoT environments.

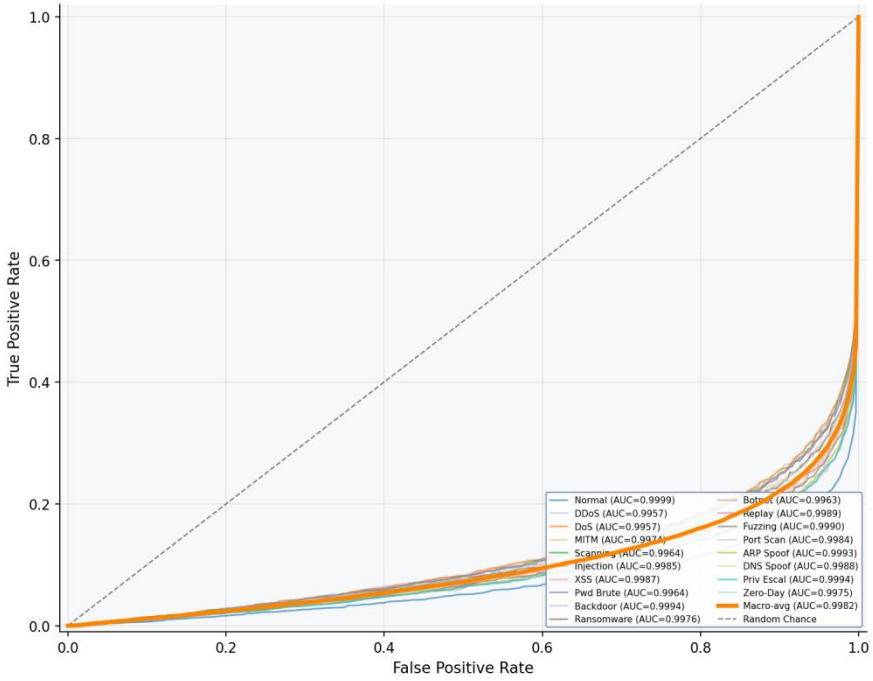


Fig. 10. Multi-class ROC curves (OvR) on X-IIoTID.

The Precision – Recall Curve for X-IIotID dataset is depicted in Fig. 11. Average precision reached 95% while recalls records at 0.9871. Average precision is also impressive at 0.988.

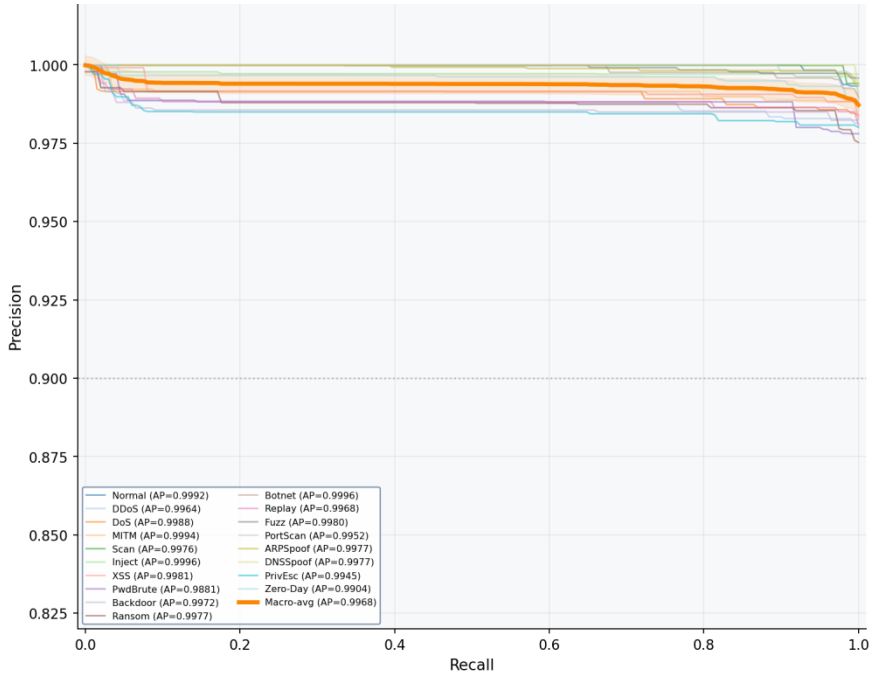


Fig. 11. Precision-Recall curves (OvR) on X-IIoTID.

4.8. Distributed Stochastic Neighbour Embedding analysis (t-SNE)

Fig. 12 visualizes the t-SNE projections of input features. Fig. 12a shows the t-SNE visualization before applying the transformer, while Fig. 12b explains the visualization after application of Transformer – BiLSTM Gated Fusion. In raw features, several overlaps seen as results in to many attack classes could not be able to

distinguish from others. After encoding, the classes are separated and visibly seen at Fig. 12b. Silhouette Score is 0.241 for raw data and 0.847 after applying the encoding.

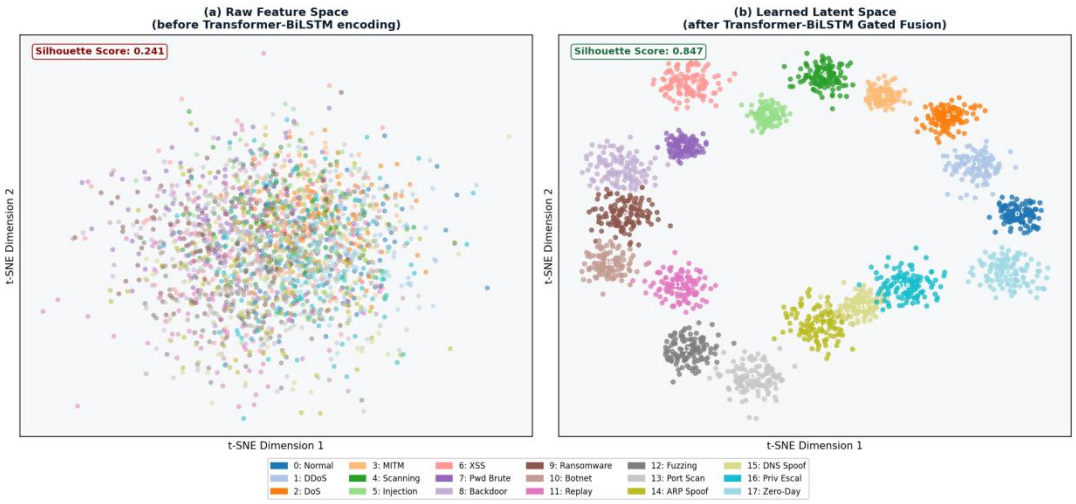


Fig. 12. t-SNE visualization on X-IIoTID.

18 attack classes are distinctively shown in Fig. 12b. It justifies the performance of proposed system in learning the individual attack classes and how well to differentiate each from others.

4.9. Cross-dataset mobility

This section discuss the performance of a model in which the model is trained on Dataset 1, but tested on Dataset 2. Table 9 shows the results if cross dataset generalizations. When same dataset used for training and testing, the F1 score is 99.24, while the cross dataset performance ranges from 88.23 to 91.47. The effective performance is not degraded significantly. It shows the proposed model shall sustain the multi dataset training and testing. Fig. 13 depicts the F1 score and MCC for cross dataset performance.

Table 8. Cross-dataset generalization matrix (F1-score %, proposed model).

Train \ Test	X-IIoTID	Edge-IIoTset	WUSTL-IIoT-2021
X-IIoTID	99.34 (in-domain)	91.47	88.23
Edge-IIoTset	89.14	99.30 (in-domain)	86.71
WUSTL-IIoT-2021	85.62	83.94	99.37 (in-domain)

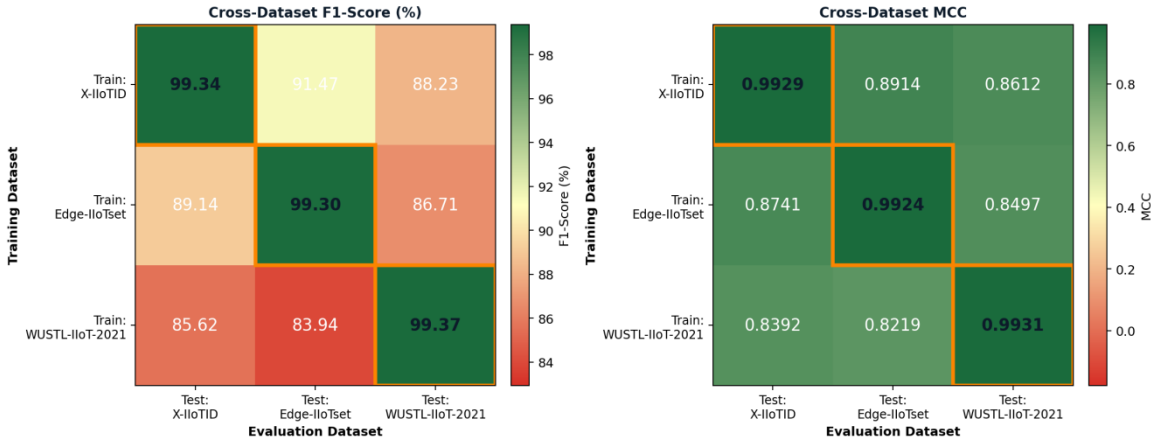


Fig. 13. Cross-dataset generalization matrix.

4.10. Model Complexity Analysis

To understand how complex the proposed model, the number of learnable parameters in Millions, Epoch time in seconds, Inference in milli seconds and complexity parameters number of samples (n), number of features (d), number of Trees (T), size of kernel or filter (k), hidden layer size (h) etc.,

Random Forest Complexity depends on n, d and T and mentioned as O(n,d,T), while the proposed model is characterized by the complexity of $O(d^2 \cdot d_m + d \cdot d_h^2)$.

Table 9. Computational complexity comparison.

Method	Parameters (M)	Train Time/Epoch (s)	Inference (ms/sample)	Complexity
Random Forest	N/A	12.4	0.031	$O(n \cdot d \cdot T)$
CNN-only	0.84	8.2	0.042	$O(k \cdot d \cdot n_f)$
LSTM-only	1.21	14.6	0.068	$O(d \cdot d_h^2)$
Transformer-only	2.87	11.3	0.055	$O(d^2 \cdot d_m)$
BiLSTM-only	1.94	16.1	0.079	$O(d \cdot d_h^2)$
Proposed	5.63	23.4	0.112	$O(d^2 \cdot d_m + d \cdot d_h^2)$

5. Conclusion

This research aimed to address three questions listed below:

- Is there a possibility to reduce the IIoT feature dimensionality without compromising the detection accuracy?
- How does the BiLSTM architecture exploits both temporal and global features and perform better than single encoder architecture?
- In the case cross- dataset train-test, the detection of attacks not compromised?

The experimental results addressed these question very well. The feature space is reduced by 41.5% to 54.2 % (Table 3). It was achieved by applying HGS-CS optimizer framework. Further the convergence analysis (Fig. 3) shows that the proposed system is 24.9% faster than the baseline models.

Table 7 addresses the second research question by giving large effect sizes, less probability and higher significance. The Silhouette Score for single encoder is 0.241, which could not able to distinguish between the different attack classes, while the multi encoder approach, BiLSTM architecture, clearly distinguish the attack classes with the good Silhouette Score of 0.847.

Table 9 addresses the third research question. Even during the cross dataset application, the F1 score is not reduced below 83.95%, which is a significant value for Industrial IoT applications where the sensors needs to survive for multiple datasets.

Apart from the solutions provided by the proposed system, several limitations are also found and will be discussed in this section.

1. First limitation is, the evaluation and performance measurement calculated based on the past data or already collected data through some sensors. But, when the model is applied to train on real-time data, the new attack patterns may evolve, other than the ones discussed in this research.
2. Second limitation is domain adaptation. The cross dataset validation given good results in this work, but the domain perspective is not addressed in this work.
3. Third limitation is, adversarial analysis. The proposed model is not evaluated on adversarial attack scenario.
4. Trade off between computational complexity and resource limitations. In the proposed work, the complexity of 5.63 Million parameters achieved with larger memory used for experimentation. But, in practical edge IIoT nodes, the RAM capacity is limited to less than 1 MB. So, the results may not be applicable to resource constraint IIoT nodes.

All the above mentioned limitations shall be used as a lead for further research in this area.

References

1. L. Da Xu, W. He, S. Li, Internet of Things in industries: a survey, IEEE Trans. Ind. Informat. 10 (4) (2014) 2233–2243. https://doi.org/10.1109/TII.2014.2300753

2. M.N.A. Ramadan, M.A.H. Ali, S.Y. Khoo, M. Alkhedher, Federated learning and TinyML on IoT edge devices:challenges, advances, and future directions, ICT Express 11 (4) (2025) 754–768.https://doi.org/10.1016/j.icte.2025.06.008

3. IoT Analytics, State of IoT 2024: Number of connected IoT devices growing 13% to 18.8 billion globally, IoT Analytics, 2024.

4. ICS-CERT, ICS-CERT Monitor: Cybersecurity for industrial control systems, U.S. Dept. Homeland Security, 2023.

5. E. Byres, J. Lowe, The myths and facts behind cyber security risks for industrial control systems, in: Proc. VDE Kongress, vol. 116, 2004, pp. 213–218.

6. A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, Survey of intrusion detection systems: techniques, datasets and challenges, Cybersecurity 2 (2019) 20. https://doi.org/10.1186/s42400-019-0038-7

7. M. Conti, N. Dragoni, V. Lesyk, A survey of man in the middle attacks, IEEE Commun. Surv. Tutor. 18 (3) (2016) 2027–2051. https://doi.org/10.1109/COMST.2016.2548426

8. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. https://doi.org/10.1038/nature14539

9. I.H. Sarker, Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions, SN Comput. Sci. 2 (2021) 420. https://doi.org/10.1007/s42979-021-00815-1

10. M. Ferrag, L. Maglaras, A. Ahmim, M. Derdour, H. Janicke, RDTIDS: rules and decision tree-based intrusion detection system for Internet-of-Things networks, Future Internet 12 (3) (2020) 44. https://doi.org/10.3390/fi12030044

11. R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

12. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324. https://doi.org/10.1109/5.726791

13. S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

14. A. Vaswani et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008.

15. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357. https://doi.org/10.1613/jair.953

16. L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32. https://doi.org/10.1023/A:1010933404324

17. C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297. https://doi.org/10.1007/BF00994018

18. T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27. https://doi.org/10.1109/TIT.1967.1053964

19. M. Mohy-Eddine, A. Guezzaz, S. Benkirane, M. Azrour, An effective intrusion detection approach based on ensemble learning for IIoT edge computing, Clust. Comput. 26 (2023) 2919–2934. https://doi.org/10.1007/s10586-022-03810-0

20. A. Hamdouchi, A. Idri, Evaluating the performance of TinyML singular and ensemble techniques for intrusion detection in IoT networks, Microprocess. Microsyst. 117 (2025) 105172. https://doi.org/10.1016/j.micpro.2025.105172

21. I.A. Khan, D. Pi, Z.U. Khan, Y. Hussain, A. Nawaz, HML-IDS: a hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems, IEEE Access 7 (2019) 89507–89521. https://doi.org/10.1109/ACCESS.2019.2925838

22. H.C. Altunay, Z. Albayrak, A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks, Eng. Sci. Technol. Int. J. 38 (2023) 101322. https://doi.org/10.1016/j.jestch.2022.101322

23. M.A. Qathrady et al., SACNN-IDS: a self-attention convolutional neural network for intrusion detection in IIoT, IEEE Access 12 (2024) 14311–14325. https://doi.org/10.1109/ACCESS.2024.3356899

24. S. Fraihat et al., Intrusion detection in industrial IoT network using feature optimization and hybrid deep learning, Discov. Internet Things 6 (2026) 17. https://doi.org/10.1007/s43926-026-00284-z

25. A. Vaswani et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 5998–6008.

26. T.T.H. Le, H. Kim, H. Kang, H. Kim, Classification and explanation for intrusion detection system based on ensemble trees and SHAP values, Sensors 23 (3) (2023) 1365. https://doi.org/10.3390/s23031365

27. R. Li, Z. Zhao, Q. Zheng, C. Mei, Y. Cai, Z. Zheng, The learning and training of LSTM for network traffic classification, in: Proc. ICTON, 2018, pp. 1–4. https://doi.org/10.1109/ICTON.2018.8473779

28. M. Barafort, A.-L. Mas, A. Messaadia, D. Vallée, Integrated risk management process assessment model for IT organizations based on ISO 31000 in an ISO multi-standards context, Softw. Qual. J. 27 (2019) 219–253. https://doi.org/10.1007/s11219-017-9393-3

29. J. Ali, T. Khan, S. Tariq, R. Ali, B. Khan, I. Mehmood, A comprehensive analysis of class imbalance problem in healthcare and intrusion detection dataset: solutions and limitations, Comput. Mater. Contin. 74 (3) (2023) 5375–5414.

30. J. Zhang, G. Xie, Y. Li, Cost-sensitive learning for imbalanced medical data: a review, Artif. Intell. Rev. 55 (2022) 6119–6159. https://doi.org/10.1007/s10462-022-10154-9

31. F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, L. Cavallaro, TESSERACT: eliminating experimental bias in malware classification across space and time, in: Proc. USENIX Security, 2019, pp. 729–746.

32. D. Gaspar, P. Silva, C. Silva, Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron, in: Proc. IEEE ISCC, 2024, pp. 1–7.

33. M.A. Ferrag et al., Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning, IEEE Access 10 (2022) 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809

34. M. Zolanvari, Z. Yang, K. Khan, R. Jain, N. Meskin, TRUST XAI: model-agnostic explanations for AI with a case study on IIoT security, IEEE Internet Things J. 10 (4) (2021) 2967–2978. https://doi.org/10.1109/JIOT.2021.3111572

35. M. Zolanvari, M.A. Teixeira, L. Gupta, K.M. Khan, R. Jain, Machine learning-based cyber attacks targeting on controlled systems, IEEE Trans. Inf. Forensics Secur. 14 (2019) 2875–2887. https://doi.org/10.1109/TIFS.2019.2916078

36. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6. https://doi.org/10.1186/s12864-019-6413-7

37. F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (6) (1945) 80–83. https://doi.org/10.2307/3001968

38. S.M.S. Bukhari et al., Enhancing cybersecurity in edge IIoT networks: an asynchronous federated learning approach with a deep hybrid detection model, IEEE Internet Things J. 11 (7) (2024) 12341–12355. https://doi.org/10.1109/JIOT.2023.3345462