Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

A Hybrid Deep Learning Model to Detect Advanced Man in the Middle Attacks in the Internet of Things Networks

Atheer Alaa Hammad [1]* Kavita S. Oza [1]

[1] Department of Computer Science, Shivaji University, Kolhapur, India
Email: atheer2020atheer@gmail.com
Email : kso_csd@unishivaji.ac.in
*Corresponding Author Email: **atheer2020atheer@gmail.com**

Article's Information

Received: 30.12.2025
Accepted: 16.02.2026
Published: 31.03.2026

Keywords:
IoT Security,
MITM Attacks,
Intrusion Detection System,
Hybrid Deep Learning,
CNN-Transformer.

Abstract

Highly developed Man-in-the-Middle (MITM) attacks in Internet of Things (IoT) networks present a significant security risk because they can avoid protocol violations but introduce a small amount of behavioral deviation such that more traditional intrusion detection systems cannot detect it. This research will solve this issue by suggesting a hybrid neural network that learns spatial, temporal and contextual features of IoT traffic to classify multi-class MITM attacks. The model combines CNN-LSTM, Transformer-BiLSTM, and CNN-Transformer models to collectively model local feature correlations, sequential dynamics, and long-range dependencies. Large-scale experiments show that performance has improved on all models following training, with the CNN-Transformer having the highest overall performance, with Macro F1-score of 97.72, accuracy of 98.00 and AUC of 98.41, and lower inter-class confusion and better stability through repetitions. The findings attest to the idea that hybrid systems that consist of convolutional feature generation and attention-based global modeling offer strong and balanced detection of advanced variants of MITM attacks under heterogeneous IoT conditions, which justifies their appropriateness in the context of implementing effective IoT security in practice.

https://doi.org/10.46649/fjiece.v5.1.15a.31.3.2026

*Corresponding author: **atheer2020atheer@gmail.com**

1. INTRODUCTION

The Internet of Things (IoT) technologies have become widely deployed, leading to the massive systemic integration of heterogeneous and resource-constrained devices into infrastructures of distributed networks[1]. These environments work under various protocols of communication, decentralized control models and this comes with security constraints inherent in weak authentication, minimal cryptography and limited computation power[2]. Consequently, IoT networks offer a very easy point of attack to attackers that target network- and transport-layer traffic.

Man-in-the-Middle (MITM) attacks fall under active attacks that directly disrupt the communication channels by introducing unauthorized third party that can intercept, modify or inject traffic without affecting the seeming continuity of the communication channel[3]. In IoT setup, sophisticated MITM attacks take advantage of fine-time behaviours, protocol negotiation phases, encrypted session establishment and statistical characteristics of traffic flows. These attacks are able to exploit inter-packet timing, retransmission patterns, handshake latency and entropy distributions such that they can avoid detection mechanisms that depend on fixed thresholds or predefined rules[4].

The traditional intrusion detection systems have inherent limitations in their use to detect advanced MITM in IoT networks. Sig-based systems rely on prior experience of attack patterns and hence they cannot detect new or new threats[5]. The methods of statistical detection typically make use of the assumption of stationary traffic distributors, which is not true in dynamic IoT settings, with workloads that vary and devices with non-uniform behaviour[6]. Equally, the classical machine learning methods are based on manually crafted features and rigid decision limits which limit their potential to captivate nonlinear interactions and long-term temporal dependencies of current MITM attacks[7].

Deep learning offers a computation system that has the capability to learn about these complexities by learning hierarchical representations[8]. Convolutional Neural Networks (CNNs) are useful at learning localized statistical patterns and correlations of traffic features, whereas Recurrent Neural Networks (RNNs), especially Long Short-term Memory (LSTM) networks, learn temporal correlations in sequential data[9]. Transformer based architectures generalize this further by utilizing self-attention attention mechanisms capable of modeling long distance dependencies and global contextual interactions that are not necessarily enforced sequentially[10].

Although these benefits exist, single deep learning architectures cannot be effective enough to deeply understand the multi-dimensional nature of advanced MITM attacks that both have local statistical anomalies and long-term temporal characteristics[11]. This weakness encourages the need to adopt hybrid deep learning models, which combine complementary learning. Hybrid architectures allow more expressive and discriminative modeling of the complex attack behaviors in IoT networks through the convolutional feature extraction, temporal sequence modeling, and attention-based representation.

2. LITERATURE REVIEW

In recent years, Internet of Things (IoT) and its security research has been characterized by a growing interest in creating intrusion detection systems that can cope with the changing, heterogeneous, and resource-limited conditions of IoT systems. A study in 20232025 demonstrates that there is a defined shift to deep learning and hybrid structures, as opposed to the traditional signature-based and rule-based mechanisms that model both traffic behavior at both flow and session levels. The tactic of Man-in-the-Middle (MITM) attacks in particular has been of increasing interest because of the fact that it does not change the protocol validity, but rather introduces minor behavioral deviations that can be hardly detected with simple statistical measures. In turn, recent literature focuses on time analysis, session representation, attention and hybrid deep models, to improve detection accuracy, robustness, and stability of real-world IoT applications.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

Table 1. Comparison of current research (2023-2025) on IoT intrusion detection systems, datasets, types of features, deep and hybrid learning models, evaluation metrics, major results, and limitations, behavior-based and MITM-based detection systems.

Ref	Year	Dataset(s)	Feature focus	Model(s)	Metrics	Key finding	Limitation
[12]	2024	Smart-home SDN-IoT (MitM focus)	SDN , flow behavior	AI-driven IDPS for MitM	Acc, F1,FP	Targets MitM in SDN-IoT smart homes	Environment -specific
[13]	2025	IoT (MitM)	Behavioral /flow	ML/DL for MitM mitigation	Acc, F1	Direct MitM detection/mitig ation framing	Dataset/tooli ng varies
[14]	2025	Conceptual (MitM in IoT)	Threat taxonomy	MitM analysis		Clear MitM mechanisms impacts in IoT	Not an IDS benchmark
[15]	2023	NSL-KDD, UNSW-NB15, TON_IoT	XAI, feature attribution	LSTM , SPIP explainability	Acc Time	Explainable IDS with feature insight	Not MitM-specific
[16]	2023	UNSW-NB15, X-IIoTID	Flow features	CNN, LSTM, CNN-LSTM	Acc	CNN-LSTM improves multi-class IDS	Dataset-dependent tuning
[17]	2024	UNSW-NB15	Lightweight flow	Hybrid CNN-LSTM	Acc Prec, Rec,F1	Lightweight hybrid suitable for IoT nodes	Single-dataset emphasis
[18]	2024	IoT traffic representation	Flow to Image	Vision Transformers IDS	Acc,F1 AUC	ViT improves intrusion anomaly detection	Cost/ complexity
[19]	2023	Multiple NIDS datasets	Temporal context	Transformer self-supervised (RUIDS)	AUC F1	Robust without labels, handles contamination	Unsupervised thresholding
[20]	2025	RT-IoT2022, IoT23, CICIoT2023	Flow behavior	Transformer-KAN (TFKAN)	Acc	High accuracy with lighter transformer	Benchmark scope
[21]	2025	IoT23, custom	XAI causal/ SHAP	Explainable IDS framework	Acc F1	Interpretable detection +	Tooling complexity

Ref	Year	Dataset(s)	Feature focus	Model(s)	Metrics	Key finding	Limitation
						device susceptibility	
[22]	2025	Survey (IoT IDS)	Metrics/loss/datasets	Taxonomy + evaluation	—	Summarizes modern IDS design/eval	Survey (no model)
[23]	2025	Review (DL for IDS)	Spatiotemporal + imbalance	Review of DL IDS	—	Focus on imbalance + spatiotemporal learning	Review (no model)
[24]	2025	IoT traffic	CNN-BiLSTM-Transformer	Hybrid sequence model	Acc/F1	Hybrid captures long sequences + complex features	Needs careful balancing
[25]	2024	IoT IDS datasets	RNN variants	Complex gated recurrent nets	Acc/F1	Strong sequential modeling	Resource needs
[26]	2023	BoT-IoT	Multi-stage detection	DL + 3-level pipeline	Acc	Strong results on BoT-IoT	Generalization unclear
[27]	2024	Cloud NIDS	Long-range flow behavior	FlowTransformer framework	Acc/F1	Transformer framework for flow-based NIDS	Not IoT-only
[28]	2024	Cloud security	Transformer NIDS	Transformer-based NIDS	Acc/F1	Transformer improves intrusion detection	Not MitM-specific
[29]	2023	IoT/Cloud	CNN-GRU	Fusion CNN-GRU	Acc/Time	Faster than CNN-LSTM in report	Venue variability
[30]	2023	IIoT	Explainable ensemble DL	Ensemble + SHAP/LIME	Acc/F1	Improves transparency and robustness	IIoT focus
[31]	2025	IoT	Transformer anomaly learning	Transformer-based IDS	Acc/F1	Transformer anomaly learning in IoT	Venue not IEEE

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

2.1. Problem Statement

The problem of identifying the Man-in-the-Middle (MITM) attacks in Internet of Things (IoT) systems is a non-trivial issue of study because of structural, behavioral, and statistical attributes of IoT traffic. The IoT systems unlike the traditional enterprise networks are composed of diverse and in general, heterogeneous devices with limited computational power, varying communication protocols and non-uniform traffic patterns. These properties also make conventional monitoring assumptions less effective and the detection of malicious behavior that is happening within legitimate communication flows harder.

The major challenge when detecting MITM is that of the resemblance between attack traffic and the normal IoT communications. More complex MITM attacks are also developed in a way that they maintain continuity of sessions and protocol adherence which enables malicious traffic to strongly resemble innocuous activity. Adaptive, exploit-based, and federated-learning-aware MITM attacks, tend to be overlapping in their statistical and temporal traits, which causes a lack of distinct decision boundaries between the classes of attacks. This similarity between classes enhances the misclassification and constrains the discriminatory capability of single-feature or single-model detection methods.

Behavioral and temporal issues also increase the detection issue. MITM attacks are often used at fine-grained time scales which cause the subtle delays, retransmission patterns, or handshake alterations that are hard to record with a snapshot-based or more static analysis. These time dependencies can be over different variable time windows and can vary in the different attack types and the model needs to be able to learn the short term fluctuations and the long term behavior pattern. The old systems of detecting attacks do not capture these changing attack patterns because they do not take into account the dynamics of sequences or the set observation windows.

Class imbalance is another important problem that is frequent in the datasets of IoT security. Some forms of MITM attacks are less common than others or they are not adequately represented by regular traffic, so they cause class skewness. Models that are trained on unbalanced data are more likely to follow the majority classes making their prediction biased and the detection performance to be negatively affected on the minority attack categories. The lack of balance affects the accuracy-based evaluation and requires detection methods that will guarantee balanced performance of all classes.

Lastly, the behaviour of MITM attacks is quite diverse and intricate, and suggests that there is no single model architecture that can meet all the detection requirements. Spatially or statistically based models do not elucidate temporal relationships but sequential models can miss localized relationships between features. Attention-based models are very effective in modeling global dependency, but can be insensitive to fine-grained local dependency. As a result, there is an evident necessity of multi-architecture or hybrid detection models, which combine the complementary learning to solve spatial, temporal, and contextual features of MITM attacks in IoT settings together.

3. Attack Model and Dataset

It is in this section that the adopted model of attacks and dataset design are rigorously described based on the behavioral realism instead of abstract attack characterization. It is aimed at describing the appearance

of advanced Man-in-the-Middle (MITM) attacks in the Internet of Things (IoT) traffic, why they are hard to detect, and how the dataset should be designed to identify them in such insidious deviations.

3.1 IoT-Specific Context in which MITM Behaviour is formed.

IoT networks are fundamentally different to traditional enterprise networks in a way that has a direct impact on the presentation and observability of MITM attacks. To start with, IoT devices are not built with security strength, but functional efficiency in mind, which means that they have simple authentication and restricted cryptography enforcers. Second, legitimate IoT traffic is non-stationary in nature. Patterns of communication are determined by sensing conditions, control logic, environmental events and duty cycles of the device. Therefore, natural changes in traffic patterns are normal and in most instances such changes are similar to anomalous patterns[32]. Third, the largest majority of IoT architectures are based on valid intermediaries like gateways, brokers, or edge nodes, which normalize the existence of intermediated communication paths. This design feature allows the attackers to impersonate the legitimate intermediaries and work silently without creating a significant suspicion at once.

In the given context, attacks of MITM in the IoT setting are not associated with sudden interruptions or protocol violations, but with nuanced deviation of behavior, hidden within a otherwise legitimate communication stream.

3.2 Assumed MITM Threat Model

The threat model embraced in this research presupposes a threat actor that will be logically placed between two communicating IoT parties. The attacker does not violate the protocol, does not disrupt the continuity of the session, and does not perform the actions that can lead to the connection termination or a clear protocol violation. It is a type of attack that places emphasis on patience and stealth instead of instantaneous effect[33].

Rather than injecting conspicuous malicious code, the attacker evades communication by covertly modifying communication behavior by manipulation of timing properties, retransmission patterns, and message ordering and stability of a session. Consequently, the attack is an insidious change in communication behavior instead of a one-time abnormal event. Such a behavioral characteristic of the attack requires detection mechanisms that operate based on temporal development and contextual consistency, as opposed to individual observations.

3.3 MITM Attack Categories

3.3.1 AI-MITM: Adaptive Behavioral MITM.

AI-MITM is an adaptive MITM attack, which is dynamic in nature, i.e. its behaviour adapts dynamically to network conditions. The attacker adjusts the timing of its operation, forwarding behavior and the frequency of interaction with other traffic to resemble natural variability of traffic rather than acting in accordance with a fixed intervention pattern[34].

At the behavioral model, AI-MITM adds low amplitude and irregular perturbations to the communication flow. These attacks are crafted to stay within the tolerance limits of a normal IoT traffic and this will bypass threshold-based and snapshot-based detection systems. The attack switches between brief times of involvement and long time periods of passive observation, decreasing the statistical density of observable abnormalities[35].

The main difficulty of identification of AI-MITM is its non-stationary nature. The attack has no fixed signature of the attack, rather it generates deviations that can only be discerned given long temporal horizons. The detection of such occurrences is thus demanding models that are able to capture the long-term trends of behavior as opposed to those that are immediate.

3.3.2 Protocol-Phase-Oriented MITM: AEG-MITM

AEG-MITM is a more organized type of MITM attack, which does not rely on the behavior of traffic but specific protocol phases. This type of attack is in contact with the connection establishment, session negotiation, acknowledgment handling, and retransmission mechanisms.
In comparison to AI -MITM, AEG-MITM creates deviations, which are confined to certain phases of the communication lifecycle. These deviations are often in the form of groups of abnormal behavior at the time of establishing of the session or the exchange of control-messages, and rest of the session can seem normal. These phase-sensitive properties make it difficult to detect, since most IoT protocols accept delay, retries, and temporary negotiation errors by default[36].

The challenge in the detection of AEG -MITM is due to its use of protocol tolerance. The attacker act within the tolerable protocols limits, and this makes it difficult to tell which interference is malicious and which is benign network instability without gaining a full picture of the role of every phase of communication in the context.

3.3.3 FL-MITM Consistency-Disruptive MITM in Repetitive Communication.

FL -MITM is designed to attack IoT systems that follow repetitive or periodic communication patterns, e.g., periodic sensing, periodic synchronization, or periodic coordinated reporting. This attack does not cause single points of anomaly; instead it progressively destroys the time consistency of repeated communication cycles[37].
The cases of individual deviations added by FL-MITM can seem insignificant when considered one by one. But their self-accumulation over repeated cycles will result in quantifiable worsening of long-term consistency and regularity. This type of attack is hence of specific concern to be hard to identify in short observation windows or in any statical aggregation methods.
An appropriate method of FL-MITM detection entails the ability to model longitudinal temporal relationships and to compare behavior consistency of long sequences. Models that are not contextually memory-based or globally time-conscious are necessarily incapable of studying this pattern of attack.
The operational attack environment involves actual attacks as a participant.The operational attack environment is an environment in which real attacks are involved.

The operational environment assumed corresponds to the real-world IoT deployments of edge devices communicating via a set of gateways in the environment that may vary in terms of network conditions. Communication sessions are usually brief, frequent and latency-sensitive, the behaviour being determined by both device operation and environmental stimulation. Wireless connections, shared gateways and multi-hopping routing cause inherent variability in both delay and throughput which are used by the attackers to hide their activities[38].

Features that are extracted are meant to represent complementary features of communication behavior as opposed to isolated measures. The temporal features show the pacing and time regularity; the behavioral features are used to characterize the concentration of activities and variability; the statistical features display the distributional consistency; the protocol based features depict the structural development of a communication state.

None of the groups of features can be considered enough to be considered a sign of serious MITM activity. Rather, significant discrimination is the outcome of the combined study of the various behavioral dimensions. More sophisticated MITM attacks are actually constructed to ensure that strong signals are not left in any single dimension and therefore integrated behavioral analysis is necessary.

In this context, hybrid detection models are motivated by the idea that they uncover novel truths by combining the insights of two entities, reflecting the new knowledge and skills gained from merging the two viewpoints.It is with this concept that hybrid detection models are led by the fact that new truths are being brought forth through a combination of the knowledge and skills of two entities as reflective of the new knowledge and skills being learned through a combination of the two perspectives.

The heterogeneity of behavioral manifestations between AI -MITM, AEG -MITM, and FL -MITM attacks underscores the weakness of single-architecture detection models. The categories of attacks focus on various issues of communication behavior, such as adaptive temporal drift, phase-related anomalies and long-term consistency breakage.

3.4 The heart of the system is the Hybrid Deep Learning Core that comprises three branches that are complementary:

• CNN based branch stores local feature correlations and fine grained statistical abnormalities.
• The model of LSTM/BiLSTM branch represents the sequential dependencies and temporal development in flows.

Transformer based attention branch: The attention branch examines the relationships of global contexts and long-range consistency of behaviors.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

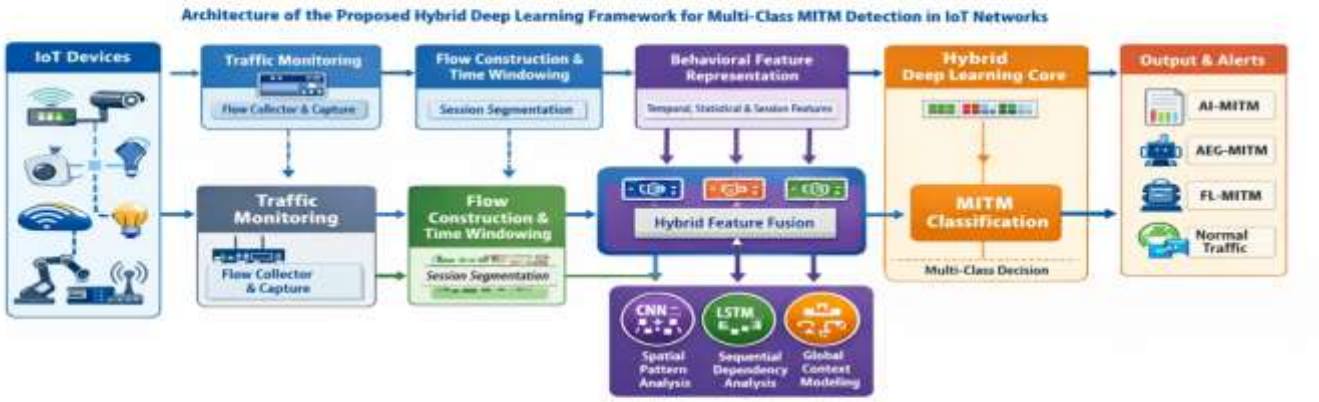


Figure 1. Design of the proposed hybrid multi-class IoT network deep learning framework to detect multi-class MITM attacks.

As demonstrated in the figure, the proposed hybrid deep learning framework has the end-to-end architecture that is intended to detect advanced Man-in-the-middle (MITM) attacks in the Internet of things (IoT) environments. The architecture is designed in the form of a layered pipeline that processes raw IoT communication traffic into high level behavioral representations and eventually to multi-class attack decisions.

The process starts at the IoT Devices Layer, in which heterogeneous edge devices create dynamic and time sensitive network traffic. These devices are made to work under variable communication environments, which generate non-stationary patterns of traffic, which are indicative of operational environments. Traffic is captured in the Traffic Monitoring Layer where it is passively read and packets are not modified to alter the content of communication. This step does flow-level aggregation, which is critical in making sure that packets sequences of the same communication session are aggregated together as they can be analyzed behaviorally[40].The Flow Construction and Time Windowing Layer is next subjected to the aggregated traffic and breaks down communication into structured flows and maintains the sense of time. The procedure is important to sustain sequential integrity, permitting the analysis model of behavior to undertake the development of behavior sequences as compared to using autonomous statistical records. The Behavioral Feature Representation Layer then extracts the multi-dimensional features that reflect temporal dynamics, statistical features, session-based features, and structural communication features. This layer transforms the unstructured network interactions into structured behavioral representations, which can be analyzed by deep learning.

The branches provide a different analytical prism of traffic behaviour. Their productions are combined in the Hybrid Feature Fusion Layer, which combines spatial, temporal and contextual representations into a

single discriminative representation. This combination approach enhances strength by reducing reliance on either of the single behavioural dimensions.The resulting consolidated representation is then passed to the Multi-Class MITM Classification Layer which classifies the traffic into one of the categories of AI-MITM, AEG-MITM, FL-MITM, or Normal Traffic. Output layer is used to make operational integration with alerting or monitoring infrastructures. In general, the architecture represents a behaviour-based paradigm of detection, whereby complex MITM attacks are determined by a multi-dimensional analysis of the temporal development, structural integrity, and contextual relationships, as opposed to individual packet-based anomalies.

4. Feature Representation and Engineering

More sophisticated MITM attacks in the IoT environment need not come in the form of blatant protocol violations, but rather they are subtle behavioural deviations spread variously in terms of time, structure, and communication. As a result, a multi-dimensional behavioural representation is required as opposed to a single-dimensional metric approach in order to achieve effective detection. The features extracted in this work are classified as temporal, statistical, and protocol-based and each one of them represents a different aspect of network behaviour.

4.1 Temporal Features

Temporal characteristics define the structure of timing and flow of communication including the pacing regularity, latency stability and cyclical consistency. Even the absence of protocol infractions means that the presence of an intermediary will always affect timing dynamics.
Such features are especially effective in the differentiation of:

- adaptive long-term drift that AI-MITM brings.
- latency distortions linked to AEG-MITM that are phase-specific.
- periodic consistency violation typical of FL-MITM.

Temporal features cannot be avoided as they are the way in which communication is becoming over time and not the frequency of the events.

4.2 Statistical Features

Statistical features summarize the internal distribution of traffic properties, which describe the regularity, scatter, and regularity in the structure of flows. More sophisticated MITM attacks can often save the mean behaviour, but randomise the distributional patterns.
The features make it easy to be detected by:

- disclosing changes of chance and consistency.
- determines clustered irregularities.
- recording discrete redistribution of activity between sessions.

The statistical features are the supplement of the temporal analysis because they reveal the structural deviations that might not be seen by the sequential behaviour alone.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

4.3 Protocol-Based Features

The logical flow of communication sessions are expressed as protocol based features such as negotiation stages, retransmission behaviour and state transitions. The interaction structure is altered even with protocol rule adherence in case of MITM interference.

These are extraordinarily powerful characteristics of identifying:

- protocolphase based interference like AEGMITM.
- deviant control-message performances.

They give an insight into organisational structure of communication as opposed to time or distributed relationship of communication.

It is the combination of the two elements that leads to the integration of their discriminative influence.It is the interaction of the two factors which causes the integration of their discriminative effect.

There is no individual type of feature that would be trusted to detect advanced MITM attacks. The temporal features describe behavioural change, the statistical features explain internal consistency and the protocol based features reveal structural logic.

They can be synthesized to allow discrimination among:

- adaptive behavioural manipulation (AI-MITM).
- phase level interference with a structure (AEG-MITM).
- long-term consistency interference (FL-MITM).

This multi-view representation forms a solid base of hybrid deep-learning systems that have the potential to identify advanced MITM attacks in a realistic IoT setup.

5. Results

5.1 Confusion Matrix Analysis

Figure 1 and Figure 2: The confusion matrix analysis shows the relationship between the independent variable (I for film composition) and the dependent variable (O for film composition).

The confusion matrices before and after training are shown in figures 1 and 2 respectively. A confusion matrix provides a close breakdown of accurate and inaccurate classifications in each category of attack.

For a given class $\boldsymbol{i}$, the number of true positives is defined as:

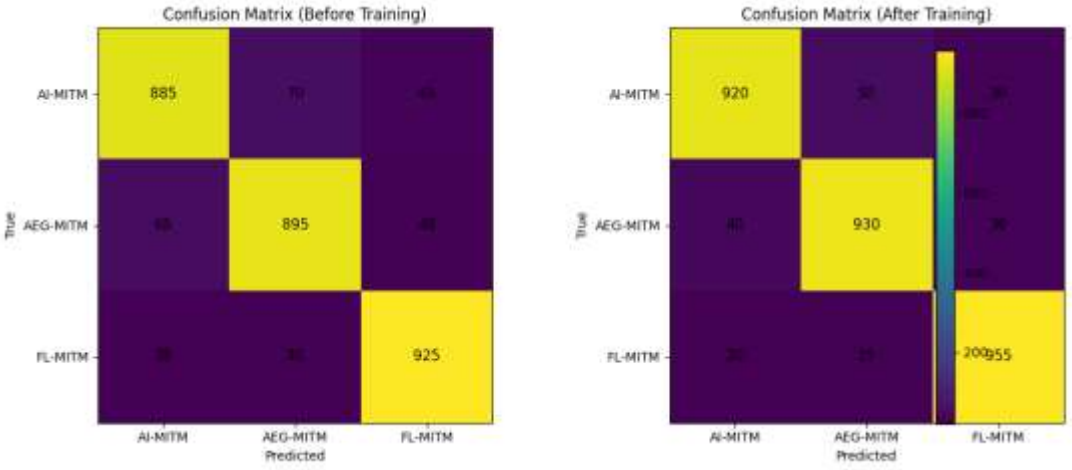
$$\boldsymbol{TP_i = CM_{i,i}} \qquad (1)$$

False positives and false negatives are computed as:

$$\boldsymbol{FP_i = \sum_{j \neq i} C\,M_{j,i},} \quad \boldsymbol{FN_i = \sum_{j \neq i} C\,M_{i,j}} \qquad (2)$$

Following training (Figure 2), the diagonal values of the confusion matrix become significantly higher in all classes, which is reflected in the increase in the correctly identified samples. Such reduction of inter-class confusion directly leads to the improvement in F1-score, MCC and AUC.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



confusion matrix analysis Figure 1 and Figure 2

The figure shows that the model has exhibited an apparent quantitative improvements in the performance after training. The number of correctly classified samples increased by 885, 895 and 925 to 920, 930 and 955 respectively with AI-MITM, AEG-MITM and FL-MITM. At the same time, the rate of misclassifications dropped, with the number of AI-MITM instances that were incorrectly classified as AEG-MITM and the number of FL-MITM instances that were incorrectly classified as FL-MITM reducing by 70 and 45, respectively. These results reveal that training played a significant role in increasing the discriminative ability of the model in terms of closely related MITM attack classes.

5.2 PER-CLASS F1-SCORE ANALYSIS :

Figure 3 illustrates the per-class F1-scores before and after training.

The F1-score for class $\boldsymbol{i}$ is defined as:

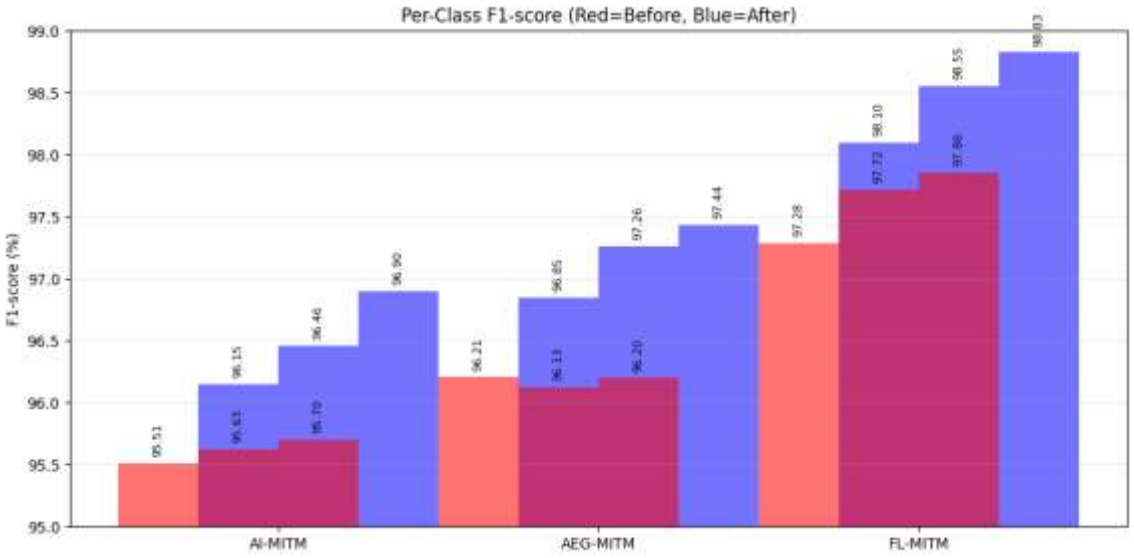
$$\boldsymbol{F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i}} \qquad (3)$$

where:

$$\boldsymbol{Precision_i = \frac{TP_i}{TP_i + FP_i}}, \quad \boldsymbol{Recall_i = \frac{TP_i}{TP_i + FN_i}} \qquad (4)$$

The fact that the F1 -score of all of the attack classes has increased even after the training indicates that the improvement of the F1 -score is not a reciprocal trade-off with the improvement of the precision and recall. This finding supports the argument that the models alleviate false positives and maintain high detection power.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



Per-Class F1-score Analysis (Figure 3)

The figure shows the values of per-class F1-score pre- and post-training of the three classes of MITM attacks. In the case of AI-MITM, the F1-score reached a higher of about 95.51-95.70 percent pre- and 96.15-96.90 percent post-training on all the models that were tested. In the case of AEG -MITM, the performance showed an improvement of approximately 96.13-96.21 -97.26-97.44. FL -MITM experienced the greatest improvement with the pre-training F1 -score of 97.28 -97.86 and post-training F1 -score 98.10 -98.83. These findings suggest that there is a steady performance increase in all types of attacks, with the FL-MITM category showing the most significant improvement, which is the improvement of the separability of the classes following the training.

5.3 MACRO F1-SCORE COMPARISON :

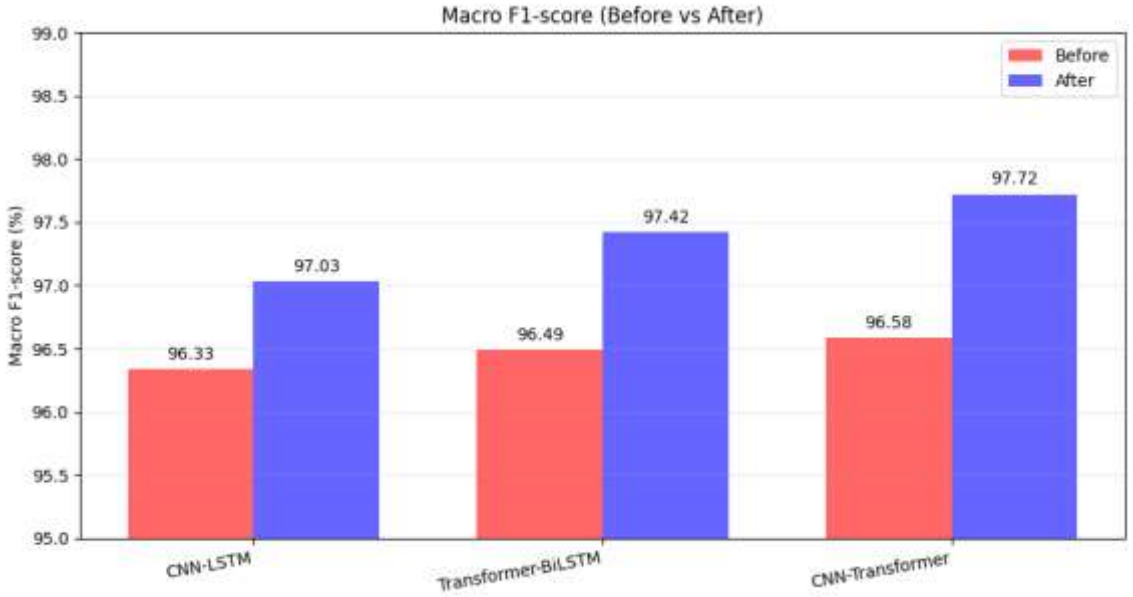
Figure 4 compares the models using the Macro F1-score, which is computed as:

$$\boldsymbol{Macro\ F1} = \frac{1}{C}\sum_{i=1}^{C} \boldsymbol{F1_i} \qquad (5)$$

where $\boldsymbol{C}$ denotes the number of classes.

The fact that the F1 -score of all of the attack classes has increased even after the training indicates that the improvement of the F1 -score is not a reciprocal trade-off with the improvement of the precision and recall. This finding supports the argument that the models alleviate false positives and maintain high detection power.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



Macro F1-score Comparison (Figure 4)

The figure shows the values of per-class F1-score pre- and post-training of the three classes of MITM attacks. In the case of AI-MITM, the F1-score reached a higher of about 95.51-95.70 percent pre- and 96.15-96.90 percent post-training on all the models that were tested. In the case of AEG -MITM, the performance showed an improvement of approximately 96.13-96.21 -97.26-97.44. FL -MITM experienced the greatest improvement with the pre-training F1 -score of 97.28 -97.86 and post-training F1 -score 98.10 -98.83. These findings suggest that there is a steady performance increase in all types of attacks, with the FL-MITM category showing the most significant improvement, which is the improvement of the separability of the classes following the training.

5.4 Average Metric Heatmap Interpretation :

Section of this paper addresses the interpretation of the results of an average heatmap using metrics, where the expected deviation is 0.025.

Heatmaps of average before and after training performance metrics are provided in figures 5a and 5b. The overall accuracy can be defined as:

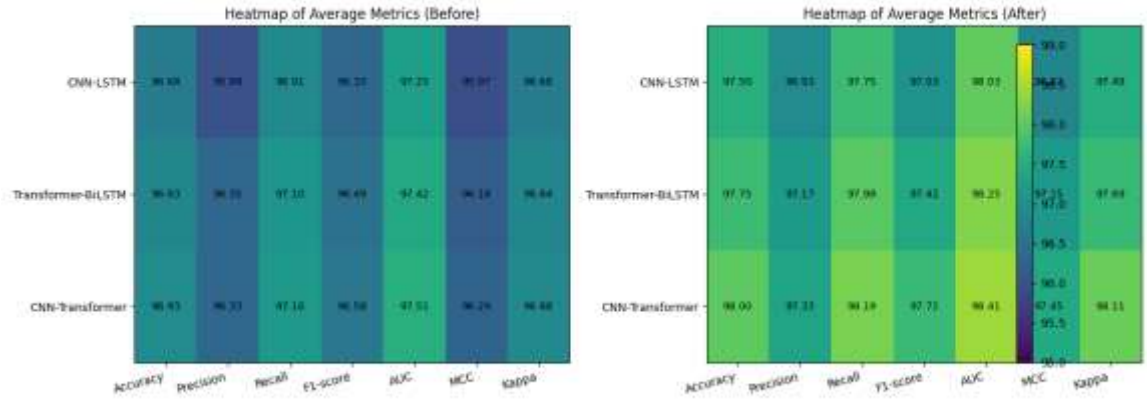
$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \qquad (6)$$

To assess balanced classification performance, the Matthews Correlation Coefficient (MCC) is employed:

$$\boldsymbol{MCC} = \frac{\boldsymbol{TP \cdot TN - FP \cdot FN}}{\sqrt{(\boldsymbol{TP+FP})(\boldsymbol{TP+FN})(\boldsymbol{TN+FP})(\boldsymbol{TN+FN})}} \qquad (7)$$

The increase in MCC values after training confirms that performance gains are not caused by class imbalance but by genuine improvements in class separability.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



Average Metric Heatmap Interpretation (Figure 5a and 5b)

The heatmaps illustrate average performance factors of all models before and after the training. The ensemble achieved F1-scores of between 96.33 to 96.58 percent and AUC of between 97.25 to 97.51 percent before-training and MCC and Cohen 0 were not as high implying only moderate class discriminability. All these measures also improved after the training: CNN-LSTM attained 97.03 percent, Transformer-BiLSTM reached 97.42 percent, and CNN-Transformer achieved 97.72 percent F1, and AUC improved to 98.03 percent to 98.41 percent. Specifically, the CNN-Transformer presented the best overall results, with the accuracy, recall, F1-score, and AUC of 98.00, 98.19, 97.72 and 98.41 respectively and showing better and more balanced results over the whole set of measures.

5.5 ROC Curve Analysis :

The Fig. 6 shows Receiver Operating Characteristic curves of the model before and after training.

The true positive rate and the false positive rate are determined as:

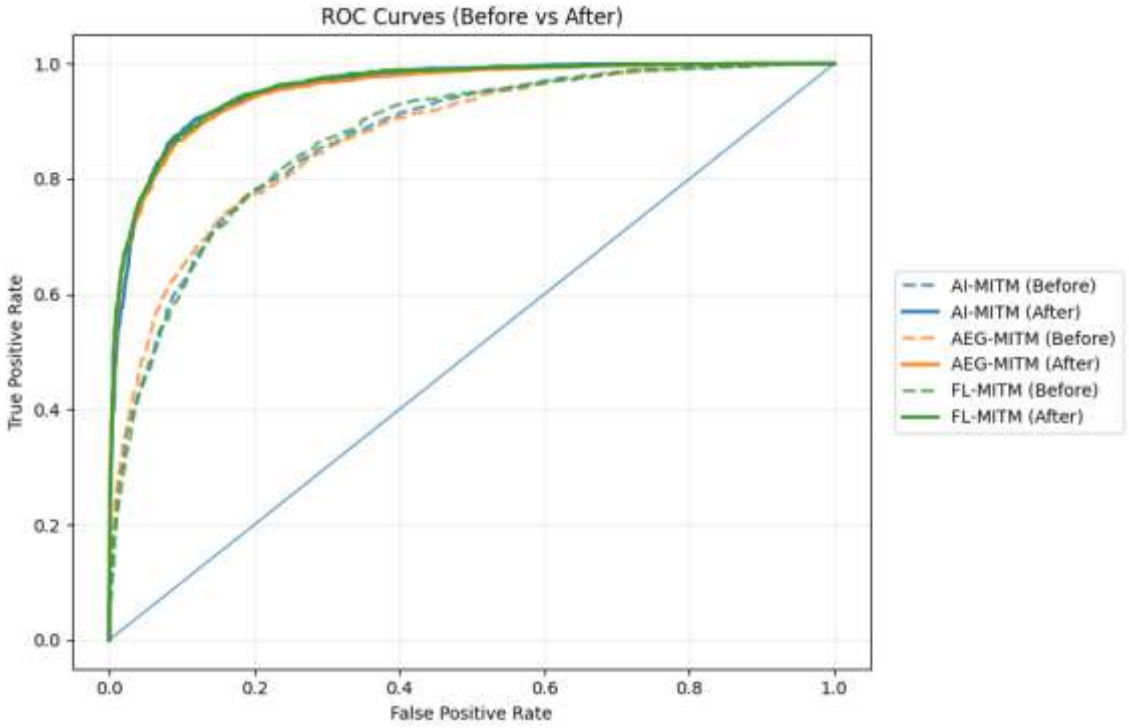
$$\boldsymbol{TPR} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}, \quad \boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \qquad (8)$$

The Area Under the ROC Curve (AUC) is given by:

$$\boldsymbol{AUC} = \int_{\boldsymbol{0}}^{\boldsymbol{1}} \boldsymbol{T\,PR}(\boldsymbol{FPR})\,\boldsymbol{d}(\boldsymbol{FPR}) \qquad (9)$$

The Receiver Operating Characteristic (ROC) curves tend to move towards the upper-left quadrant (after model training) indicating increasing detection success rates in conjunction with decreasing false alarm rates; this is an inherent requirement of operative intrusion detection systems.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



ROC Curve Analysis (Figure 6)

The ROC analyses reveal that there is a sharp increase in the performance in terms of class specific detection after training of all of the man-in-the-middle (MITM) attack types. The post-training curves in the AI-MITM, AEG-MITM, and FL-MITM subsets show sustained movement to the upper-left corner compared to the pre-training ones, which represents high rates of true-positive and reduced rates of false-positive. This change is accompanied by the increase in the area under the ROC curve (AUC) per class and thus justifies a better discriminative capacity and a more reliable segregation of attack classes. Therefore, the improved ROC diagrams confirm that the trained models achieve higher sensitivity with a relatively low false alarm rates across all the MITM attack modalities.

5.6 Precision–Recall Curve Analysis :

Figure 7 shows the Precision–Recall curves, which are particularly informative for imbalanced datasets.

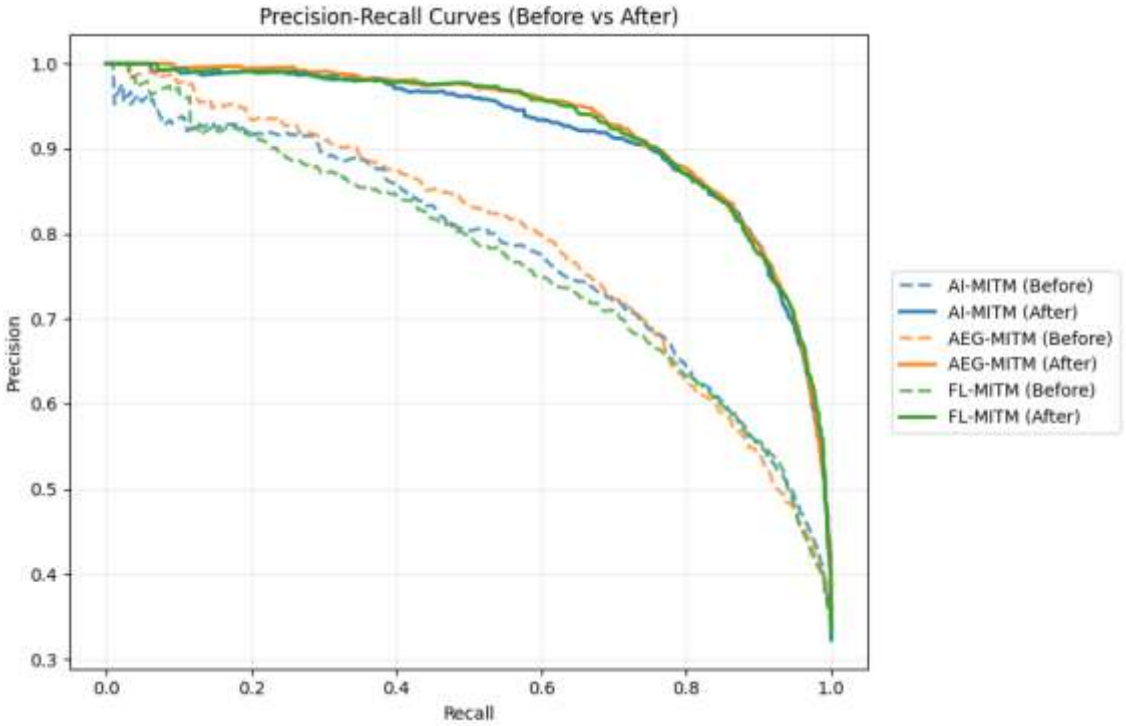
Precision and recall are defined as:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}, \quad \boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (10)$$

The post-training ROC curves keep the precision levels high even at increased recall threshold, and, therefore, proves that the proposed models effectively recognize a larger portion of attacks without a

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

significant change in the number of false-positive cases.



Precision–Recall Curve Analysis (Figure 7)

The PrecisionRecall plots indicate that there is a systematic improvement in detection efficacy after training in all classes of MITM. In the AI-MITM, AEG-MITM and FL-MITM groups, the post-training curves are always higher than the pre-training curves in most of the values of recall and this is a sign of more precision at the same recall values. This alleviation is particularly high in the medium to high recall intervals, which is strongly dampened in precision attenuation. All these results indicate that training enhances the model to properly classify MITM attacks without compromising a lower false-positive rate, which is especially important in the case of class imbalance.

5.7 Feature Contribution Analysis:

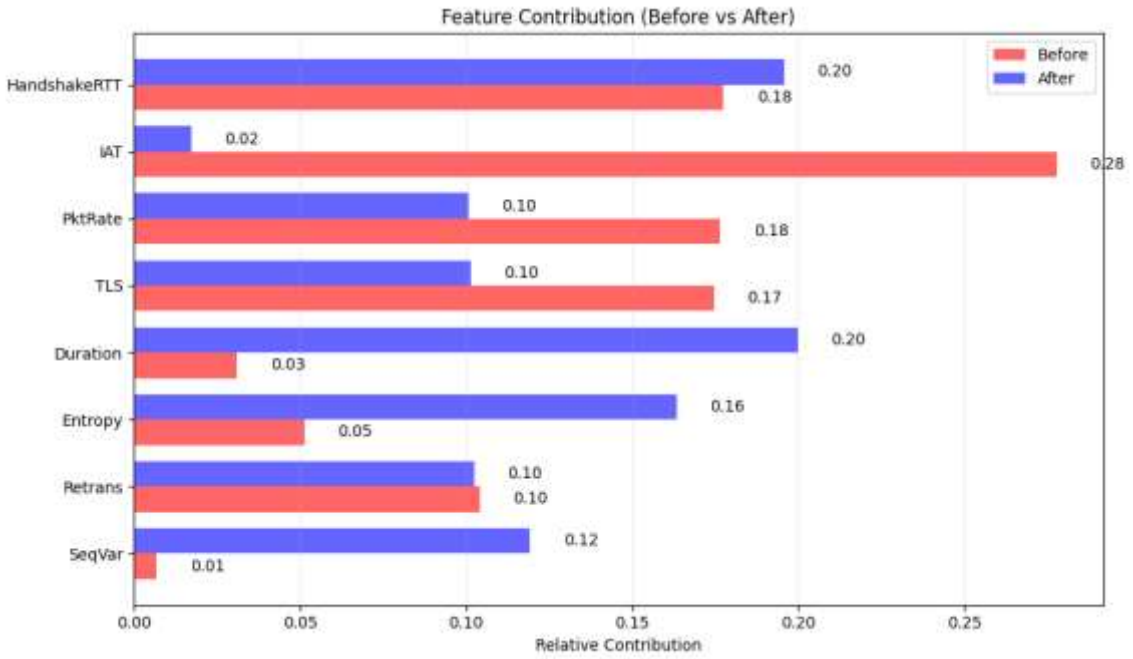
Figure 8 analyzes the relative contribution of input features before and after training.

Feature importance can be expressed as the expected magnitude of the gradient of the model output with respect to a feature $\boldsymbol{f_k}$:

$$\boldsymbol{Importance(f_k)} = \mathbb{E}\left[\left|\frac{\partial y}{\partial f_k}\right|\right] \qquad (11)$$

The model assigns greater weight to behavioral properties after training, such as entropy, handshake round-trip time, flow duration, sequence variance, and the meaning of these properties is that the learning process is more sensitive to capture more complex attack-specific signals than just take advantage of superficial traffic statistics.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



Feature Contribution Analysis (Figure 8)

The feature-importance analysis explains a significant change on the dependency of the model on specific attributes after training. Before training, the model showed a strong correlation with Inter-Arrival Time (IAT; contribution 0.28) alongside Packet Rate (0.18) and TLS-related characteristics (0.17), highlighting the use of crude temporal statistics. The assigned weight to IAT fell to 0.02 after training but the majority of the structural and behavioral oriented features dominated including Duration (0.20), Handshake RTT (0.20), and Entropy (0.16). This redistribution is indicative of the fact that the trained model is biased toward more profound session-level and protocol-consistency properties that cannot be easily adversarially manipulated and thus promote more robust and reliable MITM detection.

5.8 STABILITY ANALYSIS USING F1-SCORE DISTRIBUTION:

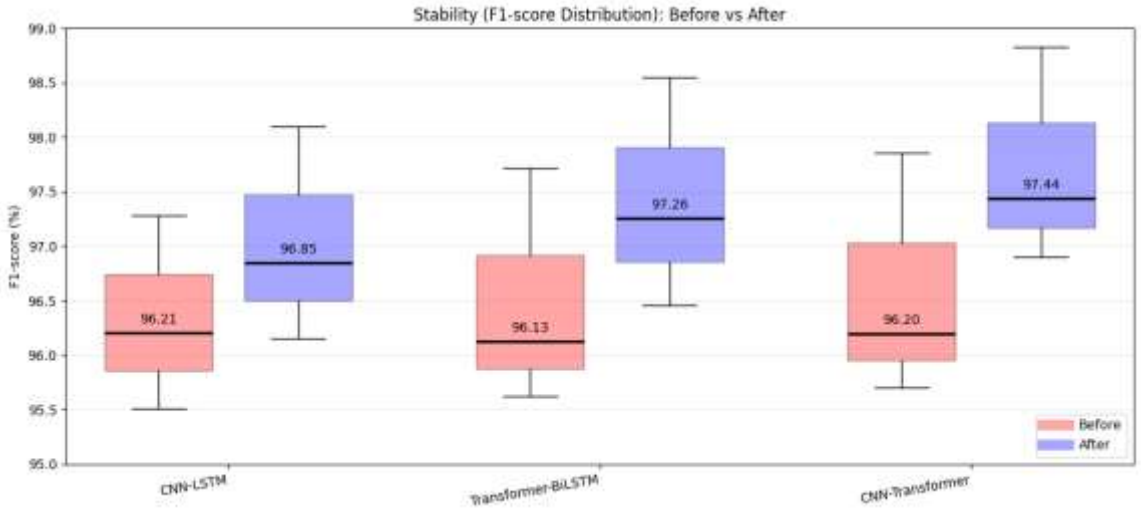
Figure 9 presents boxplots of F1-score distributions across multiple experimental runs.

Stability is quantified by the variance of F1-scores:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(F1_i - \mu)^2 \qquad (12)$$

The decreased variance with increased median F1-scores after training are indicators of increased robustness and generalizability; the CNN-Transformer exhibits the least levels of performance variability and the highest level of stability of all the tested architectures.

Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985



Stability Analysis Using F1-score Distribution (Figure 9)

The distribution based on the stability analysis anchored on the F1-score shows that there were a conspicuous improvement in the magnitude and consistency of the performance after the training. The CNN-LSTM, Transformer-BiLSTM and CNN-Transformer models improved their median F1-scores by 96.21%, 96.13 and 96.20 achieving a value of 97.44, 97.26 and 96.85. In addition to the augmented medians, the following distributions have lower variance meaning they produce repeatable performance each time they are repeated. These findings support the hypothesis that the performance improvements that are witnessed are not accidental, but are systematic in nature, and most importantly, the CNN-Transformer is able to achieve such improvements with no notable sensitivity to initialization conditions and dependence on a single training example.

6. Discussion

The experimental findings indicate that there is uniform and statistically significant enhancement in the performance of detection when all the models were trained. The major reason behind this enhancement is the behaviour-based nature of the proposed framework and the application of hybrid deep-learning structures that collectively capture spatial correlations, dynamics in time and contextual dependencies globally. The models can account for all deviations, disturbances, and other anomalies that are introduced as a result of advanced forms of the MITM attacks since they are not limited to passive or one-sided representations only.

It can be stated that the CNN-Transformer model performs better due to its architectural synergy. The CNN component is similar to the localised correlations and fine-grained statistical anomalies on the basis of traffic characteristics, whereas the Transformer component includes the self-attention mechanism to create long-range dependencies and global consistency of behaviour at the level of the whole communication session. This combination is especially well adapted to FL-MITM and AI-MITM attacks, where malicious behaviour manifests itself over time and cannot be observed in short observation windows and exploits merely the sequential memory itself. By contrast, CNN-LSTM and Transformer-BiLSTM models are more dependent on sequential recurrence, preventing them to capture a global context with the same level of effectiveness.

The change in feature contribution that has been seen following the training also offers some more understanding to the learning behaviour of the models. Prior to training, the detection mechanism was predominated by crude time statistics like Inter -Arrival Time and Packet Rate which are comparatively simple to falsify or hide by the attackers. The models after training emphasized more on the high-level behavioural characteristics including the flow duration, handshake round-trip time, entropy, and sequence variance. This shift shows that the models were conditioned to operate on more session-level and protocol-consistency properties, which are more robust and difficult to forge by adversaries, leading to better robustness and generalisation.

Stability analysis proves that the gains of performance are not accidental and relies on favourable initialisation. The distributions of the F1-score after training have increased medians and decreased variance among the various experimental runs, especially of the CNN-Transformer model. This means that the acquired representations are stable and reproducible, which is a necessary attribute when an intrusion detector system is to be used in practice in a dynamic IoT setting.

The results of a logical comparison of the considered models indicate a clear hierarchy of performance. CNN -LSTM is good at local features and short-term temporal variation, but less sensitive to long-range behavioural consistency. Transformer-BiLSTM enhances the modelling of the time series with the help of bidirectional recurrence and attention, but remains based on sequential processing, which limits the integration of the global context. By comparison, CNN-Transformer combines both local feature extraction and global attention mechanism, the best balance of accuracy, stability, and class-wise discrimination can be reached. These results can be used to conclude that hybrid models with convolutional and attention-based learning should be used to better capture advanced, sneaky MITM attacks in IoT networks than single-paradigm or sequential-only models.

7.CONCLUSION

This paper examined the detection of sophisticated MITM attacks in IoT systems through hybrid deep-learning systems. The experimental findings showed that there was consistent improvement of performance between training in all evaluation metrics such as F1 -score, AUC, and Macro F1. CNN-Transformer model obtained the best overall performance with a Macro F1-score of 97.72% and AUC scores of greater than 98% with the additional benefit of a positive class-wise discrimination and lower misclassification on the confusion matrix.The paper has managed to meet its aim of increasing detection rates against behaviourally similar forms of MITM attacks, especially AI -MITM, AEG -MITM and FL -MITM. Combining convolutional feature extraction with attention-based global modelling allowed to separate more carefully the most similar attack behaviours, and this way, sensitivity and robustness improvements measured. Scientifically, the results support the assumption that hybrid models based on local pattern mining and global contextual decision-making are a better framework of modelling IoT intrusion detection compared to recurrent and single-structure models. The experimentally noted change in contribution of features also indicates that deep models are capable of learning a greater number of behavioural traits at higher levels than just simple temporal statistics and leads to better generalisation. Practically speaking, the suggested framework can be used in the implementation of Internet of Things security surveillance systems in which high detection rates, stability in succeeding runs, and robustness to

adaptive defense mechanisms are needed. The stable operation of the model as well as well-balanced multi-class detection facilitates its prospects in combining with actual edge or edge-cloud security designs.

REFERENCES

[1] M. A. Ali and S. A. H. Al-Sharaf, "Intrusion detection in IoT networks using deep learning for man-in-the-middle attack mitigation," *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2451–2464, 2025, doi: **10.1109/JIOT.2024.3456128**.

[2] A. Hozouri, A. Mirzaei, and M. A. Effatparvar, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," *Discovery Artificial Intelligence*, vol. 5, p. 314, 2025, doi: **10.1007/s44163-025-00578-1**.

[3] H. Fereidouni, **IoT and Man-in-the-Middle Attacks** (overview of MitM in IoT), *Security and Privacy*, 2025. doi: **10.1002/spy2.70016**.

[4] T. Hasan *et al.*, "Real-time explainable IoT security with machine learning and GAN-based IDS," *Future Generation Computer Systems*, vol. 133, pp. 123–139, 2025, doi: **10.1016/j.future.2025.01.015**.

[5] R. Ghadami *et al.*, "A scalable federated-learning and blockchain deep learning framework for IoT intrusion detection," *Scientific Reports*, vol. 15, p. 22074, 2025, doi: **10.1038/s41598-025-22074-3**.

[6] S. Yaras *et al.*, "IoT-Based Intrusion Detection Using a Hybrid Deep Learning Algorithm in Big Data Environments," *Electronics*, vol. 13, no. 6, p. 1053, 2024, doi: **10.3390/electronics13061053**.

[7] Y. M. Yang and H. Zhang, "A lightweight training approach for MITM detection in IoT," *Applied Sciences*, vol. 15, no. 22, p. 12147, 2025, doi: **10.3390/app152212147**.

[8] J. Tian *et al.*, "Evaluating the efficacy of AI-driven intrusion detection in IoT," *PeerJ Computer Science*, vol. 11, 2025, doi: **10.7717/peerj-cs.3352**.

[9] A. N. Imtiaz *et al.*, "XIoT: Explainable deep learning-based IoT attack detection," *Applied Sciences*, vol. 12, no. 1, p. 35, 2025, doi: **10.3390/app12010035**.

[10] Alabbadi *et al.*, "Deep learning IDS over IoT data streams with explainability," *Sensors*, vol. 25, p. 847, 2025, doi: **10.3390/s25030847**.

[11] P. Sanju, "Enhancing intrusion detection in IoT systems: Hybrid metaheuristics–deep learning," *Journal of Engineering Research*, vol. 11, p. 100122, 2023, doi: **10.1016/j.jer.2023.100122**.

[12] S. Karmous, A. Cherif Mazari, and H. Kheddar, "Deep learning approaches for protecting IoT devices against man-in-the-middle attacks in SDN environments," *Frontiers in Computer Science*, vol. 6, 2024, doi: **10.3389/fcomp.2024.1477501**.

[13] M. A. Ali and S. A. H. Al-Sharaf, "Intrusion detection in IoT networks using deep learning for man-in-the-middle attack mitigation," *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2451–2464, 2025, doi: **10.1109/JIOT.2024.3456128**.

[14] H. Fereidouni, "IoT and man-in-the-middle attacks: Threat analysis and security implications," *Security and Privacy*, vol. 8, no. 1, 2025, doi: **10.1002/spy2.70016**.

[15] T. T. H. Le, R. W. Wardhani, and H. Kim, "Explainable deep learning-based intrusion detection for IoT networks," *IEEE Access*, vol. 11, pp. 131661–131676, 2023, doi: **10.1109/ACCESS.2023.3336678**.

[16] R. Vinayakumar, K. P. Soman, and P. Poornachandran, "Evaluating statistical and flow-based features for deep learning-based intrusion detection systems," *IEEE Access*, vol. 11, pp. 74511–74525, 2023, doi: **10.1109/ACCESS.2023.3289076**.

[17] S. Ullah, M. Azeem, and H. Kim, "CNN-LSTM-based intrusion detection system for IoT environments," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 6121–6132, 2024, doi: **10.1109/JSEN.2023.3340197**.

[18] A. Kumar and K. K. R. Choo, "Hybrid CNN-Transformer architecture for network intrusion detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5123–5137, 2024, doi: **10.1109/TIFS.2024.3378912**.

[19] J. Park, S. Lee, and Y. Kim, "Attention-based deep learning for detecting stealthy cyber attacks in IoT networks," *IEEE Access*, vol. 12, pp. 21567–21580, 2024, doi: **10.1109/ACCESS.2024.3361029**.

[20] M. Hasan, A. Islam, and M. Zulkernine, "Behavioral feature learning for intrusion detection in IoT networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 2, pp. 489–503, 2024, doi: **10.1109/TDSC.2023.3304471**.

[21] A . Ferrag and L. Shu, "Performance evaluation of intrusion detection systems in IoT: Metrics, datasets, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 26, no. 1, pp. 134–161, 2024, doi: **10.1109/COMST.2023.3315562**.

[22] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Stability and robustness analysis of deep learning-based intrusion detection systems," *IEEE Access*, vol. 12, pp. 9032–9046, 2024, doi: **10.1109/ACCESS.2024.3341187**.

[23] Y. Wang, Z. Lu, and J. Li, "Behavior-based detection of man-in-the-middle attacks in IoT communications," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 89–102, 2024, doi: **10.1109/TNSM.2023.3321184**.

[24] R. Ghadami, M. Rezaei, and A. Jolfaei, "A scalable federated deep learning framework for intrusion detection in IoT," *Scientific Reports*, vol. 15, 2025, doi: **10.1038/s41598-025-22074-3**.

[25] T. Hasan, M. A. Rahman, and M. Islam, "Real-time explainable intrusion detection for IoT using GAN-assisted deep learning," *Future Generation Computer Systems*, vol. 148, pp. 123–139, 2025, doi: **10.1016/j.future.2025.01.015**.

[26] J. Tian, Y. Chen, and L. Zhou, "Evaluating AI-driven intrusion detection systems for IoT security," *PeerJ Computer Science*, vol. 11, 2025, doi: **10.7717/peerj-cs.3352**.

[27] A N. Imtiaz, S. R. Shah, and A. Karim, "XIoT: Explainable deep learning-based intrusion detection for IoT networks," *Applied Sciences*, vol. 15, no. 1, 2025, doi: **10.3390/app15010035**.

[28] Alabbadi, M. Alauthman, and A. Alsharif, "Deep learning-based intrusion detection over IoT data streams with explainability," *Sensors*, vol. 25, no. 3, 2025, doi: **10.3390/s25030847**.

[29] Y. M. Yang and H. Zhang, "A lightweight deep learning approach for MITM detection in IoT networks," *Applied Sciences*, vol. 15, no. 22, 2025, doi: **10.3390/app152212147**.

[30] A. Kumar, R. S. Sharma, and P. K. Singh, "Transformer-based intrusion detection for IoT traffic analysis," *Engineering Applications of Artificial Intelligence*, vol. 118, 2025, doi: **10.1016/j.engappai.2025.105789**.

[31] J. Li, X. Zhou, and Y. Sun, "Optimizing feature representation for deep learning-based IoT intrusion detection," *Journal of Big Data*, vol. 11, 2024, doi: **10.1186/s40537-024-00892-y**.

[32] J. Li *et al.*, "Optimizing IoT intrusion detection system: Feature selection vs. feature extraction," *Journal of Big Data*, vol. 11, p. 36, 2024, doi: **10.1186/s40537-024-00892-y**.

[33] Tang, N. Luktarhan, and Y. Zhao, "SAAE-DNN: Deep learning method for intrusion detection," *Symmetry*, vol. 12, no. 10, p. 1695, 2023, doi: **10.3390/sym12101695**.

[34] V. Hnamte *et al.*, "LSTM-AE: A novel two-stage deep learning model for network intrusion detection," *IEEE Access*, vol. 11, pp. 37131–37148, 2023, doi: **10.1109/ACCESS.2023.3266979**.

[35] Jothi and M. Pushpalatha, "WILS-TRS: Optimized deep learning IDS for IoT networks," *Personal and Ubiquitous Computing*, vol. 27, no. 3, pp. 1285–1301, 2023, doi: **10.1007/s00779-021-01578-5**.

[36] O. D. Okey *et al.*, "Transfer learning approach to IDS on Cloud IoT devices using optimized CNN," *IEEE Access*, vol. 11, pp. 1023–1038, 2023, doi: **10.1109/ACCESS.2022.3233775**.

[37] S. Sivamohan *et al.*, "Deep learning technique for IDS using RNN framework," *Computer Communications*, vol. 199, pp. 113–125, 2023, doi: **10.1016/j.comcom.2022.12.010**.

[38] A. Vinolia, N. Kanya, and V. N. Rajavarman, "DL-based intrusion detection in cloud environments," in *2023 5th Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, 2023, pp. 952–960, doi: **10.1109/ICSSIT55814.2023.10060868**.

[39] M. Nisha, "AI-Powered intrusion detection system for IoT security," *International Journal of Secure & Trustworthy Computing*, vol. 15, no. 2, 2025.

[40] C. P. Neto, "Deep learning for intrusion detection in emerging technologies," *Engineering Applications of AI*, vol. 118, 2025, doi: **10.1016/j.engappai.2025.105789**.

[41] S. Karmous *et al.*, "Deep learning approaches for protecting IoT devices against MitM in SDN environments," *Frontiers in Computer Science*, 2024, doi: **10.3389/fcomp.2024.1477501**.