Investigating the Impact of Adversarial Evasion Attacks on Model Explanation in IoT-based DDoS Detection Systems

Hafsah Mahmood, Masroor Ahmed, Nadeem Anjum
Capital University of Science and Technology, Islamabad.
*Correspondence: Hafsah.mahmood@cust.edu.pk, masroor.ahmed@cust.edu.pk, Nadeem.anjum@cust.edu.pk

Citation | Mahmood. H, Ahmed. M, Anjum. N, "Investigating the Impact of Adversarial Evasion Attacks on Model Explanation in IoT-based DDoS Detection Systems", IJIST, Special Issue pp 637-655, May 2026
Received | April 06, 2026 **Revised |** May 13, 2026 **Accepted |** May 15, 2026 **Published |** May 18, 2026.

Explainable Artificial Intelligence (XAI) is increasingly being used by intrusion detection systems (IDS) to enhance transparency and enable human-centered cybersecurity decision-making. However, adversarial evasion attacks present a dual threat, as they can mislead both the models themselves and their interpretability outputs intended to explain IDS model predictions. This study investigates how vulnerable DDoS detection in IoT networks is to such adversarial manipulations using the CICIoT2023 dataset. Machine learning and deep learning models, including Random Forest and LSTM, were analyzed using LIME and SHAP frameworks to generate understandable explanations after being trained to detect DDoS traffic. Adversarial perturbed examples were introduced to test the resilience of the explanation, integrity, and accuracy of prediction. The findings indicate that small perturbations can significantly reduce the accuracy of detecting attacks, resulting in false feature attributions. In adversarial settings, LSTM false negatives went up as much as 24,876 to 27,993, decreasing the accuracy from 87.13% to 86.86%, whereas random forest misclassifications went up by 363 to 12,195; however, the accuracy drop was from 99.96% to 98.95%. The overlap in top-5 feature rankings for LSTM was 50%, and that of RF was 45%, respectively. Also, the SHAP cosine similarity declined to 35% for LSTM and 55% for RF, indicating important differences in interpretability. The results of this study highlight limitations in existing explainable IDS approaches and the necessity of adversarial-resistant XAI methods to ensure reliable, trustworthy, and understandable cybersecurity analytics when using the Internet of Things.

Keywords: Explainable IDS, IoT, DDoS Detection, XAI Explanations, Adversarial Perturbations.

Introduction:

The exponential expansion of the Internet of Things (IoT) has transformed the contemporary digital landscape by enabling large-scale automation, connectivity, and real-time data transfer across various sectors, such as smart cities, healthcare, transportation, and industrial control systems. However, this expansion has also made the attack surface bigger, exposing the IoT systems to more advanced cyber-attacks. One of the most common and damaging types of criminal activity amongst them is the Distributed Denial of Service (DDoS) attacks. They can overpower critical elements by drowning them with illegal traffic [1]. Due to their heterogeneous and resource-constrained nature, IoT devices are especially vulnerable. This enables attackers to exploit networked endpoints to make coordinated, large-scale attacks that can completely bring down vital services.

To counter such threats, Intrusion Detection Systems (IDS) that can scan network traffic, detect anomalies, and provide real-time alerts to users have become an important element of network security. Artificial Intelligence (AI) and Machine Learning (ML) have significantly advanced IDS architecture and performance (AI) and Machine Learning (ML) to provide sophisticated data-driven approaches that can autonomously identify intricate attack patterns using immense network data [2]. Pattern recognition and temporal correlation models have demonstrated impressive effectiveness in flow-based data, such as CICIoT2023, with the use of such models as Random Forest (RF) and Long Short-Term Memory (LSTM) networks. [3]

The AI-based intrusion detection systems (IDS) have demonstrated high detection rates, yet have posed additional issues of trust, transparency, and interpretability. Complex AI models' opaque nature has raised concerns about their dependability, especially in critical security scenarios, where understanding the reasoning behind a model's identification of certain traffic as harmful is just as important as the detection itself. This difficulty has inspired Explainable Artificial Intelligence (XAI), an approach to making the decisions made by a model more understandable and interpretable to humans. Such frameworks as Shapley Additive Explanations (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME) can help researchers and practitioners learn more about model predictions at the feature level. This enhances the responsibility and transparency of AI-based cybersecurity. [4]

Nevertheless, regardless of the increased use of XAI methods in IDS, there is an important and unresolved issue: Can such explanations be trusted when the model can be attacked by adversarial examples? In other words, also become unreliable or misleading when an attacker discreetly modifies network characteristics in an effort to trick the IDS. This problem causes a two-level weakness; an attack can harm interpretive tools that are meant to encourage openness and trust, not to mention the prevention of detection.

This is particularly urgent in the context of DDoS detection in the IoT networks, where the integrity of data and real-time reliability play a crucial role. DDoS attacks primarily depend on volumetric patterns and flow-based statistical properties (including packet rate, flow length, and inter-arrival time), most of which can be slightly manipulated by attackers with minor changes that may not be noticeable to human operators but are enough to confuse deep learning or machine learning classifiers [4]. When adversarial samples mimic benign behavior, they can mislead both the model and its explanations given, it makes it more difficult to determine the reality of the situation, as they are making it seem like a benign situation is occurring when a DDoS attack is in progress. [5]. Adversarial evasion attacks, consequently, present a twofold risk, breaching the essence of the reliable AI in cybersecurity by compromising the accuracy of the prediction as well as the integrity of the explanations.

While machine learning and deep learning models are increasingly applied in intrusion detection systems, research primarily focuses on model performance in terms of prediction accuracy or in the form of an independent explainability framework. Although adversarial

evasion attacks affect model performance, the effects on the quality and consistency of model explanations are yet to be explored. Existing explainable models do not evaluate the reliability of explanations under adversarial attacks. Hence, there is a need to understand the effect of adversarial attacks on model predictions and explanations in the context of IoT-based DDoS detection systems.

This study uses the CICIoT2023 dataset to evaluate the robustness of explainable IDS models to adversarial evasion strategies to fill in this gap. This data provides a comprehensive and versatile representation of modern IoT traffic, including benign cases and various types of attacks, including different types of DDoS attacks that simulate real-life hostile network conditions. The paper compares two complementary learning methods: LSTM, a deep learning approach that compares temporal relationships and long-term dependencies between sequential network data, and Random Forest, a more traditional ensemble-based machine learning methodology that is known to be interpretable and resistant to feature noise [6]. These models can be used together to give a comprehensive evaluation of the effects of adversarial modifications on both traditional and deep learning-based detection systems.

Objectives:

This study has three main objectives.

To assess the vulnerability of explainable IDS frameworks, in particular, LIME and SHAP, to adversarial evasion efforts on Internet of Things-based DDoS detectors.

To evaluate the impact of adversarial perturbations on classification performance and explanation stability to classification performance and consistency of explanations with such metrics as confusion matrices, false-negative ratios, and feature attribution changes.

To provide qualitative visualizations that indicate the difference in explanations between adversarial and clean examples, illustrating how it is possible to modify explanations to encourage erroneous model behavior.

By achieving these objectives, this study aims to address an important gap between explainable cybersecurity and machine learning, contributing to the growing discussion of trustworthy AI. The next generation of IDS architectures that should not only resist adversarial inputs but also be able to offer explanations that maintain interpretative integrity even when manipulated with malicious intent will be heavily influenced by the information obtained in the given work.

Novelty of the Study:

This research offers a new perspective on the effect of adversarial evasion attacks on explanations in IoT DDoS detection. This study stands out from current literature that either focuses on prediction accuracy or uses explainability in isolation by examining the interplay between adversarial attacks and explainability.

This study makes the following contributions:

A systematic assessment of evasion attacks on the performance and explanation stability of IoT DDoS detection systems

An investigation of the adversarial robustness of machine learning (Random Forest) vs deep learning (LSTM)

A detailed analysis of explanation variation achieved through SHAP and LIME with and without adversarial attacks

Unveiling model-specific vulnerabilities concerning prediction performance and explanation stability

Experimental results that point to the importance of reliable, explainable IDS

The remaining paper is organized in the following sections. Section II contains the summary of the existing literature on explainable Artificial Intelligence in Intrusion Detection Systems (IDS), adversarial Evasion techniques, and previous work on interpretability in IOT and cybersecurity. This is followed by the proposed methodology in Section III covering the

dataset preprocessing, the architecture of models, and sample generation of adversarial perturbations. The explainable models are also discussed in this section. Section IV contains experimental results and analysis. Section V discusses the conclusion and future work of the research.

Literature Review:

ML/DL-based IoT Intrusion Detection:

The CICIoT2023 dataset has been widely used in recent research for benchmarking machine learning (ML) and deep learning (DL) models for IoT-based intrusion detection, especially for DDoS attacks. Ensemble methods, particularly Random Forest (RF), have performed well on flow-based features with high dimensionality and noise. For example, research tested various ML classifiers on CICIoT2023 and found that RF had an accuracy over 95% in binary DDoS attacks and was more stable and generalizable than conventional classifiers like Decision Trees and k-NN. Likewise, another study showed that RF exhibits stability across different traffic distributions, making it robust for IoT networks.

In terms of deep learning, techniques like LSTM and CNN have been extensively used to model sequential and spatial features of network traffic. A study used LSTM-based models on CICIoT2023 and WUSTL-IIoT datasets and reported a high level of detection accuracy (greater than 96%) as the model captures temporal traffic patterns [7]. Similarly, another study developed a hybrid DL-based IDS with CICIoT2023 and IoTID20 datasets, achieving better detection accuracy over basic ML approaches. But such deep learning models are more complex and harder to interpret [8].

Despite the high classification accuracy, there are some shortcomings. Most of the research is limited to predictive accuracy and does not assess the robustness of the model under adversarial attacks. In addition, although DL models such as LSTM are good at modelling temporal relationships, they can be more sensitive to minor changes in the data. On the other hand, ML models like RF are more robust to small changes in data but might not be robust to strong adversarial attacks. As a result, although RF and LSTM perform well for IoT intrusion detection, little is known about their adversarial robustness, especially in terms of explanation performance.

Explainability in IoT IDS:

To provide transparency regarding the decision-making of ML/DL models, explainable AI (XAI) methods have been increasingly used in intrusion detection systems. Two of the most popular techniques are SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) due to their model-agnostic property and the ability to explain both the overall and local predictions of the model.

Several studies have used SHAP and LIME for interpreting feature contributions and decision-making in IoT IDS. For instance, a study applied SHAP to explain deep learning predictions in IoT, showing better feature explainability [9]. Likewise, another study evaluated SHAP and LIME methods on neural network-based IDS and reported that SHAP offers more stable global explanations, whereas LIME is better at instance-level explanations [10]. Another paper used SHAP explanations in real-time IoT IDS and found improved confidence in analysts through interpretable results [8].

But these techniques also have their drawbacks. Theoretically sound, SHAP may be computationally intensive for large datasets, such as CICIoT2023. LIME, meanwhile, is prone to instability due to sampling and can provide different explanations for similar data points. More significantly, most prior work assumes explanations are consistent under normal circumstances and does not assess their robustness under adversarial attacks. A study showed that adversarial samples can alter feature importance rankings, resulting in explanations that prioritize irrelevant features while providing plausible explanations [11].

Adversarial Evasion in Intrusion Detection:

Adversarial machine learning has become a major concern in intrusion detection systems, especially in IoT networks, where network traffic features are susceptible to slight modifications that can impact classification. Evasion attacks are particularly relevant here, where adversaries manipulate feature values during the inference phase to evade detection without changing the intent of the attack.

Recent research has investigated different adversarial attacks on IDS. Gradient-based attacks like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are common due to their speed and efficacy in creating adversarial examples. A study found that deep learning IDS models suffer from a substantial drop in performance with gradient-based adversarial attacks [11]. Likewise, another study found that even small changes in flow-based attributes can result in misclassification for IoT DDoS detection [12].

Besides white-box attacks, black-box adversarial attacks, which do not rely on model internals, have also been investigated. These attacks are more applicable in the IoT context. But most of the existing research measures the effectiveness of adversarial attacks based on the drop in accuracy or misclassification rate, without studying how these attacks affect the underlying decision-making of models.

Interaction between Adversarial Attacks and Explainability:

A new but relatively unexplored research direction explores the interaction between adversarial attacks and explainable AI in intrusion detection. This is an important intersection because explanation methods like Shapley Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are often assumed to be robust, even in the presence of adversarial examples.

Research has revealed this is not necessarily the case. A study found that adversarial attacks can substantially change feature attribution maps, resulting in misleading explanations that humans find acceptable [13]. Likewise, another study found that under adversarial attacks, explanation methods can attribute high importance to non-relevant or manipulated features, which can render IDS less trustworthy [14].

While some studies aim to assess adversarial robustness and explainability together, they are limited. They typically test small datasets or single models without offering a comprehensive comparison across models. Similarly, there is little research on the consistency across different explainers under adversarial attack.

Research Gap:

In recent years, IoT-based intrusion detection has been greatly improved through the integration of deep learning models (e.g., LSTM, CNN, and their ensembles) with large datasets (e.g., CICIoT2023). Moreover, explainable AI (XAI) methods like SHAP and LIME have been integrated to enhance model interpretability, and other studies have investigated adversarial attacks to assess model robustness.

But most of these works consider classification performance, explainability, and adversarial robustness as separate research challenges. A few studies are exploring the joint impact of adversarial evasion attacks on predictive accuracy and explanation stability. Moreover, current approaches seldom offer a comparative study on both machine learning (e.g., Random Forest) and deep learning (e.g., LSTM) models under the same adversarial attack conditions.

However, this research fills this gap by measuring the impact of adversarial attacks on both classification performance and explanation stability (using SHAP and LIME) on the CICIoT2023 dataset for RF and LSTM models.

Methodology:

The experimental design used to evaluate explainable intrusion detection systems' (IDS) susceptibility to adversarial evasion attacks is explained in this section. Preprocessing

datasets, designing models, creating adversarial perturbations, and assessing explainability are explained. Figure 1 shows the complete methodology. The suggested methodology starts with the first step of data preprocessing, which contains feature cleaning, feature scaling, label encoding, and temporal structure for LSTM. The second step is training and testing on the clean dataset and evaluating the results of ML and DL classifiers based on accuracy, precision, recall, and F1-score. The next step is to train the models again and then add a small percentage of adversarial perturbations on the Test set and reevaluate the ML and DL models to understand the impact of adversarial perturbations on detection accuracy. The next phase is to apply XAI on a clean dataset and on an adversarial perturbed dataset and then evaluate the results by comparing them to understand the shift in the explanations.

Dataset Description:

The CICIoT2023 dataset, a comprehensive and current benchmark created especially to assess IoT network security procedures, is used in the experiments in this paper. With smart cameras, sensors, and home automation devices interacting over a realistic network topology, the dataset mimics a heterogeneous Internet of Things environment. A wide range of legitimate and malicious traffic, including DDoS, DoS, Mirai, UDP flood, TCP SYN flood, and port scanning attacks, are included in CICIoT2023. Since DDoS attacks are among the most common and dangerous dangers in IoT ecosystems, we explicitly focus on them for this study. DDoS attacks result in coordinated traffic floods that overwhelm the IoT devices with limited capacity leading to network unavailability, device failure, and causing disruption of services on a large scale [11].

Stratified sampling was applied to the dataset to balance classes and ensure representativeness by dividing it into training (70%), validation (10%), and testing (20%) subsets. The test sub-group (containing both benign and DDoS samples) contained 1.17 million flow records. Later, both adversarial and clean evaluations were conducted using this test segment.

Data Preprocessing:

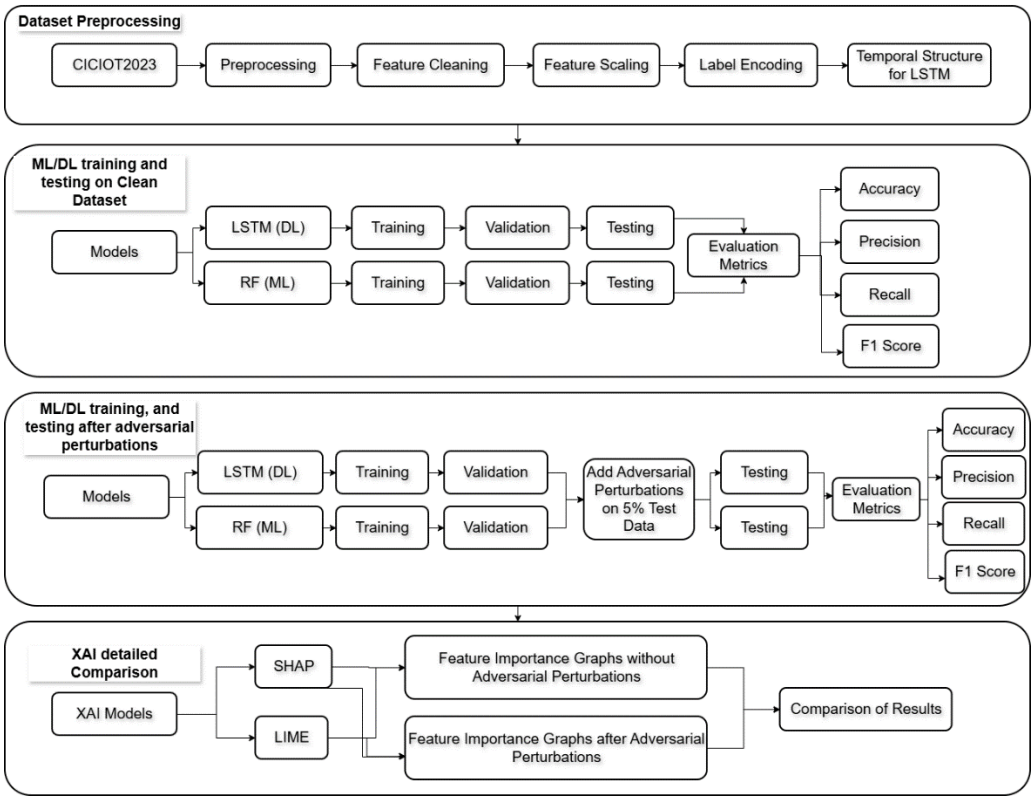

Figure 1. Proposed Methodology

Each flow record in CICIoT2023 contains more than forty statistical and behavioral attributes such as the packet size, flow duration, inter-arrival time, and the count of protocol-specific attributes. The preprocessing steps used to achieve uniform model training and reduce bias based on scale were as follows.

Feature Cleaning: To deal with missing or infinite values, every feature was cleaned. To ensure that the dataset was safe to train the model, 0 was used to replace NaN (Not a Number) and INF (Infinity) values.

Scaling of the Numbers feature The Standard Scaler was used to scale the numbers feature; to give each feature a mean of 0 and a standard deviation of 1, which ensures that every feature plays an equal part in the training process.

Label Encoding: The labels on attacks were changed to a binary form, where 0 was benign traffic and 1 was DDoS. Using common binary classification criteria, this approach streamlines performance evaluation.

Temporal Structuring for LSTM: To maintain sequential dependencies between flow-level features, feature sequences were transformed into a 3D tensor of size (samples, 46, 1) for LSTM trials

To ensure repeatability and computational efficiency, all preprocessing was done in Python using the Pandas and Scikit learn modules

Model Architectures:

To compare adversarial and interpretability behavior across learning methods, two different model families were selected to represent machine learning and deep learning paradigms.

Random Forest:

The machine learning baseline was the Random Forest (RF) model, a popular ensemble technique. RF reduces variation and mitigates overfitting by building several decision trees and aggregating their predictions via majority voting. The model setup used in this investigation comprised:

There are two hundred trees.

Maximum depth: 20

Gini impurity as a criterion

Enabling bootstrap sampling

The Scikit-learn implementation was used to train the RF classifier, and class balancing was enabled to address a slight dataset imbalance between DDoS and regular traffic.

Long-Short-Term Memory (LSTM):

As the LSTM network can capture temporal correlations in sequential network traffic, it was chosen as the deep learning baseline. The architecture employed in this investigation included:

Flow features are represented by the input layer (46, 1).

There are two LSTM layers, each with 128 and 64 hidden units.

To avoid overfitting, use dropout regularization (0.3).

Sigmoid activation in a fully linked (Dense) layer for binary categorization.

Binary cross-entropy was used as the loss function, and the Adam optimizer with a learning rate of 0.001 was employed to train the model. Thirty epochs of training were conducted, with early termination determined by the convergence of the validation loss.

Adversarial Perturbation Generation:

We created adversarial flow records by adding small, feature-space perturbations in order to evaluate how vulnerable the trained IDS models (RF and LSTM) were to evasion attacks. To maximize the loss function with respect to input perturbation while preserving statistical proximity to the initial benign or DDoS flow record$x$, adversarial instances x' were created. The Fast Gradient Sign Method (FGSM), a first-order attack technique selected for

its effectiveness and efficiency against deep learning architectures, served as the model for this procedure.

Core Mathematical Framework:

The trained model's fundamental function, F, is parameterized by the fixed weights θ and maps an input feature vector x to a projected output $\hat{y}$ (the likelihood of being a DDoS flow):

$$\hat{y} = F(x;\theta)(1)$$

The attack is presented as an optimization problem in which the perturbation η is made to maximize the loss function $J(\theta, x, y)$, which in this study is usually Binary Cross-Entropy (BCE), which is defined as follows:

$$J(\theta, x, y) = -[ylog(y) + (1-y)log(1-y)](2)$$

Where the input feature is denoted by x

y Belongs to $\{0,1\}$ representing a binary class problem where 0 refers to the benign class, and 2 refers to a DDoS attack.

Fast Gradient Sign Method (FGSM) and L_∞ Constraint:

To assess the model's performance under adversarial attacks, we use the Fast Gradient Sign Method (FGSM) to generate adversarial examples. The technique generates adversarial perturbations by adding a small, targeted perturbation to the input in the direction that maximizes the loss.

For an input$x$, a true label y, model parameters θ, and the loss function $J(\theta, x, y)$ defined as binary cross-entropy, the adversarial perturbation η is given by:

$$\eta = \epsilon \cdot sign\big(\nabla x J(\theta, x, y)\big)(3)$$

Where:

η is the adversarial perturbation to the input

ϵ is a small constant

$\nabla x J(\theta, x, y)$ is the gradient of the loss function with respect to x

$sign(\cdot)$ is the sign of the gradient

The adversarial example x_{adv}

$$x_{adv} = x + \eta\ (4)$$

Can then be generated by adding the perturbation to the input sample:

By doing this, we ensure that the perturbation is computed in the direction of the gradient of the loss function, and thus the adversarial sample is misclassified.

In this paper, the perturbation is generated using the gradient of the binary cross-entropy loss, which means that the adversarial examples specifically increase the classification error between the model's output $\hat{y}$ and the ground truth$y$.

The single-step FGSM approach allows efficient generation of adversarial examples, which can be applied to large-scale datasets like CICIoT2023 and can be used to simulate evasion attacks in IoT-based intrusion detection systems.

Most existing studies evaluate the model's degradation in classification after applying perturbations. However, this study focuses on the evaluations of the explanations generated by explainable models and their consistency and reliability after being influenced by the perturbations.

Explainability Framework:

Two post-hoc explainability frameworks, LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations), were used to assess the impact of adversarial perturbations on model interpretability.

Lime Explanations:

By using altered samples to train a linear surrogate model, LIME approximates the local decision boundary of a model around a particular instance. LIME determined the top 5 most important characteristics influencing the anticipated label for every test instance (clean and hostile). The explanation consistency was measured by comparing the overlap between these properties for adversarial versus clean occurrences. A notable decrease in overlap suggests that explanation stability is reduced under adversarial perturbations.

SHAP Explanations:

We used Shapley Additive Explanations (SHAP) as a game-theoretic framework to precisely measure the contribution of each input feature to the model's output to supplement the feature set overlap analysis offered by LIME. SHAP provides an attribution vector $A = [\phi_1, \phi_2, \ldots, \phi_D]$ for each test instance, where ϕ_i is the SHAP value of the feature i, indicating its degree and direction of effect on the prediction. The Cosine Similarity between the SHAP attribution vector computed on the clean sample (A_{clean}) and the equivalent vector on the adversarial altered sample (A_{adv}) is the main metric used to evaluate interpretability robustness. By measuring the angular difference between the two vectors, regardless of their magnitude, this metric evaluates the extent of Explanation Drift. The normalized inner product is used to determine the Cosine Similarity (CS):

$$Cosine\ Similarity = \frac{A_{clean} \cdot A_{adv}}{\| A_{clean} \| \| A_{adv} \|} \quad (6)$$

Where the dot product of the clean and adversarial SHAP vectors is denoted by:

$A_{clean} \cdot A_{adv}$. The L_2 norms (Euclidean magnitudes) of the corresponding vectors are represented by $\| A_{clean} \| \| A_{adv} \|$. High consistency is indicated by a Cosine Similarity value near 1.0, indicating that the adversarial approach had little effect on the attributes that contributed to the prediction (the model's "reasoning"). On the other hand, a value close to zero indicates a substantial change in the direction of the attribution vector, indicating susceptibility to misleading interpretability. Reduced interpretive stability and successful deception of the model's explanations are consequently indicated by lower cosine similarity scores.

Explanations Impact Evaluation:

To evaluate the impact of adversarial perturbations on explanation stability, to analyze the impact of the adversarial perturbations on model explanation, we compared the explanations generated before and after the adversarial perturbations.

The explanations of both SHAP and LIME are recorded under two scenarios.

Explanations on the clean dataset

Explanations on the Adversarial dataset

The features that are contributing towards the explanations are then compared among the above two instances to produce quantifiable results shown in the explanation patterns.

Similarity-based analysis is conducted to measure the variations produced in pre- and post-attack explanations. The score of the similarity index represents the deviation of the explanations. If the score is less, the deviation is high, and if the similarity score is high means the deviation is less and the explanations are not impacted to a greater extent. This step is particularly taken to ensure that the experimental steps are not only conducted to measure the perturbation effect on the classification performance but also on the reliability and stability of XAI.

Evaluation Metrics:

A two-layer evaluation technique was used to evaluate both interpretability and classification performance.

Metrics for Classification:

The model prediction robustness against adversarial perturbations was evaluated using conventional performance metrics:

Accuracy (ACC)
Precision (P)
Recall (R)
F1-Score (F1)
False Negative Rate (FNR) for DDoS

The degree of DDoS detection failure under adversarial settings was revealed by visualizing misclassifications using a confusion matrix.

Explainability Metrics:

To measure deceptive explanations, two complementary metrics were introduced:

The cosine SHAP vector similarity: Evaluates the degree of consistency in feature attributions following adversarial disruption. Stable explanations are indicated by values close to 1, whereas abrupt alterations are suggested by values close to 0.

Top 5 Feature Overlap (LIME): The top five features were recorded for comparing the results of LIME with clean and adversarial instances. The % overlap was used to determine the consistency.

Experimental Setup:

The experimental configuration for conducting the experiments utilizes a high-performance computing environment.

CPU: 2.40 GHz Intel Xeon
GPU: Tesla T4 from NVIDIA
16 GB of memory
Frameworks: SHAP 0.44, LIME 0.2, Scikit-learn 1.4, TensorFlow 2.15

To guarantee uniformity, each experiment was conducted three times, and all numerical results were averaged. For reproducibility, the trained models, SHAP values, and LIME explanations were automatically archived into a downloadable ZIP file and kept in an organized folder hierarchy.

Results and Discussion:

The results are discussed in this section in terms of classification accuracy and false negatives for both the Random Forest and LSTM on both the clean and adversarial datasets. The explanations of both models using LIME and SHAP are further assessed under both scenarios, including a dataset without any perturbations and with perturbations. DDoS detection is highlighted, and both explanation fidelity loss and prediction accuracy degradation under adversarial circumstances are investigated.

Classification Performance on Clean and Adversarial Datasets:

LSTM Results:

To address objective 2, the classification performance of LSTM was evaluated. On the clean CICIoT2023 test set, the LSTM model demonstrated excellent baseline performance. Figure 2 (Confusion Matrix LSTM Clean) illustrates the model's overall accuracy of 87.13%. Of the total DDoS occurrences, 24,876 DDoS flows were incorrectly categorized as benign, resulting in a false-negative rate of 2.1%. Table 1 shows the classification results on LSTM.

Performance loss was visible when small adversarial perturbations were introduced (Figure 3, Confusion Matrix LSTM Adversarial). The number of incorrectly categorized DDoS flows increased to 27,993, indicating a 12.5% rise in false negatives, despite the overall accuracy slightly declining by 0.26%.

This shows that in IoT systems, even little disruptions can have a disproportionate effect on the detection of important attack-traffic. Table 2 shows the classification results on LSTM for adversarial attacks.

Table 1. Classification Report of LSTM on Clean Dataset

Category	Precision	Recall	F1-Score	Samples
Class 0	88.65%	60.55%	71.95%	320,870
Class 1	86.78%	97.09%	91.65%	855,981
Overall Accuracy	–	–	**87.13%**	1,176,851
Macro Average	87.72%	78.82%	81.80%	1,176,851
Weighted Average	87.29%	87.13%	86.28%	1,176,851

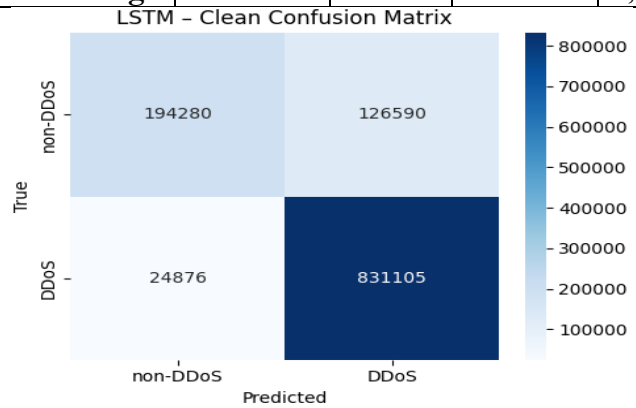

Figure 2. Confusion Matrix LSTM Clean

Table 2. Classification Report of Adversarial Set on LSTM

Class / Average	Precision	Recall	F1-Score	Support
0	0.8741	0.6055	0.7154	320,870
1	0.8674	0.9673	0.9146	855,981
Accuracy	–	–	**0.8686**	1,176,851
Macro Average	0.8707	0.7864	0.8150	1,176,851
Weighted Average	0.8692	0.8686	0.8603	1,176,851

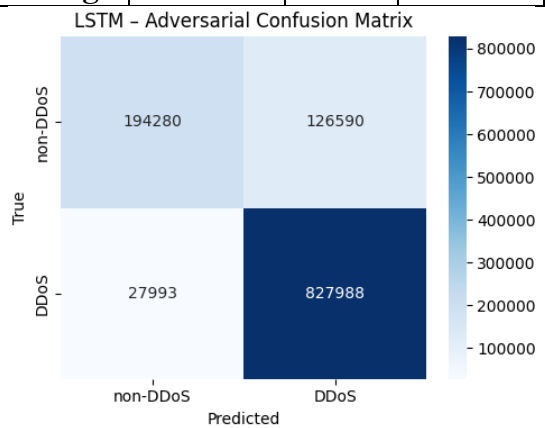

Figure 3. Confusion Matrix of LSTM on Adversarial Set

Random Forests Results:

With an overall accuracy of 99.96%, the Random Forest classifier demonstrated similar clean-data performance (Figure 4, Confusion Matrix RF Clean). Only 363 DDoS flows were incorrectly identified as regular on the clean test set. Table 3 shows the classification results on Random Forest for the clean dataset.

Adversarial changes, however, resulted in a sharp spike in false negatives to 12,195 flows (Figure 5, Confusion Matrix RF Adversarial), signifying a 3,260% increase in the rate of misclassification. Even when the perturbations are small, this great sensitivity highlights how susceptible tree-based models are to targeted feature perturbations. Table 4 shows the classification results on Random Forest for adversarial attacks.

Table 3. Classification Report of Random Forest on Clean Set

Class / Average	Precision	Recall	F1-Score	Support
0	0.9989	0.9995	0.9992	320,870
1	0.9998	0.9996	0.9997	855,981
Accuracy	–	–	**0.9996**	1,176,851
Macro Average	0.9993	0.9995	0.9994	1,176,851
Weighted Average	0.9996	0.9996	0.9996	1,176,851

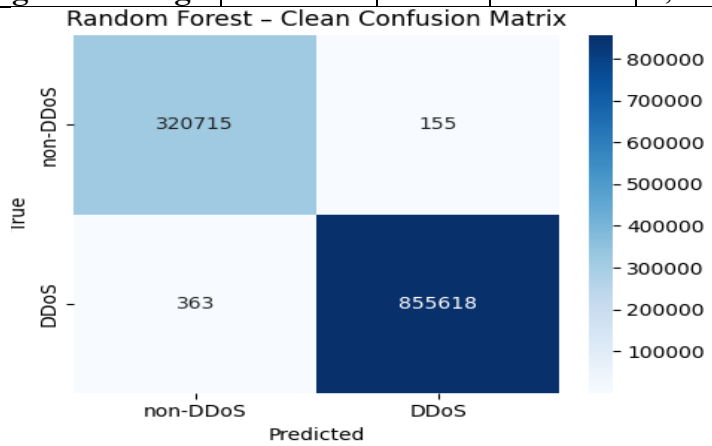

Figure 4. Confusion Matrix of Random Forest on Clean Set

Table 4. Classification Report of Random Forest on Adversarial Set

Class / Average	Precision	Recall	F1-Score	Support
0	0.9634	0.9995	0.9811	320,870
1	0.9998	0.9858	0.9927	855,981
Accuracy	–	–	**0.9895**	1,176,851
Macro Average	0.9816	0.9926	0.9869	1,176,851
Weighted Average	0.9899	0.9895	0.9896	1,176,851

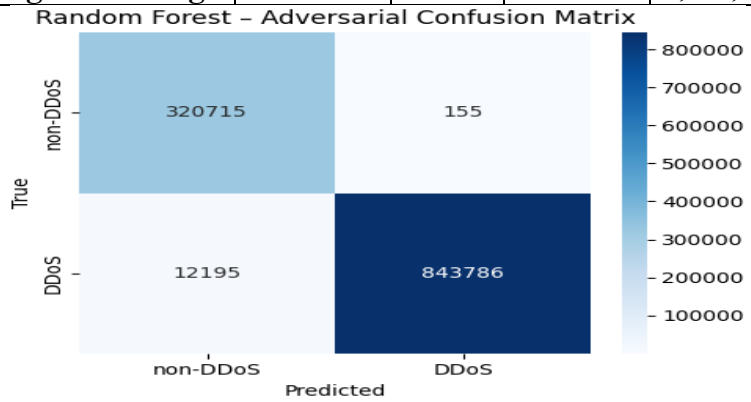

Figure 5. Confusion Marix of Random Forest on Adversarial Set

Receiver Operator Characteristics Curve (ROC):

ROC curves were used to further assess how robust each model's discrimination ability was. Figures 6, 7, 8, 9 show the LSTM and RF ROC curves for clean versus adversarial data, respectively. Under adversarial conditions, the area under the curve (AUC) for both models decreased somewhat, from 0.93 to 0.92 for LSTM and does not change for RF.

Based on the results from the ROC analysis, it is evident that the performance of the LSTM model is highly impacted by adversarial perturbations, while at the same time, the Random Forest model shows relative stability in terms of its ROC attributes. This can be explained by analyzing the underlying architecture of the two approaches used for the experiments.

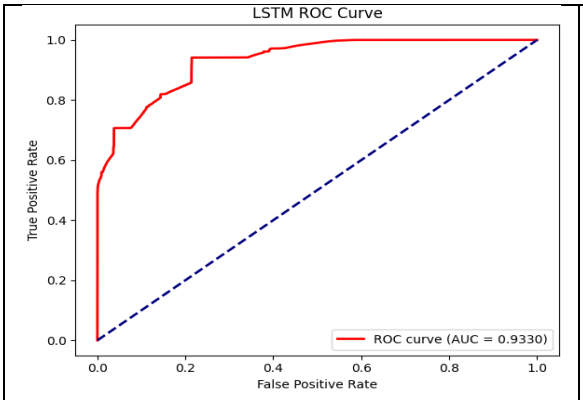

Figure 6. ROC Curve of clean LSTM

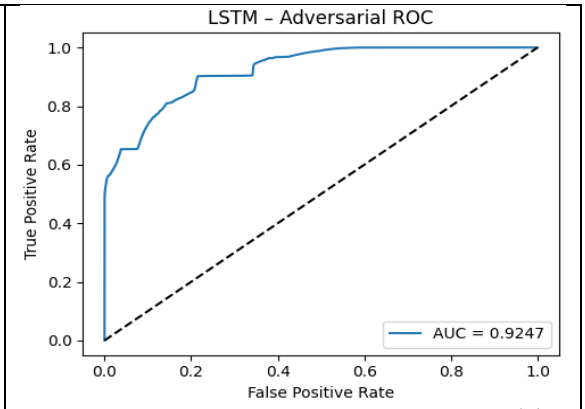

Figure 7. ROC Curve for adversarial LSTM

Since LSTM is a deep learning algorithm aimed at capturing sequence dependencies, it greatly depends on learned representations. Adversarial perturbation causes a disruption of the model's capacity to make a distinction between benign examples and attacks because it's learned patterns are distorted by even minor perturbations. As a consequence, a decrease in true positive rates and an increase in false positives cause performance degradation.

At the same time, the Random Forest model is characterized by an ensemble of decision trees, with the final prediction being produced as the result of multiple splits performed on different feature subspaces independently. Therefore, due to its unique architecture, Random Forest is not significantly impacted by the input alterations.

The above observations suggest that while deep learning models are highly susceptible to adversarial perturbations, machine learning approaches are more robust to such distortions.

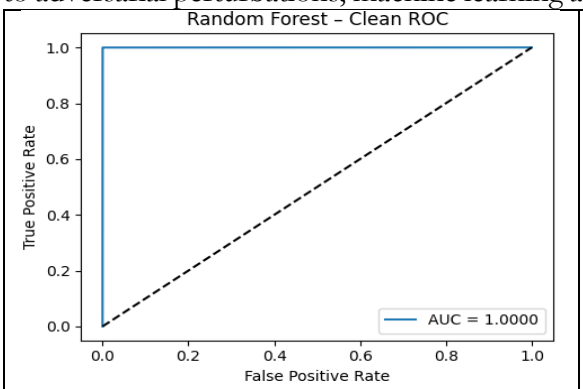

Figure 8. ROC Curve for clean RF

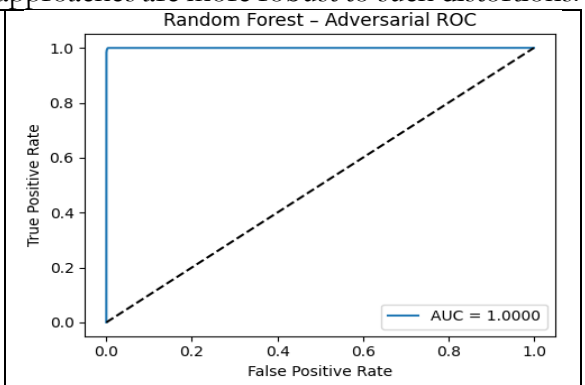

Figure 9. ROC Curve for Adversarial RF

Explanations Analysis using LIME:

LIME was used on both adversarial and clean cases to evaluate interpretability degradation. The evaluation on the explanations of LIME answers the first objective of this study.

Explanations of Random Forest:

The major characteristics influencing DDoS detection for three sample indices (284, 290, 298) are highlighted in representative LIME visualizations for clean RF predictions (Figures 10, 11, 12). Feature contributions that highlight high packet speeds, irregular flow durations, and traffic bursts are in line with domain expertise. LIME explanations for the same indices (Figures 13, 14, 15) demonstrate a significant change in feature relevance following adversarial disruption, with a number of previously dominating features either losing their influence or mistakenly implying benign traffic characteristics. Significant interpretability drift was demonstrated quantitatively, as the top 5 feature overlap between adversarial and clean explanations dropped to 40–50%.

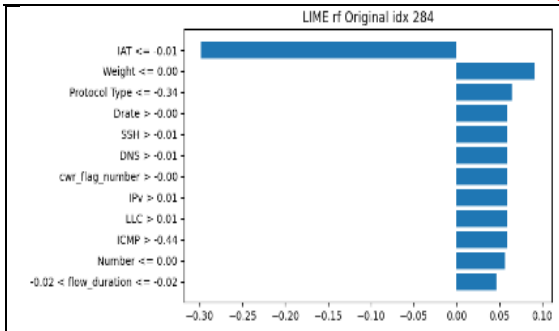

Figure 10. LIME results on idx 284 on RF clean

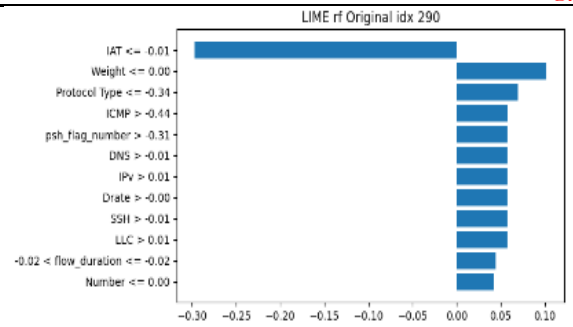

Figure 11. LIME results on idx 290 on RF clean

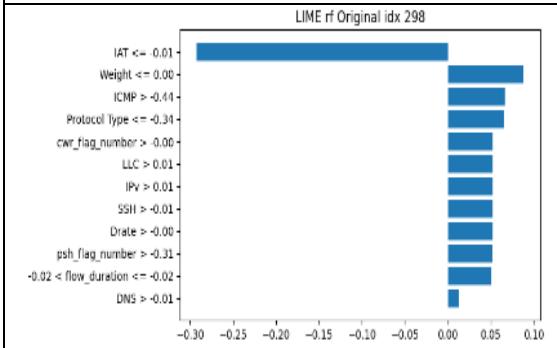

Figure 12. LIME results on idx 298 on RF clean

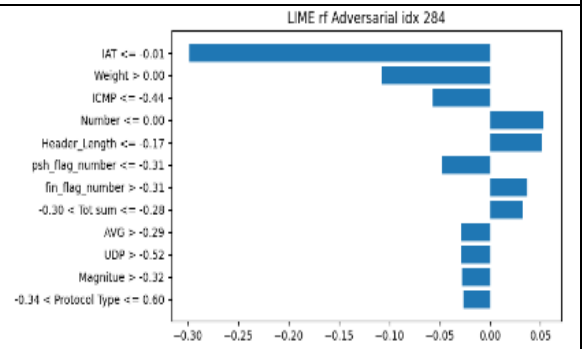

Figure 13. LIME results on idx 284 on RF adversarial

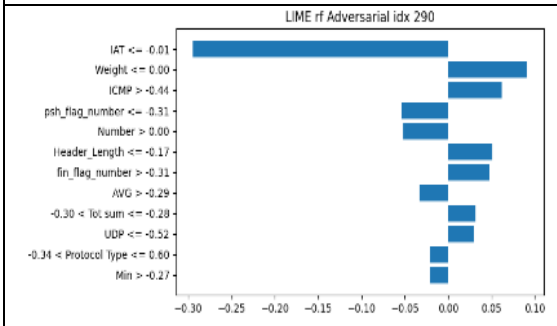

Figure 14. LIME results on idx 290 on RF adversarial

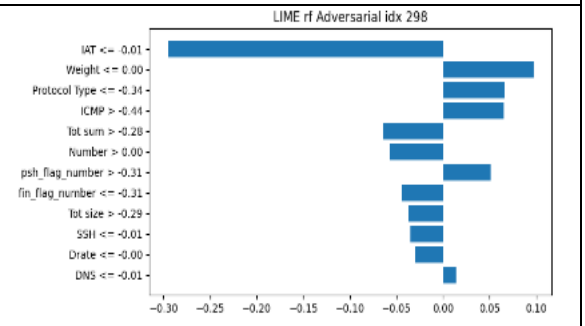

Figure 15. LIME results on idx 298 on RF adversarial

Explanations of LSTM:

Clean instances of LSTM (Figures 16, 17, 18 on indices 3, 10, 63) showed feature contributions that highlighted temporal dependencies suggestive of high-intensity DDoS flows. LIME explanations after applying the adversarial perturbations were then assessed against the same indices as shown in Figures 19, 20, and 21, showing two of the possibilities. Either their contributing top significant patterns of features become hidden, or on the other hand, they get inverted which allows the model to use coherent feature contributions that are deceptive in terms of justification. This observation draws attention to a crucial problem: explanations may be disproportionately deceptive, creating the appearance of model reliability even when predictions are somewhat impacted.

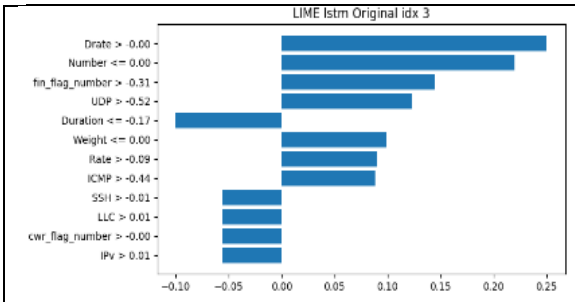

Figure 16. LIME results on idx 3 on LSTM clean

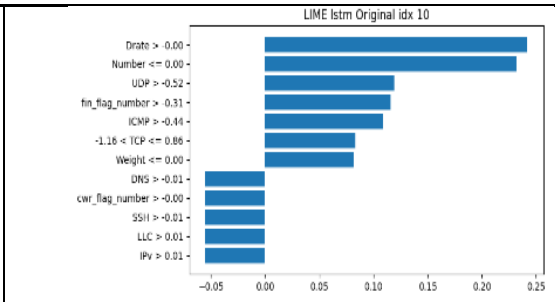

Figure 17. LIME results on idx 10 on LSTM clean

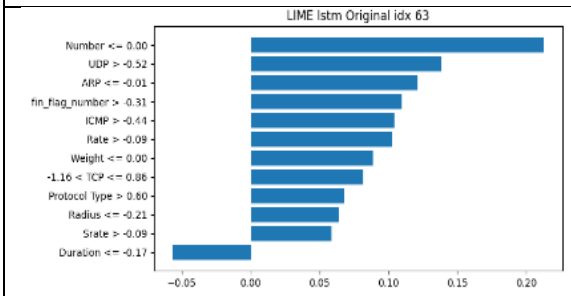

Figure 18. LIME results on idx 63 on LSTM clean

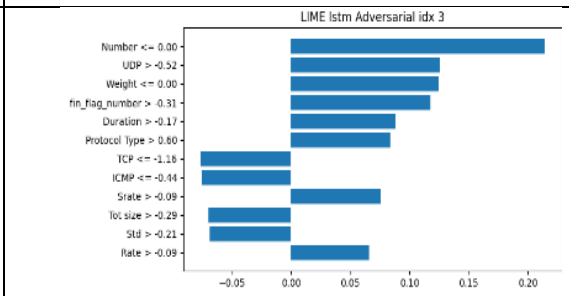

Figure 19. LIME results on idx 3 on LSTM adversarial

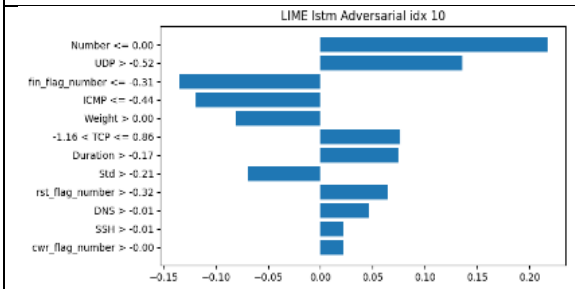

Figure 20. LIME results on idx 10 on LSTM adversarial

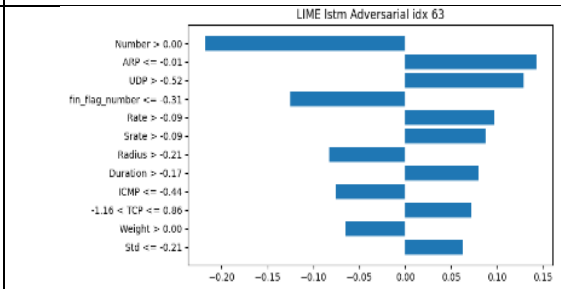

Figure 21. LIME results on idx 63 on LSTM adversarial

Explanations Analysis using SHAP:

The feature-level contributions of SHAP were recorded on the test set. The SHAP values with the clean dataset on the classification of DDoS versus non-DDoS are shown in Figure 22. When adversarial perturbations were applied, the feature contribution was disturbed, and the contribution of the top features towards the prediction decreased. The comparisons answer to the objective 2.

An average similarity of 0.55 for RF was found by cosine similarity analysis between clean and adversarial SHAP vectors, indicating a significant divergence in explanations. This suggests a shift in feature attribution patterns, indicating potential instability in the learned decision boundary under adversarial perturbations. The LSTM feature contribution in the test set was also recorded, as shown in Figure 23. It contains the average value of the features based on two classes. The results before and after adversarial perturbations showed an average of 0.35 similarity index, which depicts that LSTM explanations using SHAP were more sensitive to adversarial perturbations.

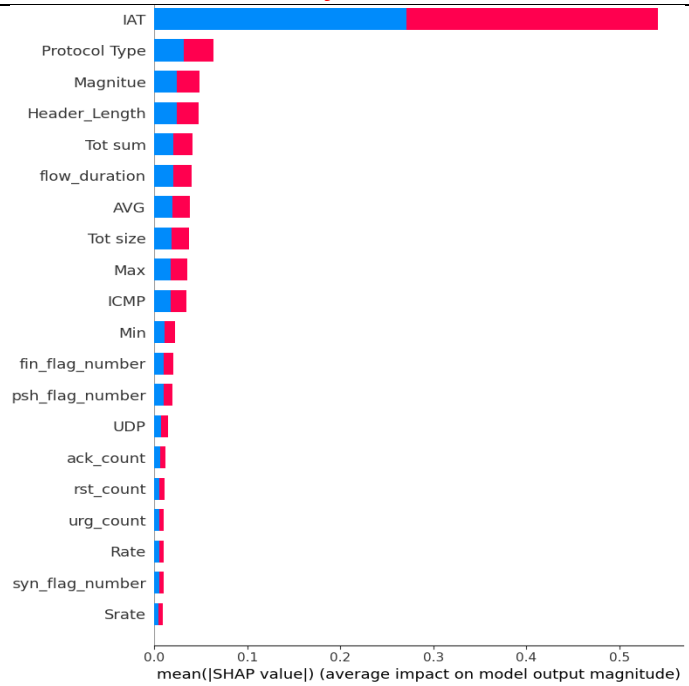

Figure 22. Mean absolute SHAP values for clean vs attack DDoS occurrences

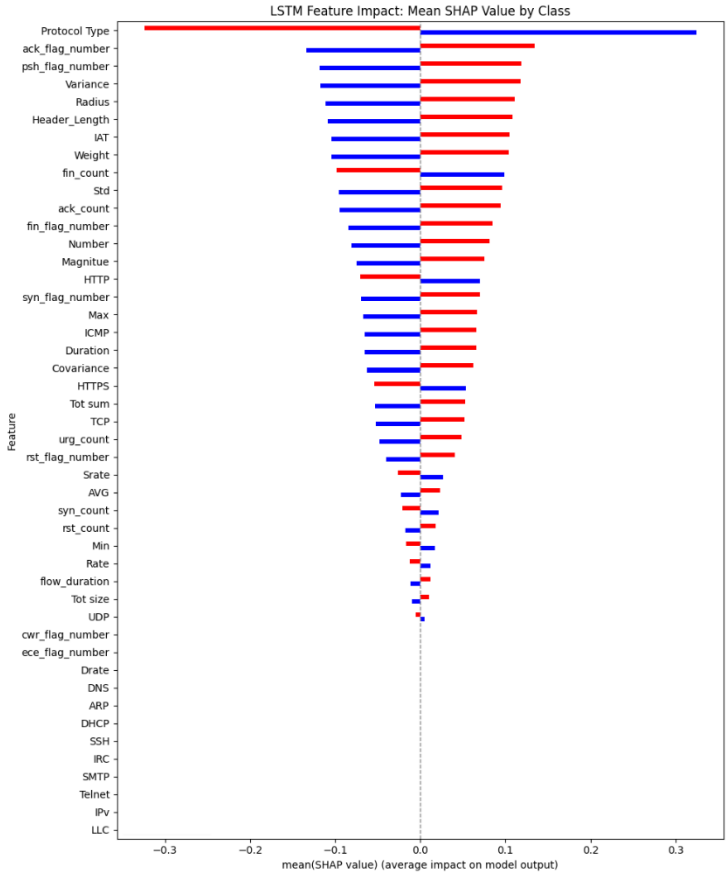

Figure 23. Mean absolute SHAP values for clean vs attack DDoS occurrences

Discussions:

The results provide a complete analysis of models' degradation in terms of detection and interpretability.

Prediction degradation: It is noted that altering only 5% of the test dataset leads to a significant increase in false negatives, as the results indicate an increase of false negatives from

24,876 to 27,993 in LSTM. This increase was even more pronounced for RF. The number increased from 363 to 12,195.

Interpretability compromise: The explanations of both LIME and SHAP were also compromised as the top contributing **feature attributions** were altered, resulting in **misleading results**.

Model-specific sensitivity: The results were model-dependent. For instance, in terms of performance or detection accuracy, the RF was more sensitive to perturbations, whereas the results were opposite in terms of explanations, which are depicted by LIME and SHAP values. Visual representations of the explanations are provided to address Research Objective 3.

The outcomes of this study are aligned with modern literature that underscores the susceptibility of intrusion detection systems to adversarial attacks. According to earlier works, adversarial perturbations can substantially affect detection accuracy; however, previous research was focused solely on detection performance without examining interpretability factors. On the contrary, the results of the current study indicate that, apart from influencing prediction accuracy, adversarial perturbations can lead to unstable explanations.

In addition, previous investigations related to explainable intrusion detection techniques found that methods based on SHAP and LIME offer valuable insights into model decision-making in regular conditions. Nevertheless, these studies did not assess the effectiveness of SHAP and LIME-based explanations under adversarial settings. As indicated in the present research, both techniques showed high variability in terms of stability when confronted with adversarial perturbations, thus questioning their applicability in security-sensitive environments.

Moreover, the discrepancy between the outcomes for Random Forest and LSTM is congruent with earlier research, according to which deep learning architectures are generally more susceptible to adversarial perturbations compared to ensemble models. For instance, in this case, the ROC-AUC performance for LSTM is considerably lower than the RF results.

Implications of the Study:

The findings indicate that explanation approaches are non-robust by nature and can be heavily impacted by changes in input features.

In turn, this contradicts the widespread belief that the explanation algorithms can adequately describe the true mechanism of decisions made by a model during the prediction process. The vulnerability of SHAP and LIME explanations calls for the development of more stable explanation methodologies that can withstand adversarial attacks.

Moreover, this paper stresses the necessity of conducting a joint analysis of the model's robustness and interpretability. This issue is an important research avenue for further exploration in the field of security-sensitive applications.

Research Contributions:

The main contributions of this study are summarized as follows:

In this paper, we evaluate adversarial evasion attacks on IoT-based DDoS detection systems and provide deep insights into their behavior using the CICIoT2023 dataset.

We conduct an experimental study comparing the robustness of machine learning versus deep learning approaches. We test the state-of-the-art Random Forest approach alongside a deep learning approach using Long Short-Term Memory (LSTM) cells.

This paper presents a quantitative measure of explanation stability using a SHAP-based similarity index. This measure allows for the calculation of changes in feature assignments before and after adversarial attacks.

We also examine local interpretability through the lens of LIME-based analyses, revealing deviations in instance-level explanations generated in the presence of adversarial perturbations.

This work provides a visual analysis of explanation patterns within machine learning models for intrusion detection, analyzing how feature importance distributions change when models are subject to adversarial attacks.

Recommendations:

From the results obtained in this study, the following recommendations are offered for the deployment and development of intrusion detection systems based on secure IoT technology:

IDS should avoid using only classification accuracy since attacks can reduce both classification accuracy and **interpretability**.

IoT applications that involve security issues should perform robustness analysis for ML and DL approaches in adversarial settings before implementation.

SHAP and LIME methods should be employed together with a stability measure for trustworthy feature attribution under attack settings.

Random forest, an ensemble approach, could be more suitable in IoT scenarios where computation resources are limited since its robustness is comparatively better than DL approaches.

In future IDS solutions, adversarial learning and explanation stability should be included in the framework development stage rather than at the end stage.

Conclusion and Future Work:

This work used the CICIoT2023 dataset to investigate explainable intrusion detection systems' (XAI-IDS) susceptibility to adversarial evasion efforts with an emphasis on DDoS detection in IoT environments. It was shown that even small adversarial perturbations can dramatically damage model predictions and the explanations that go along with them through a series of systematic tests employing Random Forest and LSTM models in conjunction with LIME and SHAP frameworks for interpretability. The number of DDoS flows that were incorrectly identified as benign traffic increased significantly, with LSTM false negatives growing from 24,876 to 27,993 and RF false negatives rising from 363 to 12,195, even though overall classification accuracy only marginally decreased. These findings quantify the threat posed by adversarial assaults in real-world IoT intrusion detection scenarios.

LIME and SHAP explanations for hostile samples that were incorrectly categorized were deceptively cohesive, often downplaying significant indications while making major contributions to irrelevant characteristics, according to tests. The top 5 feature overlap for LIME decreased to 45% for RF and 50% for LSTM, while the cosine similarity between clean and adversarial SHAP vectors decreased to 0.55 and 0.35 for RF and LSTM, respectively. These findings demonstrate that current XAI algorithms, while effective in most scenarios, they may provide false explanations when attack manipulations are present, endangering operator trust and operational judgment in real-world applications.

These findings show how vital it is to develop adversarial robust explainable IDS methods. Future research directions include creating hybrid models that combine resilience against adversarial inputs with interpretable decision-making, examining defense strategies like adversarial training created specifically for Internet of Things traffic, and creating quantitative metrics for interpretability robustness that can be incorporated into IDS evaluation pipelines. Furthermore, a more comprehensive understanding of XAI susceptibility across various cyber threat environments will be obtained by expanding our investigation to attack types other than DDoS, such as low-rate and covert attacks.

In conclusion, this study provides a critical viewpoint on the shortcomings of existing explainable IDS frameworks and lays the groundwork for developing next-generation, adversarial aware, and reliable AI-driven cybersecurity solutions for the Internet of Things.

References:

[1] M. F. Saiyed, I. Al-Anbagi, and M. Shamim Hossain, “An Explainable Deep Learning

System for Cyberattack Detection in Internet of Energy Networks," *IEEE Netw.*, 2025, doi: 10.1109/MNET.2025.3622918.

[2] A. A. G. Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu, "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," *Sensors*, vol. 23, no. 13, p. 5941, 2023, doi: https://doi.org/10.3390/s23135941.

[3] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, May 2017, Accessed: Aug. 14, 2024. [Online]. Available: https://arxiv.org/abs/1705.07874v2

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.

[5] "(PDF) Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense." Accessed: May 09, 2026. [Online]. Available: https://www.researchgate.net/publication/367979039_Adversarial_Machine_Learning_Attacks_against_Intrusion_Detection_Systems_A_Survey_on_Strategies_and_Defense

[6] M. A. Ayub, W. A. Johnson, D. A. Talbert, and A. Siraj, "Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning," *2020 54th Annu. Conf. Inf. Sci. Syst. CISS 2020*, Mar. 2020, doi: 10.1109/CISS48834.2020.1570617116.

[7] Sheikh Abdul Wahab, Saira Sultana, "A Multi-Class Intrusion Detection System for DDoS Attacks in IoT Networks Using Deep Learning and Transformers," *Sensors*, vol. 25, no. 15, p. 4845, 2025, doi: https://doi.org/10.3390/s25154845.

[8] A. Khan, Y. Li, S. Shoukat, D. Javeed, and M. Adil, "Towards secure IoT-enabled transportation: an explainable AI and deep learning-based approach for efficient threat detection," *Clust. Comput. 2025 2811*, vol. 28, no. 11, pp. 699-, Sep. 2025, doi: 10.1007/S10586-025-05473-Z.

[9] Jason Moss, Jeremy Gordon, "Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation," *Electronics*, vol. 14, no. 23, p. 4622, 2025, doi: https://doi.org/10.3390/electronics14234622.

[10] Vincent Zibi Mohale, Ibidun Christiana Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," *Front. Artif. Intell.*, vol. 8, 2025, doi: https://doi.org/10.3389/frai.2025.1526221.

[11] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," *Proc. IECON 2018 - 44th Annu. Conf. IEEE Ind. Electron. Soc.*, pp. 3237–3243, Dec. 2018, doi: 10.1109/IECON.2018.8591457.

[12] Maraz Mia, Mir Mehedi A. Pritom, "Explainable but Vulnerable: Adversarial Attacks on XAI Explanation in Cybersecurity Applications," *arXiv:2510.03623*, 2025, [Online]. Available: https://arxiv.org/abs/2510.03623

[13] Jon Vadillo, Roberto Santana, Jose A. Lozano, "Adversarial Attacks in Explainable Machine Learning: A Survey of Threats Against Models and Humans," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2024, [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1567

[14] Tharindu Lakshan Yasarathna, Nhien-An Le-Khac, "SoK: Systematic analysis of adversarial threats against deep learning approaches for autonomous anomaly detection systems in SDN-IoT networks," *arXiv:2509.26350*, 2025, [Online]. Available: https://arxiv.org/abs/2509.26350