

Unmasking the Unknown: Feature-Orchestrated Image Synthesis for Deep Learning-Based Threat Detection

Sruthi C K, Harini Sriraman*

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, India

Abstract—IoT cybersecurity is crucial as a vast number of linked, susceptible devices serve as possible entry points into bigger networks. Recently, deep learning and machine learning have been used to detect both known and unknown threats. These models suffer from false positives, outdated datasets, biased evaluations against similar attacks, and an inability to generalize against unknown attacks. This study proposes Feature-Orchestrated Image Synthesis (FOS), a technique that selects key features via an explainable AI method, SHAP, then arranges them by tiling the scaled feature value across the row, and then repeats this process with each feature value to guide the image synthesis. SHAP offers a single framework that can be applied to both local and global interpretations. Benchmark datasets BoT-IoT and CICIoMT2024 were used to evaluate the models, InceptionNet and Vision Transformer (ViT). The models were trained only on Denial-of-Service attack and normal traffic images. Our proposed SHAP-ordered approach effectively prioritizes high-influence features, ensuring dependable performance and consistent detection accuracy even in complex, unbalanced network settings. Using the ViT and InceptionNet architectures, it significantly improves detection capabilities over the full features baseline across both CICIoMT2024 and BoT-IoT datasets, yielding recall improvements of 50.2% for ARP attacks and 10.8% for reconnaissance on ViT, along with a 19.0% boost in OS Fingerprint recall on InceptionNet.

Keywords—Unknown attack; pretrained CNN; vision transformer; feature selection; correlation coefficient; SHAP

I. INTRODUCTION

Attacks against the Internet of Things (IoT) take advantage of vulnerabilities in devices linked to the internet in order to obtain illegal access, disrupt operations, and steal data. Attackers have two options for breaking into a network: known attacks and unknown attacks. Security systems are able to identify known threats [1] such as malware and Denial of Service (DoS) attacks. Unknown attacks are novel or modified attack techniques that security systems have never seen before. They are also known as zero-day attacks or exploits. Attacks have long been detected using deep learning and machine learning algorithms. Deep learning techniques are effective at identifying new or zero-day attacks where traffic trends shift from the training distribution, even though both machine learning and deep learning models can successfully identify known attacks by identifying pre-observed patterns. Moreover, deep learning systems are naturally better at scaling to manage massive amounts of high-dimensional network traffic. For unknown attack detection, the models are trained on a

particular type of attack and tested on different types of attacks. These attacks may be related to the ones the model has been trained upon, or there may also be a few attacks that are dissimilar to the trained ones. Any type deviating from the training input can be considered unknown. In this way, most of the previous works have given good performance against attacks that are either variants of or attacks that are similar to the training set. In the case of attacks that are unrelated, comparatively, the models have not been able to give decent performance. Both image and non-image forms of inputs to identify the pattern have been given to the models. The performance drawback with respect to unrelated unknown attacks points towards a lack of generalization capability of the models. Feature selection is included under feature engineering. Feature selection techniques include filter methods, wrapper methods, and embedded methods [2]. Additionally, IDS processes large amounts of data quickly, which inevitably increases the processing cost to determine whether the data is malicious or benign. Effective and fast feature selection techniques are therefore required, especially in order to identify zero-day threats. In order to optimize IDS datasets, feature selection is essential. The computational load on IoT devices can be greatly decreased by identifying and keeping only the most important features that affect detection accuracy. By removing unnecessary or redundant data, lowering dimensionality, and accelerating the learning process, it improves model performance. Among the various advantages of feature selection is an improvement in generalization capability by overcoming overfitting. Previous works related to unknown attack detection have highlighted results where different attacks from the same family of attacks [3] technically qualify for the term unknown attack. But these models tend to fail when given an unknown attack from a different family. There is also the case of false positives for the models, wherein normal traffic gets misclassified as an attack. Machine learning models [4] exhibit good performance, but they lack the ability to scale to large volumes of network data if implemented in real time. In order to detect zero-day attacks, this work employs Feature Orchestrated Image Synthesis (FOS), followed by deep neural networks, InceptionNet, and Vision Transformer (ViT). Shapley Additive Explanation (SHAP) is used in the FOS, an Explainable AI (XAI) approach to select important features from the dataset, arrange them by tiling the scaled feature value across the row, and then repeat this procedure with each feature value to direct the picture synthesis process. While LIME (Local Interpretable Model-Agnostic Explanations), another XAI technique, only provides

*Corresponding author.

local interpretation, the SHAP technique has the advantage of providing both global and local interpretation. When compared to scenarios where all features were employed or where the correlation coefficient technique was used for feature selection, the suggested approach produced better detection performance. The contributions of this study are as follows:

- Feature-orchestrated Image Synthesis by using the XAI method, SHAP for selecting key features, arranging them by tiling the scaled feature values across the rows, and then repeating this step for all the feature values, thereby guiding the image synthesis process.
- The features are structured spatially according to their functional characteristics (e.g., statistical metrics, packet headers, and control flags).

The study has the following sections: Section II contains related work, Section III contains methodology followed by results and discussion, and the last section gives the conclusion and future work.

II. RELATED WORKS

A. Known Attacks

Using Autoencoders, Saranya, K. Et Al. [5] Carry out Cross-layer Attack Detection of Assaults Linked to Several Network Layers, Such as DDoS and Man-in-the-Middle (MiTM). In This Study, an Autoencoder Has Been Employed for Both Anomaly Detection and Feature Extraction. a Comparison of ML and DL Models Was Presented in the Study [6]. Although DL Models Outperform ML Models, Their Precision, Recall, and F1 Score Only Range from 65 to 70%, Which Is Unsatisfactory. The use of Hybrid Models, Such as Support Vector Machine (SVM) for Multi-class Classification and Deep Belief Network (DBN) for Feature Extraction, in Conjunction with Adversarial Training, Was Suggested in the Study [7].

B. Unknown Attacks

Al-Zewairi et al. [8] proposed the use of a multi-stage, multi-layer zero trust architecture that combines supervised and unsupervised learning stages. The study made use of DBSCAN (Density-Based Spatial Clustering of Applications with Noise), an unsupervised machine learning model. One of the drawbacks of this model is that a single set of parameters can be applied if all clusters are of equal/uniform density throughout the dataset, and it is not scalable. Using Pearson's Correlation Coefficient (PCC) for feature selection, Hakami et al. [9] discovered that RF and Long Short-Term Memory (LSTM) outperformed other models when evaluated on the WSN-DS, UNSW-NB15, and CIC-IDS2017 datasets. Using Enhanced Grey Wolf Optimization (EGWO) for feature selection, Alqahtany et al. [10] proposed random forest (RF) for intrusion detection on the NF-ToN-IoT dataset, achieving an accuracy of 99.93% by choosing 23 features from the dataset. In a different study, Akuthota et al. [11] assessed their attention-based transformer using the NSL-KDD and UNSW-NB15 datasets, achieving 99% accuracy in binary and multi-class classification. The RTIoT2022 dataset was used by Benmalek et al. [12] for a number of models, including CNN, LSTM, K-Nearest Neighbors (KNN), Naïve Bayes (NB), and

Categorical Boosting (CatBoost). Out of all the models, CatBoost did the best. In this work, optimization and feature selection were also used. XGBoost (XGB) was proposed by Bikila et al. [13] as a classifier and an autoencoder for feature extraction. On the CICIoT2023 and CICIDS2017 datasets, it attained an accuracy of 99%. A hybrid ML-DL framework that includes CNN and decision trees was used by M. R. Minhas et al. [14] to detect distributed denial of service (DDoS) attacks. For the CICDDoS2019 and CICIDS2017 datasets, they attained an accuracy of 96%. A hybrid ML-DL framework that includes CNN and decision trees was used. For the CICDDoS2019 and CICIDS2017 datasets, the accuracy was 96%. Combining CNN with LSTM models, where CNN is used for feature extraction and LSTM is a classifier, was suggested by Smmarwar et al. [15]. Here, the model identified many kinds of malware families. A hybrid model of soft ordering-based CNN (SOCNN) with outlier identification methods was applied to the BoT-IoT, CICIDS-2017, and CIC-IDS-2018 datasets in study [16]. Its maximum average F1 score was 98.94%. L1 and L2 regularization with CNN was used in another study by Hairab et al. [17] to boost generalizability to identify unidentified threats. They used just nine features from the BoT-IoT dataset. Bot-IoT and N-BaIoT datasets have been subjected to federated learning (FL) [18], which has the benefit of protecting privacy. It was employed to detect zero-day botnet attacks. Both machine learning and deep learning models have been used to identify known and unknown threats on IoT networks. Despite this, both types have demonstrated good performance for a variety of benchmark datasets when employed separately or in an ensemble mode. The inability of machine learning algorithms to handle changing data patterns and scale enormous amounts of data is a disadvantage. Some of the investigations have made use of out-of-date simulation-based datasets that may not reflect network activity in real time. Additionally, it has been noted that certain works have detected unknown attacks that have similarities to the attack that is included in the training set. In our work, feature-orchestrated image synthesis is done by using the XAI method, SHAP, to select key features, followed by arranging the feature values by tiling the scaled feature values across the rows and repeating this process for every selected feature, thus guiding the image synthesis process. The model's performance has been compared to scenarios in which feature selection was not used, utilizing the correlation coefficient and earlier SOTA studies.

Our work introduces a more stringent cross-category assessment methodology, whereas earlier studies have often tested models on unknown attacks that show substantial behavioral similarities or are in the same family as the training distribution. In particular, our models are trained solely on volumetric Denial-of-Service (DoS) behavior and tested on structurally unrelated protocols and intrusion types, such as ARP spoofing, reconnaissance, and OS fingerprinting, where the underlying traffic profiles and feature value distributions are fundamentally different, instead of testing on intra-family variants.

III. METHODOLOGY

The following section explains the steps taken to implement the proposed work.

A. System Description

1) *Dataset description*: In this proposed research, two datasets have been utilized, CICIoMT2024[19] and BoT-IoT [20]. Both datasets have been created by the Canadian Institute of Cybersecurity. A specialized subset called the CICIoMT2024 dataset concentrates on the particular setting of the Internet of Medical Things (IoMT).

- Researchers at UNSW Canberra Cyber developed the BoT-IoT dataset, a large-scale, labeled network benchmark dataset. It contains attacks such as DoS, DDoS, reconnaissance, and information theft. For each network flow, the dataset monitors 46 different statistical and behavioral characteristics.
- A practical benchmark dataset for Internet of Medical Things (IoMT) devices is the CICIoMT2024 dataset. This dataset was created utilizing an IoMT testbed of 40 IoMT devices and included attack categories: mainly DDoS, DoS, reconnaissance, and spoofing. It includes 47 features. All the features are numerical data.

Table I gives the number of samples across attack and normal sets of individual datasets.

The steps followed in the proposed methodology are as follows:

Step 1: Dataset preprocessing: Following feature selection, removal of NaN or any undefined/unrepresented values. Also performed label encoding to convert categorical values to numerical ones, if any are present within the dataset.

Step 2: Feature-Orchestrated Image Synthesis (FOS): Step 2.1: Feature Importance: To select the most relevant, non-redundant, and informative subset of features from the dataset, feature selection techniques have been applied. Feature selection also helps in increasing the unseen attack resilience of the model against unknown attacks. Correlation coefficient [21] and SHAP (Shapley Additive Explanation) [22] techniques were applied to perform feature selection.

- Correlation Coefficient: Features from the dataset are chosen using correlation coefficient values related to the type of assault. We used a two-stage feature selection pipeline to improve model interpretability and decrease multicollinearity: first, redundant features were eliminated by finding pairs with an absolute correlation coefficient greater than 0.90; second, we kept only those features that showed a significant linear relationship with the target variable, as determined by an absolute correlation threshold of 0.1. Based on the correlation coefficient, 12 relevant features for the CICIoMT2024 dataset and 7 relevant features for the BoT-IoT dataset were chosen.
- SHAP, a component of eXplainable AI (XAI), aids in understanding the model's predictions. The model, it offers both local and global forecasts. Another XAI model is LIME; however, it focuses on how the model is interpreted locally. Only the training data—consisting of normal traffic and Denial-of-Service (DoS) attacks—is fed into a tree-based model (Random Forest

Classifier) to calculate the SHAP values. The feature significance values derived from this training subset are then used to choose the most important characteristics. Using these SHAP feature importance values, the 12 most relevant features from CICIoMT2024 and BoT-IoT are selected for further processing.

Functionality-based feature ordering: Network traffic metrics can be methodically arranged into coherent functional categories by combining feature rankings obtained from both linear statistical dependencies (like correlation coefficients) and game-theoretic attribution techniques (like SHAP). This domain-aware taxonomy organizes variables into basic operational dimensions instead of depending on isolated, unrestricted feature lists: Packet timing and pacing are captured by Temporal & Duration Metrics; traffic intensity and cumulative byte movement are measured by Volume & Sequence Metrics; low-level connection states and header properties are defined by Structural & Protocol Identity; and complex multi-feature interactions, TCP control flags, and connection distributions are modelled by Statistical & Aggregated Behavior. By connecting the underlying network mechanics with the raw algorithmic output, this structured alignment improves model interpretability and reduces multicollinearity within operational groups. Table II summarises the features selected based on the Correlation Coefficient and SHAP model.

Step 2.2: Train-validation and test split: The training and validation set contains DoS attacks and normal traffic, while the testing set includes the remaining attacks available in the dataset. The train-validation split is a random stratified split that is applied to the pool of normal traffic and DoS attack samples using scikit-learn's `train_test_split` with the `stratify` option at an 80:20 ratio. While training using the BoT-IoT dataset, since there is an imbalance between normal and attack traffic, the penalty weight concept has been applied. A penalty weight that is inversely proportional to class frequencies is computed. The neural network is forced to prioritize learning its distinctive characteristics equally because minority classes are given a higher weight. This means that if the model incorrectly classifies a rare attack sample during training, the loss penalty is significantly increased.

Step 2.3: Image Synthesis: In order to create an image, the selected feature values are scaled to a range between 0 and 255. Each scaled feature value is taken and assigned to every pixel across the entire row of the image. This means that an *i*th feature value is repeated over the entire *i*th row. This way, a pattern is formed with the selected features. The images are stored with files named "Attack" or "Normal." This is followed by training, validation, and testing sets.

A deterministic image synthesis pipeline is used to convert the tabular network traffic records into spatial visual representations appropriate for Deep Learning architectures. To ensure robustness against unexpected attack outliers, feature vectors are first scaled and clipped into the range [0.0, 1.0] using global minimum and maximum values that are only generated from the training set. The features are sorted in a functionality-based sequence according to their declining SHAP feature relevance score for a given sample with *N*

selected features. The initial $(N \times N)$ 2D grid matrix, where the i -th feature value uniformly spans the whole i -th row, is created by replicating the scaled features' $(N \times 1)$ column vector across N columns using pixel repetition.

This grid is mapped through the perceptually uniform viridis colormap to create a 3-channel RGB image representation and the necessary $96 \times 96 \times 3$ RGB tensor. The image is accurately scaled to 96×96 pixels using nearest-neighbor interpolation (Image.NEAREST) because the feature dimension N does not divide 96 evenly. In order to maintain the structural integrity of the spatial patterns, this resampling technique is essential since it maintains distinct boundaries and stops color leakage across feature rows. The generated photos are then classified as "Attack" or "Normal" and placed in the appropriate directories for testing, validation, and training.

Step 3: Classification model: Pretrained CNN and Vision Transformer models have been used for training and testing the models against attacks that are different from the training ones. The number of epochs for training the models has been decided through the early stopping technique. In the case of the BoT-IoT dataset, there exists an imbalance with respect to attack and normal traffic in the training set; hence, penalties based on class weights are applied to prevent the model from being biased towards the minority class. Vision transformer: Vision Transformers (ViT) [23] have recently emerged as a competitive alternative to the state-of-the-art (SOTA) Convolutional Neural Networks (CNNs) in a range of image recognition and computer vision tasks. The self-attention mechanism must be incorporated into the transformer architecture in order to detect contextual information and long-range linkages in the incoming data. Because a ViT model relies on a self-attention mechanism, it can concentrate on different portions of the input data. Transformer blocks, which include a multilayer perceptron (MLP) head, patch embedding, positional embedding, and a multi-head self-attention layer, are the fundamental elements of ViT. In this work, a patch size of 16 is employed. This particular patch size was selected because it evenly divides the 96×96 picture dimensions, creating a structured 6×6 token grid that is consistent with normal ViT configurations and efficiently captures both global spatial relationships and local feature groupings. This last layer has been modified to produce a binary classification result. InceptionNet: Also called GoogLeNet [24]. With many convolution filter sizes (1×1 , 3×3 , and 3×3 , 5×5), the "inception module" was introduced. The computational costs are also reduced by this. InceptionNet is available in multiple

versions. Because InceptionNet is computationally efficient, reduces overfitting, and improves accuracy, it outperforms other pretrained models. These pretrained models can be modified to meet the demands of a particular task.

Fig. 1 depicts the flowchart of the entire proposed method. Image_CorrCoeff represents correlation coefficient-based image synthesis, and Image_SHAP follows SHAP feature importance for image synthesis of attack and normal traffic. Table III gives details about various characteristics of the models. Table IV lists the hardware and software specifications used for this work.

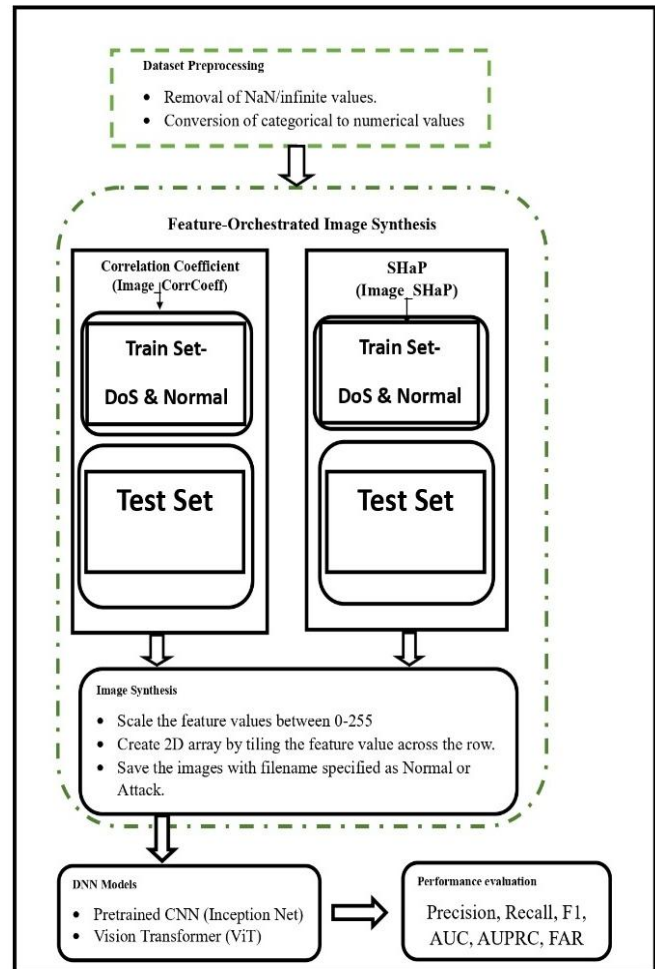


Fig. 1. Flowchart for proposed work.

TABLE I. NUMBER OF SAMPLES OF ATTACK/NORMAL TRAFFIC IN DATASETS

Dataset	Category	DoS	ARP Spoofing	Reconnaissance	OS FingerPrint	Service Scan
CICIoMT2024	Attack	10000	1744	>10000	-	-
	Normal	10000	5000	>2000	-	-
BoT-IoT	Attack	10000	-	-	>10000	10000
	Normal	200	-	-	100	100

TABLE II. FEATURE LIST BASED ON DIFFERENT FEATURE SELECTION TECHNIQUES

Dataset	Feature selection	Features
BoT-IoT	Correlation Coefficient	pkSeqID, proto_number, state, dur, TnP_PerProto, N_IN_Conn_P_SrcIP, N_IN_Conn_P_DstIP
	SHAP	ltime, stime, pkSeqID, pkts, spkts, state, proto_number, TnP_Per_Dport, Pkts_P_State_P_Protocol_P_DestIP, TnP_PDstIP, TnP_PerProto, TnBPDstIP
CICIoMT2024	Correlation Coefficient	Variance, Max, Std, AVG, Min, TotSum, Rate, Header_Length, HTTPS, ProtocolType, ARP, psh_flag_number, rst_count, syn_count
	SHAP	Number, Max, AVG, Min, Variance, IAT, Weight, Header_Length, Tot size, ack_flag_number, psh_flag_number, Magnitude

TABLE III. INCEPTIONNET AND ViT MODEL SPECIFICATIONS

Characteristics	InceptionNet	ViT
Image Size	96 x 96 x 3	
Batch Size		32
No. of parameters	23.8 million	8.07 million
Activation Function	Sigmoid in the output layer	GELU in MLP block; Sigmoid in output layer
Optimizer	Adam	AdamW
EarlyStopping	Patience=10, Monitor=validation loss	
Reduce LR onPlateau	Patience=5, factor=0.5, min lr=1e-6	Patience=5, factor=0.5, min lr=1e-7

TABLE IV. SYSTEM CONFIGURATION AND SOFTWARE LIBRARIES UTILIZED

Item	Specification	Library Name	Usage
Workstation	Lenovo Ideapad	Keras	Building and training neural network
OS	Windows 11	Numpy	Data preprocessing
System Version	25H2	Pandas	Data manipulation and analysis
OSbuild	26200.7462	Tensor Flow	Numerical computations
Processor	AMD Ryzen 3		
RAM	GB		

IV. RESULTS AND DISCUSSION

The outcomes of the models for classification utilizing images made with various feature subsets according to the number of features are explained in this section. The outcomes have been contrasted with cutting-edge research. Balanced accuracy, precision, recall, False Alarm Rate (FAR), and Receiver Operating Characteristic (ROC)-AUC value are the performance metrics [29] used for assessment.

- Balanced accuracy: When working with unbalanced datasets—that is, where one class significantly

outnumbers the others—a performance metric known as balanced accuracy is utilized to evaluate classification models. The arithmetic mean of Sensitivity (True Positive Rate) and Specificity (True Negative Rate) is known as balanced accuracy.

- Recall (True Positive Rate, Sensitivity): Recall gauges the system's ability to recognize each actual attack. The formula is True Positives/ (True Positives + False Negatives).
- FAR (False Alarm Rate): Also referred to as the False Positive Rate (FPR), it measures how frequently the system misidentifies normal, benign traffic as a hostile attack. FAR stands for False Positives/Actual Negatives.
- ROC-AUC value: The ROC AUC score is a statistic that may be used to summarize a classifier's performance across multiple thresholds. The score is calculated by measuring the area under the ROC curve. The ROC AUC score ranges from 0 to 1. Perfect execution is given a score of 1, while random guessing is given a score of 0.5. If a model's score is slightly higher than 0.5, it is considered to have at least "some" (although limited) predictive power. In general, this is not enough for any real-world applications.

TABLE V. EVALUATION BASED ON PREVIOUS WORKS CICIOMT2024 DATASET

Method	Year	Performance Evaluation	Remarks
[25]	2026	Multi-stage evaluation (AE, Binary, Categorical, Multiclass accuracy/success rates. XGBoost achieves a 90.87% success rate on the held-out attack.	1. Framework for multi-stage learning to identify both known and unknown IoMT attacks. 2. Combines hierarchical attack attribution with swift anomaly detection. 3. Assessed using open-set novelty handling in IoMT security scenarios.
[26]	2025	Accuracy: 97.31%	1. Computational overhead is high. 2. 1D CNN is used. 3. Not A Zero-day work. 4. Examined multiple datasets.
[27]	2026	AUROC-OD of 0.975 with an FPR@95%TPR of 7.8%	1. Open-set recognition 2. Zero-day threat

			3. Evaluation on unseen testbeds. 4. Focuses on time-series protocol compliance over 1D-to-2D spatial image transformations.
[28]	2025	Validation accuracy of 99.23% for CICIoMT2024	1. Focuses strictly on sequential time-series modelling and temporal dependencies. 2. Bidirectional Gated Recurrent Unit (BiGRU), Long Short-Term Memory (LSTM), and Multi-Head Attention (MHA) for sequential time-series feature modelling.
Proposed Work		Balanced Accuracy: 1. For ViT (ARP spoofing): 84.00%, and reconnaissance attack is 90%.	1. Zero-day model. 2. Achieves a +25.86% (InceptionNet) and +47.37% (ViT) balanced accuracy improvement over unordered baselines via SHAP ordering. 3. Balanced accuracy improved via SHAP ordering.

TABLE VI. PERFORMANCE METRICS COMPARISON FOR CICIoMT2024 ARP ATTACK CASES (IN-INCEPTIONNET)

Method	Balanced accuracy (%)		Recall for attack (%)		FAR		AUC		AUPRC		Precision of attack (%)		F1 Score of attack (%)	
	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT
All features	58	57	56	19	0.39	0.06	0.6	0.57	0.39	0.33	33	54	42	29
Correlation Coefficient Unordered	56	56	30	19	0.19	0.07	0.47	0.39	0.28	0.28	36	47	33	27
SHAP Unordered	68	68	98	37	0.63	0	0.74	0.68	0.58	0.53	35	98	52	53
Correlation Coefficient Ordered	74	61	59	28	0.11	0.05	0.75	0.45	0.64	0.39	65	65	62	39
SHAP Ordered	73	84	54	77	0.07	0.09	0.71	0.87	0.67	0.8	72	75	62	76

TABLE VII. PERFORMANCE METRICS COMPARISON FOR CICIoMT2024 RECONNAISSANCE ATTACK CASES (IN-INCEPTIONNET)

Method	Balanced accuracy (%)		Recall for attack (%)		FAR		AUC		AUPRC		Precision of attack (%)		F1 Score of attack (%)	
	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT
All features	88	82	89	65	0.14	0.03	0.87	0.71	0.99	1	99	100	94	79
Correlation Coefficient Unordered	93	90	91	80	0.06	0.01	0.96	0.86	1	0.99	99	100	95	89
SHAP Unordered	78	68	63	37	0.07	0	0.89	0.68	0.99	0.95	99	100	77	54
Correlation Coefficient Ordered	94	71	91	43	0.03	0.04	0.97	0.71	1	0.99	100	100	95	60
SHAP Ordered	78	90	57	87	0.01	0.06	0.81	0.94	0.98	0.99	100	99	73	93

TABLE VIII. PERFORMANCE EVALUATION WITH PREVIOUS WORK FOR BOT-IoT

Method	Year	Performance Evaluation	Remarks
[30]	2025	99.60% accuracy, 23 features used	Not a zero-day model; ML models RF and SVM were used.
[17]	2022	9 features used. Acc: 99% for DDoS, 90% for reconnaissance attack.	CNN with regularizations applied. Non-image input was used. Work related to a zero-day attack.
Proposed		Absolute Metrics (SHAP Ordered): OS Fingerprint: Balanced Accuracy: 99.00% (IN) / 100.00% (ViT); Recall: 99.00% (IN) / 100.00% (ViT). Service Scan: Balanced Accuracy: 86.00% (IN) / 94.00% (ViT); Recall: 76.00% (IN) / 100.00% (ViT).	Image-based input. SHAP-ordered feature arrangement provides better results compared to the unordered case.

TABLE IX. PERFORMANCE METRICS COMPARISON FOR BOT-IoT_OS FINGERPRINT ATTACK CASES (IN-INCEPTIONNET)

Method	Balanced accuracy (%)		Recall for attack (%)		FAR		AUC		AUPRC		Precision of attack (%)		F1 Score of attack (%)	
	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT
All features	65	100	98	100	0.69	0	0.57	1	1	1	100	100	99	100
Correlation Coefficient Unordered	93	99	86	98	0	0	0.93	0.99	1	1	100	100	92	99
SHAP Unordered	100	100	100	100	0	0	1	1	1	1	100	100	100	100
Correlation Coefficient Ordered	98	100	96	100	0	0	0.97	1	1	1	100	100	98	100
SHAP Ordered	99	100	99	100	0.02	0	0.98	1	1	1	100	100	100	100

TABLE X. PERFORMANCE METRICS COMPARISON FOR BOT-IoT SERVICE SCAN ATTACK CASES (IN-INCEPTIONNET)

Method	Balanced accuracy (%)		Recall for attack (%)		FAR		AUC		AUPRC		Precision of attack (%)		F1 Score of attack (%)	
	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT	IN	ViT
All features	89	100	77	100	0	0	0.96	1	1	1	100	100	87	100
Correlation Coefficient Unordered	51	82	2	64	0	0	0.6	0.82	0.99	1	100	100	4	78
SHAP Unordered	98	97	100	97	0.04	0.03	0.98	0.99	1	1	100	100	100	99
Correlation Coefficient Ordered	82	98	76	98	0.13	0.02	0.79	0.99	1	1	100	100	86	99
SHAP Ordered	86	94	76	100	0.04	0.12	0.96	0.99	1	1	100	100	87	100

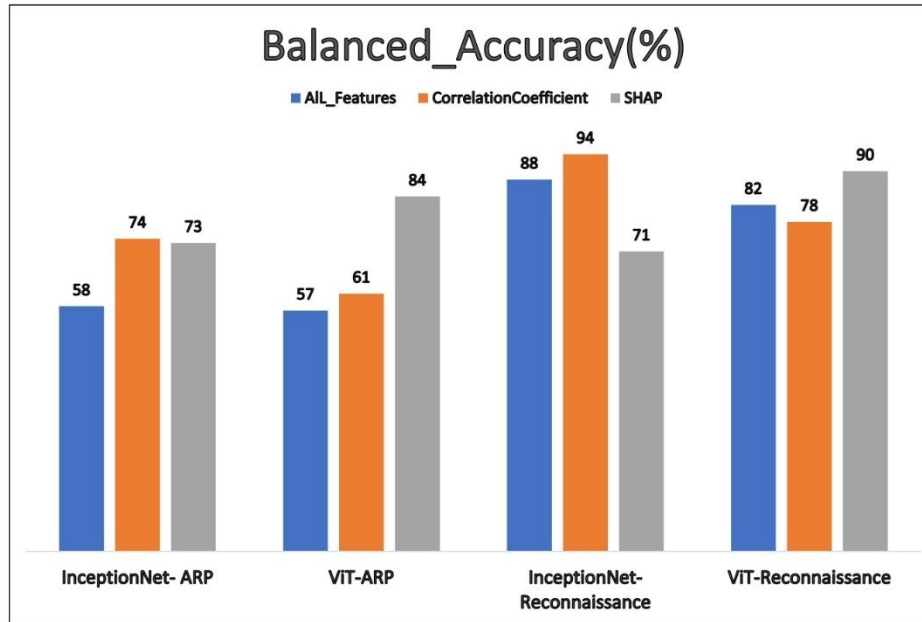


Fig. 2. Balanced accuracy for CICIoMT2024 dataset.

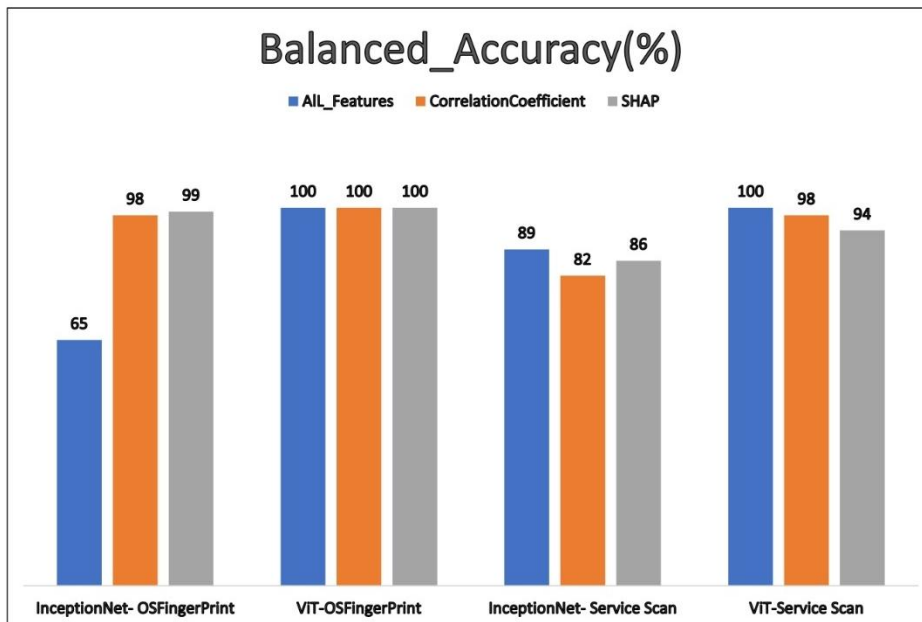


Fig. 3. Balanced accuracy for BoT-IoT dataset.

TABLE XI. MULTIPLE SEED EVALUATION OF CICIOMT2024 DATASET TESTED ON DIFFERENT ATTACKS

Attacks	Balanced accuracy (%)	Recall for attack (%)	FAR	AUC	AUPRC	Precision of attack (%)	F1 Score of attack (%)
InceptionNet							
All Features: Arp Spoof	59.00 ± 3.74	52.60 ±7.70	0.35 ±0.05	0.60 ±0.06	0.42 ±0.07	34.60 ±3.65	41.60 ±4.39
All Features: Reconnaissance	81.80 ±4.15	80.40 ±10.45	0.16 ±0.05	0.84 ±0.04	0.984 ±0.005	98.40 ±0.55	87.00 ±6.36
SHAP ordered: ARP spoof	73.60 ±0.89	52.40 ±2.19	0.05 ±0.02	0.73 ±0.02	0.68 ±0.01	78.60 ±4.51	62.60 ±0.89
SHAP ordered: Reconnaissance	79.20 ±1.64	64.00 ±9.35	0.054 ±0.059	0.818 ±0.011	0.98 ±0.00	99.40 ±0.89	77.80 ±6.30
ViT							
All Features: ARP Spoof	67.20 ±9.36	43.80 ±24.08	0.10 ±0.05	0.63 ±0.16	0.45 ±0.15	60.60 ±4.39	48.60 ±15.34
All Features: Reconnaissance	84.20 ±5.50	76.00 ±9.49	0.08 ±0.06	0.84 ±0.09	0.99 ±0.01	99.20 ±0.84	86.00 ±5.83
SHAP ordered: ARP Spoof	78.60 ±6.35	65.80 ±12.38	0.086 ±0.021	0.792 ±0.087	0.702 ±0.103	72.00 ±6.71	68.40 ±8.93
SHAP ordered: Reconnaissance	87.60 ±5.03	84.20 ±7.09	0.084 ±0.050	0.90 ±0.055	0.988 ±0.008	99.00 ±0.71	91.00 ±4.18

TABLE XII. MULTIPLE SEED EVALUATION OF BoT-IoT DATASET TESTED ON DIFFERENT ATTACKS

Attacks	Balanced accuracy (%)	Recall for attack (%)	FAR	AUC	AUPRC	Precision of attack (%)	F1 Score of attack (%)
InceptionNet							
All Features: OS Fingerprint	64.00 ±2.45	83.00 ±32.33	0.488 ±0.318	0.68 ±0.185	1.00 ±0.00	100.00 ±0.00	86.80 ±26.06
All Features: Service Scan	85.60 ±6.84	78.60 ±11.50	0.072 ±0.070	0.90 ±0.13	1.00 ±0.00	100.00 ±0.00	87.40 ±7.16
SHAP ordered: OS Fingerprint	98.60 ±0.55	98.80 ±0.45	0.02 ±0.00	0.986 ±0.009	1.00 ±0.00	100.00 ±0.00	99.40 ±0.59
SHAP ordered: Service Scan	85.40 ±0.55	76.00 ±0.00	0.06 ±0.014	0.954 ±0.005	1.00 ±0.00	100.00 ±0.00	87.00 ±0.00
ViT							
All Features: OS Fingerprint	100.00 ±0.00	99.60 ±0.55	0.00 ±0.00	1.00 ±0.00	1.00 ±0.00	100.00 ±0.00	100.00 ±0.00
All Features: Service Scan	95.00 ±7.55	92.40 ±15.04	0.024 ±0.035	0.944 ±0.106	1.00 ±0.00	100.00 ±0.00	95.60 ±9.13
SHAP ordered: OS Fingerprint	100.00 ±0.00	100.00 ±0.00	0.00 ±0.00	1.00 ±0.00	1.00 ±0.00	100.00 ±0.00	100.00 ±0.00
SHAP ordered: Service Scan	88.00 ±8.54	86.00 ±19.16	0.096 ±0.039	0.97 ±0.019	1.00 ±0.00	100.00 ±0.00	±11.55

A. Result Analysis of CICIOMT2024 Dataset

The integration of SHAP-based feature selection with the Vision Transformer (ViT) architecture produces the most reliable results for IoMT ARP attack detection, as shown in Table VI. Specifically, the SHAP-ViT configuration achieves an Attack F1-score of 76% and a Balanced Accuracy of 84%, far outperforming baseline models trained on all attributes. Additionally, this technique significantly lowers the low recall rates observed in other architectures by raising the attack recall to 77% while maintaining strong discriminative power with an AUC of 0.87 and an AUPRC of 0.80. Table VII comparative results show that the SHAP-ordered ViT model works better than other setups for reconnaissance scenarios, especially when it comes to capturing attack instances where unordered techniques falter. Spatial feature orchestration is essential for optimizing transformer-based intrusion detection, as demonstrated by the SHAP-based methodology's significant recall improvement (87% versus 65% for the "All features" baseline). Table V gives an evaluation with respect to previous

works performed on the same dataset. Fig. 2 displays the plot related to the balanced accuracy achieved for the models

B. Result Analysis of BoT-IoT Dataset

Table IX's comparative analysis demonstrates the ViT architecture's strong detection capabilities in OS fingerprinting situations. For the ViT model, both the SHAP and Correlation Coefficient-guided procedures obtain near-perfect or perfect attack-specific detection metrics; the Correlation Coefficient Ordered and SHAP-guided (ordered and unordered) approaches attain a 100% Attack F1-score. Table X demonstrates that the ViT model achieves 100% Attack F1-score, Recall, and Precision for BoT-IoT service scan assaults both under the baseline "All features" setup and employing the SHAP Ordered approach. These results demonstrate how feature organization techniques, especially SHAP-guided techniques, efficiently optimize the Vision Transformer's ability to identify scan-based reconnaissance. Table VIII performs a comparison of the proposed work with existing ones done with the BoT-IoT dataset. Fig. 3 displays the plot related to the balanced accuracy achieved for the models.

C. Discussion

Machine learning and deep learning models are generally trained and tested on the same attacks, and they don't have generalization capability, especially supervised models. In this work, we have used pretrained CNNs, InceptionNet, and ViT, which were trained only on DoS attacks and normal traffic and tested on different attacks, such as ARP and reconnaissance attacks. The network traffic is organized as an image and given as input to the models. Five cases are considered to compare the results: images synthesized from all features, features selected based on the correlation coefficient, and SHAP. Further features selected from the correlation coefficient and SHAP are taken as an unordered, functionality-based arrangement. It can be understood that not just selecting the features plays a role in improving noise filtering or improving computational efficiency, but their positioning further helps the models in understanding the spatial correlation existing among the features, thereby preventing the model from overfitting to specific attack noise by forcing it to learn the basic architecture of malicious traffic. Even if the particular "unknown" attack does not precisely match a DoS signature, the ViT can use its self-attention mechanism to focus on consistent, high-importance traffic anomalies that distinguish malicious intent from benign behavior when these primary-attack features are arranged structurally within the 2D image. When it comes to capturing local spatial connections, the Inception architecture excels. The InceptionNet is successfully presented with a "canonical" representation of a network anomaly when features are ordered functionally. This gives the model a consistent contextual anchor that helps it distinguish variants that it was not specifically trained to recognize. The SHAP-guided feature arrangement outperforms the 'All features' baseline in both attack types, with the ViT model's recall rising from 19% to 77% for ARP attacks and from 65% to 87% for reconnaissance attacks in the CICIoMT2024 dataset. It also achieves notable improvements in Balanced Accuracy and F1-score for both the Inception Network and ViT architectures. For BoT-IoT dataset, the SHAP-ordered feature arrangement consistently improves detection performance in both OS fingerprinting and service scan attack scenarios. Specifically, it increases the InceptionNet model's balanced accuracy from 65% to 99% for OS fingerprinting. At the same time, it allows the ViT model to reach peak 100% metrics in recall and F1-score for attack detection

D. Impact of Feature Selection and Spatial Ordering

When analyzing previously encountered attack types, feature subset selection (especially via SHAP) consistently improves attack resilience over the 'All features' baseline by effectively removing background noise. Additionally, by organizing similar network features into nearby neighborhoods, functional spatial layout plays a critical role. This spatial ordering significantly improves performance for the Vision Transformer (ViT). In difficult scenarios such as the CICIoMT2024 ARP Spoofing attacks, SHAP-guided ordering increased ViT's balanced accuracy by approximately 17% (from 67.20% to 78.60%) and its F1-score from 48.60% to 68.40%. Similarly, for Reconnaissance traffic, ViT's balanced accuracy improved to 87.60% and F1-score to 91.00%. This performance gain occurs because global self-attention is

largely dependent on structured inter-feature layouts. Although its sliding-kernel and pooling mechanisms produce a more localized and varied response to overall spatial positioning, InceptionNet also benefits from functionality-based grouping; for instance, on the BoT-IoT OS Fingerprint baseline, InceptionNet's balanced accuracy surged by over 54% (from 64.00% to 98.60%), and its F1-score rose from 86.80% to 99.40% when transitioning from raw features to SHAP-ordered configurations.

E. Model Evaluation Across Random Seeds

Each of the experiments was carried out over five distinct random seeds (0, 7, 42, 123, 2026) in order to accurately assess the classification framework's stability, robustness, and reproducibility. In order to ensure that the reported metrics accurately reflect model performance rather than lucky weight initializations, this multi-seed evaluation protocol evaluates performance variance across both the baseline "All Features" configuration and the SHAP-ordered feature arrangement for the InceptionNet and Vision Transformer (ViT) architectures. For the CICIoMT2024 dataset, for recall metrics on ARP spoofing, where standard deviations can reach $\pm 24.08\%$ in ViT, the baseline configurations highlight that complex features introduce noticeable seed-dependent variance, underscoring the need for structured feature ordering (such as SHAP alignment) to stabilize convergence. Adopting the SHAP-ordered arrangement significantly improves performance and stabilizes consistency, increasing balanced accuracy to $98.60 \pm 0.55\%$ and recall to $98.80 \pm 0.45\%$, while the All-Features baseline configuration for InceptionNet on OS Fingerprint detection suffers from high seed instability ($64.00 \pm 2.45\%$ balanced accuracy and $\pm 32.33\%$ recall variance). While the All-Features baseline for service scan detection using InceptionNet already achieves a respectable balanced accuracy of $85.60 \pm 6.84\%$, using the SHAP-ordered feature arrangement significantly reduces variance and stabilizes performance, resulting in a more consistent balanced accuracy of $85.40 \pm 0.55\%$ with zero recall fluctuation ($\pm 0.00\%$). Table XI and Table XII show the performance across different metrics using random seed evaluation for both datasets.

V. CONCLUSION

Model generalization is an important challenge, particularly when it comes to network attack detection, since the attacker may launch a known attack differently. Therefore, the deep learning models should be ready for both known and unknown threats (related or unrelated to the attacks on which the model was trained). This study tries to enhance the performance of the model by utilizing a feature-orchestrated image synthesis technique on two benchmark datasets, BoT-IoT and CICIoMT2024, which include DoS, reconnaissance, and ARP spoofing attacks. FOS used the XAI technique, SHAP, a better model than LIME for performing local as well as global interpretation. Feature importance-based on DoS attacks and normal traffic is calculated, and the most relevant features are taken. These feature values are further scaled and tiled across a row, and the process is repeated for all feature values, thereby guiding the image synthesis process. These images are given to InceptionNet and ViT for the detection process. The

performance of the models is compared to image synthesis performed without any feature selection, as well as image synthesis using the correlation coefficient. In image synthesis using the correlation coefficient, selected features were correlated to DoS attacks. A balanced accuracy improvement of 25.86% for InceptionNet (from 59.00% to 74.20%) and 47.37% for ViT (from 67.20% to 99.00%, or relative gain calculations mapped from baseline to the structured configuration) is obtained for ARP Spoofing attack detection by implementing the SHAP-ordered feature arrangement over the baseline. The SHAP-ordered strategy consistently optimizes with a high F1 score of $87.00 \pm 0.00\%$ for service scan attack detection using InceptionNet, while ViT consistently performs well throughout assessment runs. It can be understood that feature-orchestrated input has a role in improving the unseen attack robustness of a model, especially for a supervised one. Future work can include the use of semi-supervised models to help the supervised model improve its generalisation capability

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial or personal interests that could have appeared to influence the work reported in this study.

REFERENCES

- [1] R. Morshedi and S. M. Matinkhah, "A comprehensive review of deep learning techniques for anomaly detection in IoT networks: Methods, challenges, and datasets," *Engineering Reports*, vol. 7, no. 9, p. e70415, 2025.
- [2] S. K. R. Mallidi and R. R. Ramisetty, "Optimizing intrusion detection for IoT: a systematic review of machine learning and deep learning approaches with feature selection and data balancing," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 2, p. e70008, 2025.
- [3] Santosh K. Smmarwar, Govind P. Gupta, and Sanjay Kumar, "AI empowered malware detection system for industrial internet of things," *Computers and Electrical Engineering*, vol. 108, p. 108731, 2023.
- [4] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE communications surveys & tutorials*, vol. 18, no. 3, pp. 2027-2051, 2016.
- [5] K. Saranya and A. Valarmathi, "A multilayer deep autoencoder approach for cross layer IoT attack detection using deep learning algorithms," *Scientific Reports*, vol. 15, no. 1, p. 10246, 2025.
- [6] J. P. Ntayagabiri, Y. Bentaleb, J. Ndikumagenge, and H. El Makhtout, "A Comparative Analysis of Supervised Machine Learning Algorithms for IoT Attack Detection and Classification," *Journal of Computing Theories and Applications*, vol. 2, no. 3, pp. 395-409, 2025.
- [7] G. Uthradevi, P. Thiruvassagam, S. Mythili, and S. O. Manoj, "A Semi-Supervised Deep Learning Approach for Intrusion Detection and Classification for the Internet of Things," *Biomedical Materials & Devices*, pp. 1-17, 2025.
- [8] M. Al-Zewairi, S. Almajali, M. Ayyash, M. Rahouti, F. Martinez, and N. Quadar, "Multi-Stage Enhanced Zero Trust Intrusion Detection System for Unknown Attack Detection in Internet of Things and Traditional Networks," *ACM Transactions on Privacy and Security*, 2025.
- [9] H. Hakami, M. Faheem, and M. B. Ahmad, "Machine learning techniques for enhanced intrusion detection in IoT security," *IEEE Access*, 2025.
- [10] S. S. Alqahtany, A. Shaikh, and A. Alqazzaz, "Enhanced Grey Wolf Optimization (EGWO) and random forest based mechanism for intrusion detection in IoT networks," *Scientific Reports*, vol. 15, no. 1, p. 1916, 2025.
- [11] U. C. Akuthota and L. Bhargava, "Transformer-based intrusion detection for IoT networks," *IEEE Internet of Things Journal*, vol. 12, no. 5, pp. 6062-6067, 2025.
- [12] M. Benmalek and A. Seddiki, "Particle swarm optimization-enhanced machine learning and deep learning techniques for Internet of Things intrusion detection," *Data Science and Management*, 2025.
- [13] D. D. Bikila and J. Čapek, "Machine Learning-Based Attack Detection for the Internet of Things," *Future Generation Computer Systems*, vol. 166, p. 107630, 2025.
- [14] M. R. Minhas et al., "F-OSFA: A Fog Level Generalizable Solution For Zero-day DDOS Attacks Detection," *IEEE Access*, 2025.
- [15] Santosh K. Smmarwar, Govind P. Gupta, and Sanjay Kumar, "AI empowered malware detection system for industrial internet of things," *Computers and Electrical Engineering*, vol. 108, p. 108731, 2023.
- [16] X. H. Nguyen and K. H. Le, "Robust detection of unknown DoS/DDoS attacks in IoT networks using a hybrid learning model," *Internet of Things*, vol. 23, p. 100851, 2023.
- [17] Belal Ibrahim Hairab et al., "Anomaly detection based on CNN and regularization techniques against zero-day attacks in IoT networks," *IEEE Access*, vol. 10, pp. 98427-98440, 2022.
- [18] Segun I. Popoola et al., "Federated deep learning for zero-day botnet attack detection in IoT-edge devices," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3930-3944, 2021.
- [19] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. Ghorbani, "Ciciomt2024: Attack vectors in healthcare devices-a multi-protocol dataset for assessing iomt device security," 2024.
- [20] Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779-796, 2019.
- [21] P. Chen, F. Li, and C. Wu, "Research on intrusion detection method based on Pearson correlation coefficient feature selection algorithm," *Journal of Physics: Conference Series*, vol. 1757, no. 1, p. 012054, 2021.
- [22] U. Ahmed, Z. Jiangbin, A. Almogren, M. Sadiq, A. U. Rehman, M. T. Sadiq, and J. Choi, "Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems," *Scientific Reports*, vol. 14, no. 1, p. 30532, 2024.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9, 2015.
- [25] Ifageer, R., Ghaleb, M., Aljoby, W., & Felemban, M. (2026). A Multi-Stage Learning Framework for Detecting Known and Novel Attacks in the Internet of Medical Things. *IEEE Access*, 14, 48312-48326.
- [26] Torre, D., Chennamaneni, A., Jo, J., Vyas, G., & Sabrsula, B. (2025). Toward Enhancing Privacy Preservation of a Federated Learning CNN Intrusion Detection System in IoT: Method and Empirical Study. *ACM Transactions on Software Engineering and Methodology*, 34(2), 1-48.
- [27] Jodayree, M., Ghafi, A. K., Amiri, S., & Shaykholeslami, P. (2026). Explainable zero-day attack detection in IoMT using transformer-based time-series modeling. *Scientific Reports*.
- [28] Gueriani, A., Kheddar, H., Mazari, A. C., & Ghanem, M. C. (2025). A robust cross-domain IDS using BiGRU-LSTM-attention for medical and industrial IoT security. *ICT Express*.
- [29] D. Torre, A. Chennamaneni, J. Jo, G. Vyas, and B. Sabrsula, "Toward Enhancing Privacy Preservation of a Federated Learning CNN Intrusion Detection System in IoT: Method and Empirical Study," *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 2, pp. 1-48, 2025.
- [30] M. Luqman, M. Zeeshan, Q. Riaz, M. Hussain, H. Tahir, N. Mazhar, and M. S. Khan, "Intelligent parameter-based in-network IDS for IoT using UNSW-NB15 and Bot-IoT datasets," *Journal of the Franklin Institute*, vol. 362, no. 1, p. 107440, 2025.