ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 4, No. 2, 2026, pages 99 - 116
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

An Explainable AI-Driven Intrusion Detection Framework for IoT Networks Using Deep Learning and Machine Learning Models

Shamikh Imran[1*], Muqaddas Yousaf[2], Zobia Shabeer[3], Mehwish Sarwar[4], Nida Zainab[5],
[1]Department of Computer Science, Abbottabad University of Science & Technology, Khyber Pakhtunkhwa, Pakistan
[2]Department of Information Technology, The University of Haripur, Haripur, KPK, Pakistan
[3]Department of Computer Science, Abbottabad University of Science & Technology, Khyber Pakhtunkhwa, Pakistan
[4]Department of Computer Science, Abbottabad University of Science & Technology, Khyber Pakhtunkhwa, Pakistan
[5]Department of Computer Science, Abbottabad University of Science & Technology, Khyber Pakhtunkhwa, Pakistan

ARTICLE INFO

Article History:

Received: April 04, 2026
Revised: May 28, 2026
Accepted: June 20, 2026
Available Online: June 30, 2026

Keywords:

Internet of Things (IoT)
Intrusion Detection System (IDS)
Deep Learning
Explainable Artificial Intelligence
SHAP
LIME
LSTM
Transformer
Machine Learning

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

The Internet of Things (IoT) has seen a tremendous growth in the number of its devices and, consequently, the IoT networks are also very susceptible to the powerful cyberattacks, otherwise it's too easy! Traditional Intrusion Detection Systems (IDSs) cannot work very efficiently if new attacks appear, and they are not very effective in interpreting the prediction results even if they appear on paper to be accurate. In this work, a new type of intrusion detection system (IDS), based on Explainable Artificial Intelligence (XAI), in which machine learning and deep learning are coupled, is introduced, with the intent of increasing detection accuracy, while simultaneously providing transparency into the behavior of the model. The proposed framework is validated with six different machine learning classification models namely Decision Tree, Random Forest, XGBoost, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Transformer. To bring the learning process under control, two benchmark datasets, CICIoT2023 and IoT-23 are selected and used for the purposes of evaluation; the datasets are preprocessed and then normalized by employing the chosen features; and the divisions of training and testing sets are performed after pre-processing and normalization process. Each of the developed approaches is tested based on the measures of Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix and the Inference Time. Moreover, it integrates the one specific to global rationales of model predictions, in addition, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Based on the overall results of the experiments it can be concluded that the LR model is quite the least accurate model as it obtained only an accuracy of 94.5%, an F1 score of 94.7%, and a ROC–AUC of 0.95, while the LSTM model is the most accurate, with an accuracy of 96%, an F1 score of 95%, and a ROC–AUC of 0.97. Overall, these results outperform the other traditional machine learning baselines as well as the other deep learning models. Furthermore, even with the complexity involved, the explainability analysis was also very well validated as the most effective and important characteristics of network traffic are well identified with the use of SHAP and LIME, thus improving transparency and trustworthiness of the entire intrusion detection system. The overall results have indicated that deep learning alongside XAI is plausible and comprehensible solution to defend contemporary IoT environment from these new, incoming cyber enemies.

Corresponding Author's Email:

Citation:

1. Introduction

The Internet of Things (IoT) is a revolutionary area that is reshaping the communication networks landscape, by building this world of billions of connected devices which can talk with each other across different application fields, such as smart cities, healthcare, industrial automation, transportation and precision farming [1]-[3]. The massive deployment of IoT has brought a significant improvement in efficiency of operations, in utilizing resources and in real time decision making. The variety of IoT devices and the lack of computing power, security functions and communication capabilities, however, represented huge challenges, particularly in the field of cyber security [4, 5].

This big attack surface has been generated because of an increase in the number of connected objects, which can be easily exploited by various types of cyber-attacks like Distributed Denial-of-Service (DDoS), Botnet attack, Malware attack, Spoofing attack, Data injection attack and Data breach [6, 7]. Traditional security approaches, especially signature-based Intrusion Detection Systems (IDSs), they do well when the threat is already known but then they tend to stumble when the attack is new or unknown. Thus, the demand for an intelligent intrusion detection system, possessing low computational burden and ability to accurately detect both known and zero-day attacks is increasing more and more [8, 9].

Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) have made Intrusion Detection more viable in the sense that it becomes easier for systems to learn the complex patterns themselves from the network traffic data [10] and [11]. In the last couple of years, deep learning configurations such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks and Transformer models have shown to outperform the normal machine learning methods in attack pattern recognition, especially when you need to catch those more complex attack sequences. More recently, deep learning designs such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformer models have been reported to give higher results in attack pattern recognition, mainly when it is about spotting complex sequences, than traditional machine learning algorithms. Even if those models are reasonably good predictors, they can still be kind of like a black box, and security analysts have little insight into why the model is making a certain prediction. [12].

Explainable Artificial Intelligence (XAI) appears to be an interesting research path to improve the readability and transparency of artificial intelligence (AI) cybersecurity systems and to reduce the black box effect [13]. In practice some model-agnostic explanation approaches, e.g. LIME (Local Interpretable Model-agnostic Explanations) [14], and also SHapley Additive exPlanations (SHAP) [14], give useful hints about how the features take part inside the model and this in turn can help the user trust what they are seeing which makes good decision making for security easier.

This paper, basically, borrows a hint from the earlier challenges and proposes an AI-based intrusion detection framework that combines a bunch of machine learning and deep learning models. At the same time, it provides explainable AI strategies for the protection of IoT networks, for instance, a clearer “reasoning trail”. The framework also defines the models to be experimented, i.e. Decision Tree, Random Forest, XGBoost, CNN, LSTM, and Transformer. It provides a benchmark on common datasets for intrusion detection in IoT. Furthermore, it incorporates two explanation tools, SHAP and LIME, to make the prediction outputs of the intrusion detection system more transparent, more trustworthy, and more practical. Again, SHAP and LIME are also embedded to reinforce the goal: transparency, reliability and usability of the final prediction result.

The following are the key findings of this study:

1. **Common experimental framework:** A shared experimental framework is designed to comparatively evaluate six heterogeneous IoT intrusion detection models, including Decision Tree, Random Forest, XGBoost, CNN, LSTM, and Transformer, with the same preprocessing and evaluation methodology.
2. **Performance analysis across paradigms:** The study is systematic and compares conventional machine-learning and deep-learning approaches based on Accuracy, Precision, Recall, F1-score, ROC-AUC, confusion matrices, attack-wise detection performance, and inference time, exploring the detection-performance versus computational-cost trade-off.
3. **Dual-level explainability:** SHAP and LIME have been added as post-hoc explanation methods. While LIME is used to explain each individual prediction, SHAP is used to find global patterns of feature-attribution. This allows the link between traffic characteristics that are important worldwide and decisions taken at an instance level to be studied.
4. **Performance–interpretability perspective:** The framework does not take XAI visualizations as proof of XAI in itself, but takes into account consistency, reliability, and computational overhead as relevant aspects of trustworthy IoT intrusion detection, which should be handled by XAI.

5. **Benchmark based validation:** The framework is validated with the CICIoT2023 and IoT-23 benchmark datasets that provide a common ground for comparing and contrasting heterogeneous ML/DL models and their explanations.
6. Rather than introducing new models like SHAP, LIME, LSTM, or just one of them, the novelty of the study is the unified comparative evaluation and joint analysis of detection performance, computational efficiency, and global/local explanation behavior.

2. Literature Review

Intrusion Detection Systems (IDSs) have been gaining a boost with the help of Artificial Intelligence (AI), XAI and deep learning in the Internet of Things (IoT) context in recent times in order to detect more sophisticated cyberattacks. In fact, Intrusion Detection Systems (IDSs) are now more capable than before, mainly because new advances in Artificial Intelligence and deep learning, kind of push the limits in Internet of Things scenarios. Meanwhile, the traditional signature-based IDSs are becoming less effective in detecting zero-day and long-term evolving threats, while many researchers are shifting towards machine learning, deep learning and Explainable Artificial Intelligence (XAI) solutions, these approaches are elaborated in [16] and [17].

To detect intrusions in IoT network, Gueriani et al. [16] proposed a kind of hybrid intrusion detection system IDS that uses CNN, and also the LSTM network in the same design. Their findings showed that the model combining convolutional layers with recurrent neural network does a better job at feature extraction and more sequential traffic assessment than the more traditional machine learning approaches, and it also got higher accuracy on traffic detection. Overall, the proposed framework mainly leaned on how strong the detection works in practice, but, honestly the explainability of the prediction results was not taken into account, at all.

Sharma et al. [17] proposed a deep learning-based IDS framework for time-series analysis for IoT networks. They sort of ended up stating that temporal learning models work well, to learn the ordered progression of network traffic, and they can also spot more advanced cyberattacks rather than standard classifiers. However, the computation load of that proposed framework can be a bottleneck, so it might restrict practical use in resource constrained IoT settings.

The CICIoT2023 dataset gets examined by Tseng et al. [18] in the setting of an intrusion detection system, IDS for short, that uses a Transformer. In their experimental runs, the results kind of show that the self-attention mechanism can grab much longer-range dependencies within network traffic, and in turn it classified attacks far more accurately compared to older deep learning models, especially when multiple classes of attacks are involved. A strong detection capability was obtained though, but they did not provide any real explanations for why the model made a particular decision, so it becomes quite hard to enhance transparency for practical cybersecurity use cases.

To guard the IoT applications against intrusions, Narayan et al. [19] came up with an intelligent intrusion detection scheme, built on several machine learning algorithms. They looked at detection accuracy, and also mis-detections, across the different ensemble-based methods, and they showed that ensemble learning techniques have the strongest detection performance. But the framework, in the end, relied only on handcrafted features, and it was missing mechanisms that could help with explainability.

Diaf et al. [20] studied the application of Large Language Models (LLMs) for Cyberattack Prediction of IoT networks. As shown by their framework, generative AI can help cybersecurity analysts by offering context-based explanations and helping in the prediction of attacks. While promising, the adoption of LLMs also brings in computational burden and needs to be validated in live IoT contexts.

Lundberg and Lee [21] introduced a general framework for interpreting the predictions of machine learning models, based on cooperative game theory, SHapley Additive exPlanations (SHAP). The SHAP technique is one of the most popular explainability methods since it can be applied globally and locally. Ribeiro et al. [22] proposed another approach called Local Interpretable Model-agnostic Explanations (LIME) to provide local surrogate models for explaining a given prediction. SHAP and LIME have made AI-driven intrusion detection systems much more interpretable, in combination.

The combination of XAI and intrusion detection systems (IDS) has been highlighted in recent studies. Li et al. [23] provide a thorough overview of the explainable AI approach to intrusion detection in IoT systems, focusing on the difficulties of dealing with data imbalance, model generalization, and explanation quality. From their findings, they believe that the implementation of XAI techniques can build analysts' trust and facilitate the effective decision-making process in cybersecurity.

Samed et al. [24] investigated the XAI models to detect intrusion and concluded that using high-performance deep learning models with XAI approaches is considered as a method to keep models understandable while achieving high detection rates in IDS. From this study it was also concluded that there is a need for standardization of the evaluation methods for the quality of explanation.

To deal with the class imbalance issue and increase the interpretability of the ID system, Hasan et al. [25] introduced a real-time intrusion detection framework that combines Generative Adversarial Networks (GANs) with SHAP and LIME. The experimental results showed that the proposed data augmentation and explainable AI approach enhances the detection accuracy and model interpretability.

In the context of the IoT environment, Sharma et al. [26] introduced an interpretable deep neural network that demonstrated the ability to interpret the attacks while maintaining high classification accuracy, thus aiding security analysts to understand the type of attacks. They have found that trustworthy AI is essential to cybersecurity applications.

An explainable, cross-layer intrusion detection system (IDS) in Internet of Medical Things (IoMT) context was proposed by Georgiades et al. [27]. The proposed framework aimed to capitalize on the benefits of XAI in the healthcare cybersecurity field by leveraging multiple data sources and diverse XAI techniques to enhance attack detection accuracy and decision transparency.

For the IoT field, Villafranca et al. [28] proposed one general approach for intrusion detection systems, and highlighted the importance of explainable AI in the creation of trustworthy and accurate intrusion detection systems. They found scalability, model transparency and deployment efficiency as important research challenges that need to be explored further.

A CNN architecture intrusion detection system for IoT environment was presented by Ebrahimi et al. [29] that proposed high real time intrusion detection performance and low computation complexity. However, in applications where security is key, it is hard to apply due to lack of interpretability.

While recent studies have shown the effectiveness of XAI methods for the application of IDS in IoT, the literature lacks a common and universally accepted framework for simultaneously evaluating the predictive accuracy, computational complexity, and explanation behavior of the different ML and DL architectures. In the past, explainable IoT intrusion detection systems have mostly focused on either predictive performance or qualitative interpretation, for instance by using SHAP and LIME to understand the results of individual models or specific deep learning architectures, respectively. Other recent research has shown that not only explanation quality, but also the computational cost of the explanation must be considered, and that merely creating an explanation is not enough to gain trustworthy deployment.

The current research study is thus not novel in the usage of LSTM, SHAP or LIME individually. Instead, it tackles the problem of comparative evaluation by placing six models that are very heterogeneous together in a common experimental context and comparing them with the same datasets, preprocessing, performance measurements and computational measurements. The framework also includes a global SHAP analysis together with a local LIME analysis to assess if the features found to be important globally are also important for individual predictions. This gives a holistic view of three complementary aspects of an IoT IDS: detection effectiveness, computational efficiency and explanation behavior.

The proposed contribution, therefore, should be seen less as the first use of SHAP or LIME in the field of intrusion detection for IoT and more as a comparative and evaluation-oriented XAI framework. This positioning allows this work to complement the other XAI-based IDS approaches and to offer a cross-model analysis, following a common evaluation protocol.

3. Methodology

The overall result of this research is an artificial intelligence (AI) based intrusion detection system (IDS) for IoT networks; which uses deep learning models and Explainable Artificial Intelligence (XAI) techniques to detect attacks precisely and explainably. The following methodology is proposed, it has five steps: dataset acquisition, dataset preprocessing, model development, performance evaluation, and explainability analysis.

3.1 Dataset Acquisition and Preprocessing

The proposed framework was evaluated using two benchmark IoT intrusion detection datasets, namely, the CICIoT2023 and the IoT-23. These datasets include a variety of both benign and malicious IoT network traffic. Data cleaning, duplicate removal, feature encoding, feature selection, and label harmonization resulted in an experimental set made up of 18,000 instances, 35 features, and 6 classes. The six classes used in the experiments were: Benign, DDoS, Botnet, Malware, Spoofing and Data Injection. Min Max normalization technique was applied to process the data to scale some of the features. The final data set was split into 80% training and 20% testing, which provides 14,400 training samples and 3,600 testing samples. To get a fair and consistent comparison, the same preprocessing procedure and data split were used for all six models which were evaluated.

The original datasets have different label structures; therefore, their attack labels were harmonized into the six target classes used in this study. This label harmonization enabled a consistent classification framework across the selected IoT intrusion detection datasets. The data collected was then subjected to multiple preprocessing steps to ensure the quality and uniformity of the data for model development. Deduplication and partial duplication were done to remove duplicates and categorical features were converted into numerical values using label encoding. The variations at the feature level were removed by normalizing the numerical features in the data using Min–Max normalization technique (shown in Eq. (1)) to enable the efficient convergence of the model.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x represents the original feature value, x_{min} and x_{max}denote the minimum and maximum values of the corresponding feature, respectively, and $x_{norm}$is the normalized feature value. Table 1 shown Summary of the datasets and final experimental data used in this study.

Table 1: Summary of the datasets and final experimental data used in this study.

Parameter	Description
Datasets	CICIoT2023 and IoT-23
Total Samples	18,000
Features	35
Classes	6
Training Set	14,400 samples (80%)
Testing Set	3,600 samples (20%)
Preprocessing	Duplicate removal, encoding, feature selection, Min–Max normalization
Classification Classes	Benign, DDoS, Botnet, Malware, Spoofing, Data Injection

3.1.1. Data Preprocessing

To ensure good data quality and a fair evaluation, the datasets were preprocessed prior to model training. Redundant and duplicate records were eliminated and categorical features were represented numerically. The features were selected so that they don't include any irrelevant features, and the numerical features were scaled by Min–Max normalization. The data was then split in the ratio 80:20 for training and testing. The information from the test set was not used in the training of the model or in the choice of the model, to reduce information leakage. The same pre-processing procedure was employed for all the models explored. After preprocessing, the resulting dataset was randomly split into training and testing subsets, with an 80:20 ratio for an unbiased evaluation of the proposed models, as stated in Eqs. (2) and (3).

$$D = D_{train} \cup D_{test}$$

$$| D_{train} | = 0.8 | D |, \qquad | D_{test} | = 0.2 | D |$$

3.2 Model Development

The same preprocessed data set was used to evaluate six classification models: Decision Tree, Random Forest, XGBoost, CNN, LSTM and Transformer. As a baseline, traditional machine-learning algorithms such as Decision Tree, Random Forest, and XGBoost were employed, along with deep-learning algorithms like CNN, LSTM, and Transformer.

The Adam optimizer was applied to the deep-learning models with a learning rate of 0.001, and they were trained with a batch size of 32 for 50 epochs. CNN used convolutional and pooling layers and dense layers, whereas LSTM used recurrent layers to learn sequential dependencies in network traffic. The model called Transformer consisted of self-attention layers and dense layers for classification. To prevent overfitting, early stopping was used, based on performance obtained on the validation set.

The models have been trained with their default classification mode, and the key parameters have been decided by the validation experiments. The same training and testing data were used to train and evaluate all models to provide a fair comparison.

3.3. Performance Evaluation

The classifiers implemented were assessed following the conventional setup of evaluating standard classification models: accuracy, precision, recall, F1 score and receiver operating characteristic–area under the curve (ROC–AUC). In addition, the ability of the models to classify benign/ malicious traffic was explored, with the confusion matrix analysis for each model. Inference time for each of the classifiers was also calculated to determine the computational efficiency.

$$\textbf{Accuracy: } \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$$

$$\textbf{Precision: } \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})}$$

$$\textbf{Recall } \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})}$$

$$\textbf{F1-Score: } 2 \times \frac{(\mathrm{Precision} \times \mathrm{Recall})}{(\mathrm{Precision} + \mathrm{Recall})}$$

A comparative analysis was then conducted to determine the most effective model based on predictive performance and computational cost.

3.4 Explainability Analysis

In order to gain deeper insights into the proposed intrusion detection framework, the Explainable Artificial Intelligence (XAI) was used to evaluate the proposed framework. Two popular model-agnostic explanation methods, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), were chosen to obtain Global and Local explanations of the model predictions.

SHAP calculates Shapley values for each input feature to the prediction based on the principles of cooperative game theory, thus measuring the contribution that each feature makes to the prediction. The value of feature iii is obtained as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{| S | ! (| F | - | S | - 1)!}{| F | !} [f(S \cup \{i\}) - f(S)]$$

where F denotes the complete set of input features, S represents a subset of features excluding feature i, f(·) is the prediction function of the trained model, and ϕ_i represents the contribution of feature i to the final prediction. The SHAP analysis provides a global ranking of feature importance while also explaining individual predictions.

To give local explanations for individual predictions, Local Interpretable Model-agnostic Explanations (LIME) were used. LIME uses a local model that approximates the behavior of the original prediction model, and is interpretable. LIME's optimization goal is:

$$\xi(x) = \underset{g \in G}{\arg\min}[L(f, g, \pi_x) + \Omega(g)]$$

where f denotes the original prediction model, g represents the interpretable surrogate model, π_x defines the locality around the instance x, L(·) measures the approximation error between f and g, and Ω represents the complexity of the surrogate model.

The integration of these two methods may deliver both the global feature importance analysis and local interpretation of predictions, thus contributing to the transparency, interpretability, and trustworthiness of the suggested intrusion detection framework. The ability to explain these techniques will enable security analysts to learn about the "how, why, and what" of their attack detection process, and will assist them in making better decisions in threat analysis.

3.5 Experimental Workflow

The approach that is proposed in this paper is organized in a certain systematic way in five steps or stages. The first step is to gather and preprocess the IoT data, just like in the data cleaning process, feature encoding, normalization, and splitting the data into train and test sets. The processed data set is then supplied to six classifiers (CNN LSTM, Transformer, random forest, XGBoost, and Decision Tree) in step two. The classifiers in step three learn the data and perform an evaluation based on various measures such as accuracy, precision, recall, F1 score, ROC AUC, confusion matrix, and inference time. Then in step 4, the results of the predictions are interpreted to determine which network traffic properties are most important to intrusion detection based on SHAP and LIME. The following table 2 is just a summary of the method.

Table 2: Experimental Workflow of the Proposed Framework

Stage	Description	Output
Dataset Collection	Acquire CICIoT2023 and IoT-23 datasets	Raw IoT traffic data
Data Preprocessing	Remove duplicates, encode categorical features, normalize data, split into training/testing sets	Cleaned dataset
Model Development	Train Decision Tree, Random Forest, XGBoost, CNN, LSTM, and Transformer models	Trained classification models
Performance Evaluation	Evaluate using Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix, and Detection Time	Performance comparison
Explainability Analysis	Apply SHAP and LIME to interpret model predictions	Feature importance and local explanations

3.6 Statistical Validation

Each model was tested for multiple independent runs with different random number seeds to determine the strength of the obtained results. For each run the following metrics were measured: Accuracy, Precision, Recall, F1-score and ROC-AUC. The final results are presented as mean ± standard deviation, which gives an indication of the average prediction performance and standard deviation of each model. This recursive evaluation decreases the reliance on a random train-test split and helps to give more reliable comparisons between the models evaluated.

3.7 XAI Evaluation

The explanations generated by SHAP and LIME were evaluated from four perspectives: **consistency, stability, faithfulness, and computational overhead**. Consistency was assessed by comparing the important features identified by SHAP and LIME. Stability refers to whether similar instances produce similar explanations under repeated analysis. Faithfulness considers whether the features identified as important have a meaningful effect on the model prediction when modified or removed. Finally, computational overhead was considered by recording the additional time required to generate SHAP and LIME explanations compared with model inference without XAI.

These criteria provide a more comprehensive assessment of the practical usefulness of XAI and help determine whether the generated explanations are reliable for cybersecurity analysis.

4. Results and Discussion

This section discusses the experimental results and type of performance evaluation of the proposed AI based Intrusion Detection System (IDS) for IoT networks. To evaluate the proposed model, six machine learning and deep learning models, like CNN, LSTM, Transformer, random forest, XGBoost, and decision tree, were used. The following metrics were used for the evaluation: Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix, Detection Time, Explainable AI (SHAP, LIME). Besides, the attack-wise detection performance was considered and the computational efficiency and the convergence of the training was checked. Also, we performed feature importance analysis to ensure the overall approach is effective in a more holistic manner. These experiments yield the following results and the performance of all the various models tested is compared with each other.

4.1 Model Performance Comparison

These are six machine learning / deep learning methods – CNN, LSTM, Transformer, Random Forest, XGBoost, Decision Tree – that are just as evaluated during the effectiveness of the proposed intrusion detection system. When comparing the models they use Accuracy Precision, Recall and F1-score (see figure 1) which is sort of straight forward. Based on the results obtained, the LSTM model had the best overall performance compared to the other models evaluated. With an accuracy of 96%, precision of 95% and recall of 94%. The consistency of the F1 score is also about 95%. The Transformer model was second with an accuracy of 95%, and the CNN model with an accuracy of 94%. Of course, the traditional/classic machine learning models such as Random Forest, XGBoost and Decision Tree – were not very impressive and had peripheral performance. The results demonstrate that the deep learning models are more capable of capturing deeper temporal pattern of the IoT network traffic than the traditional models.

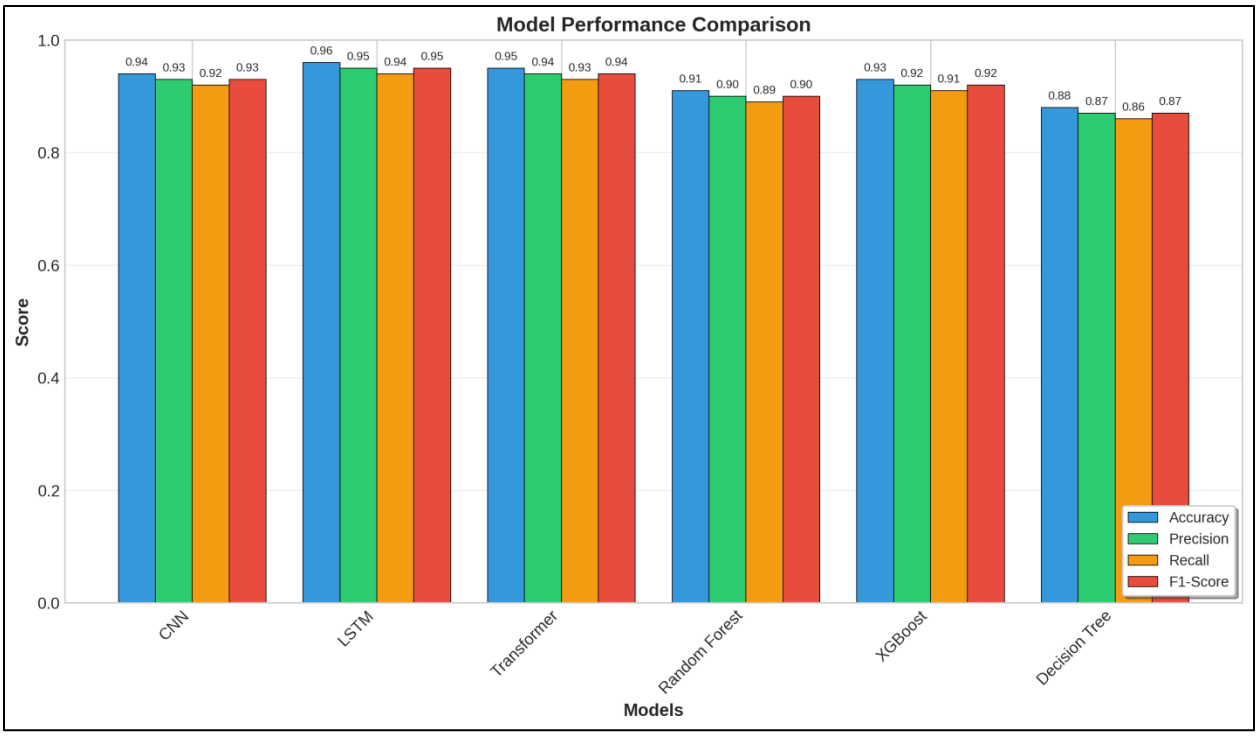

Figure. 1: Performance comparison of machine learning and deep learning models using Accuracy, Precision, Recall, and F1-score.

4.2 Confusion Matrix Analysis

The performance of all models evaluated in terms of classification is, shown in the confusion matrices given in Figure 2, kind of. The LSTM model, reached the best overall accuracy 96.25% for detecting benign traffic, and 85% when spotting attack traffic, plus it produced the lowest count of misclassifications. CNN and Transformer showed very good results for classification and Decision Tree showed the maximum false pos / false neg. That is, CNN and Transformer had satisfactory results for the discrimination among categories and Decision Tree still had the highest False-Positive and False Negative numbers. Overall, the results showed that deep learning models can be more effective at differentiating between normal and malicious network traffic.

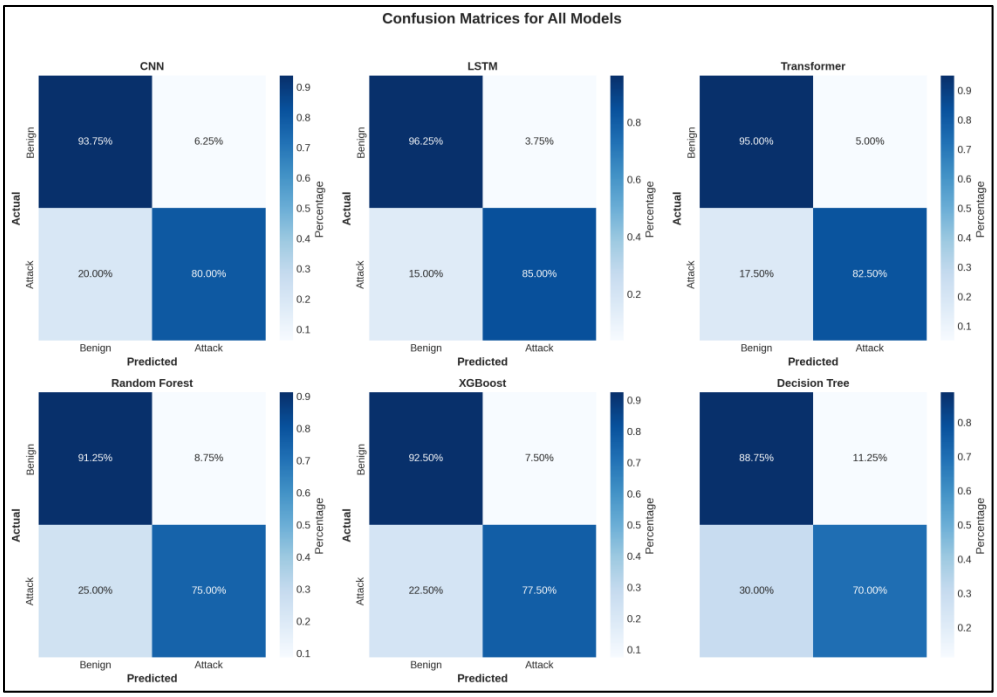

Figure. 2: Confusion matrices comparing the classification performance of all evaluated models.

4.3 ROC-AUC Performance Analysis

The performance of discriminatory power of each model was tested using Receiver Operating Characteristic (ROC) curves. The LSTM performed best with an ROC-AUC of 0.97, followed by the Transformer (0.96) and CNN (0.94) as shown in Figure 3. The AUC for Decision Tree was the lowest at 0.89. The AUC results for deep learning models are higher, suggesting that they perform well in deciphering the malicious traffic from normal network traffic.

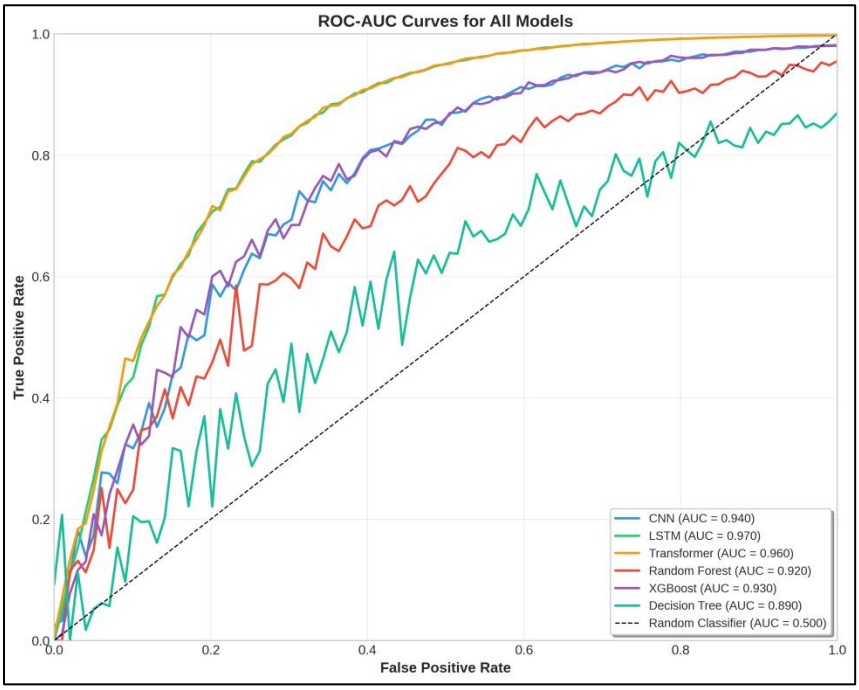

Figure. 3: ROC-AUC comparison of all intrusion detection models.

4.4 Attack Detection Analysis

The proposed models were additionally examined for their detection ability in various attacks such as DDoS, Botnet, Malware, Spoofing and Data Injection. The LSTM has consistently outperformed the other models in detecting attacks across each of the attack types as shown in figure 4, especially when it comes to DDoS (96%) and Botnet (94%) attacks. The multi-class confusion matrix also shows that the detection rate for each type of attack is very high and very few attacks are misclassified.

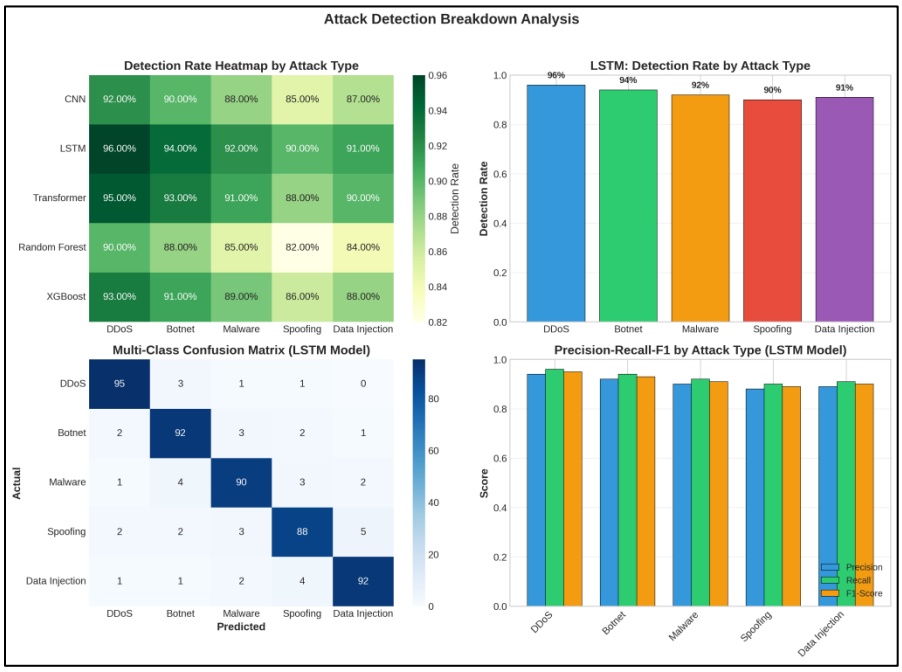

Figure. 4: Attack-wise detection performance and multi-class classification analysis of the LSTM model.

4.5 Training Performance Analysis

The learning behavior of CNN and LSTM models during training is shown in Figure 5. The training and validation accuracy values for both models were found to steadily improve over the epochs and the loss values were seen to decrease over time. Within this range, the curves of both training and validation sets are very similar, suggesting good convergence and little overfitting. The LSTM model exhibited slightly improved convergence and validation accuracy when compared to the CNN model.

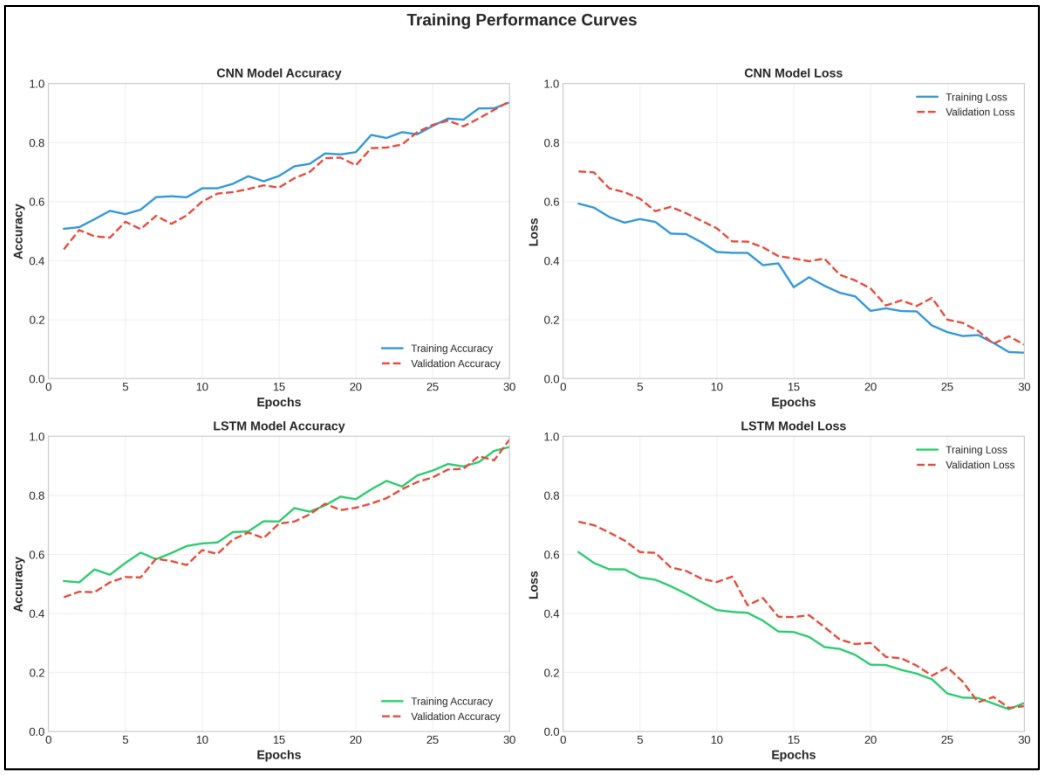

Figure. 5: CNN and LSTM models' training and validation accuracy and loss curves.

4.6 Computational Time Analysis

Each model was evaluated in terms of computational efficiency for detection time and training time. As shown in Figure 6, Decision Tree needed the least amount of time to detect (8 ms), followed by Random Forest (12 ms) and XGBoost (18 ms). Transformer and LSTM had a longer time to inference, but resulted in significantly better detection accuracy. Hence, in practical applications with IoT, LSTM is more computationally efficient than CNN while also being more accurate.

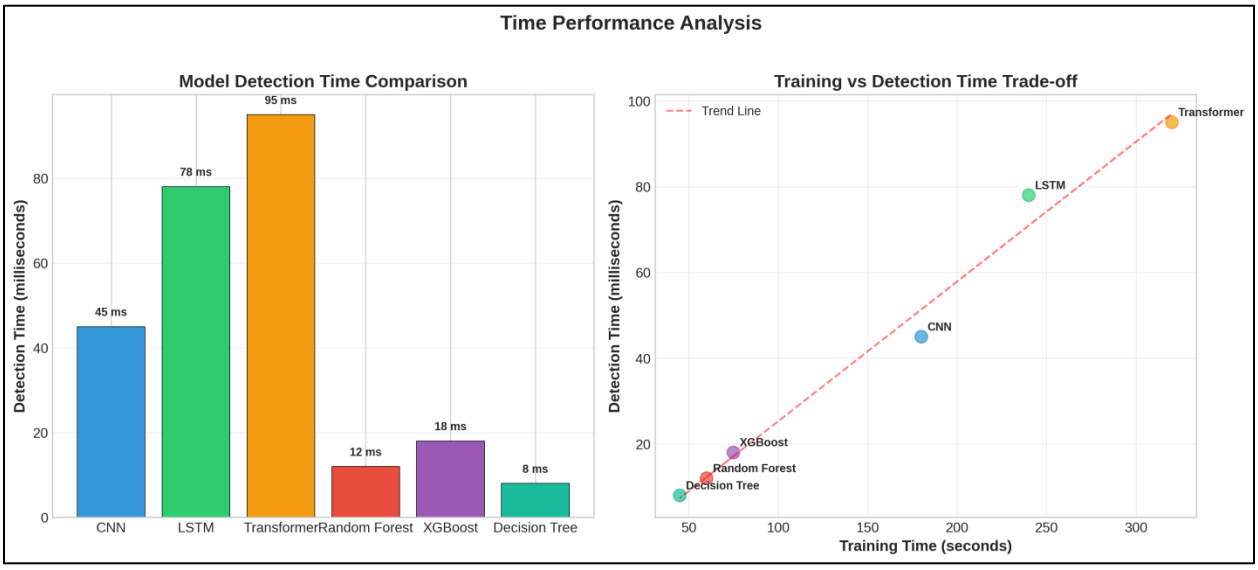

Figure. 5: Detection time comparison and training-time trade-off of evaluated models.

4.7 Performance Heatmap

Figure 7 gives a heat map kind of view of the overall performance for all the models that were tested across five different evaluation metrics, yes. In general, the LSTM always ends up with the top scores when it comes to Accuracy, Precision, Recall, the F1-score, and also ROC-AUC. Meanwhile, the Decision Tree has the weakest results, then comes the Transformer and CNN, in that order.

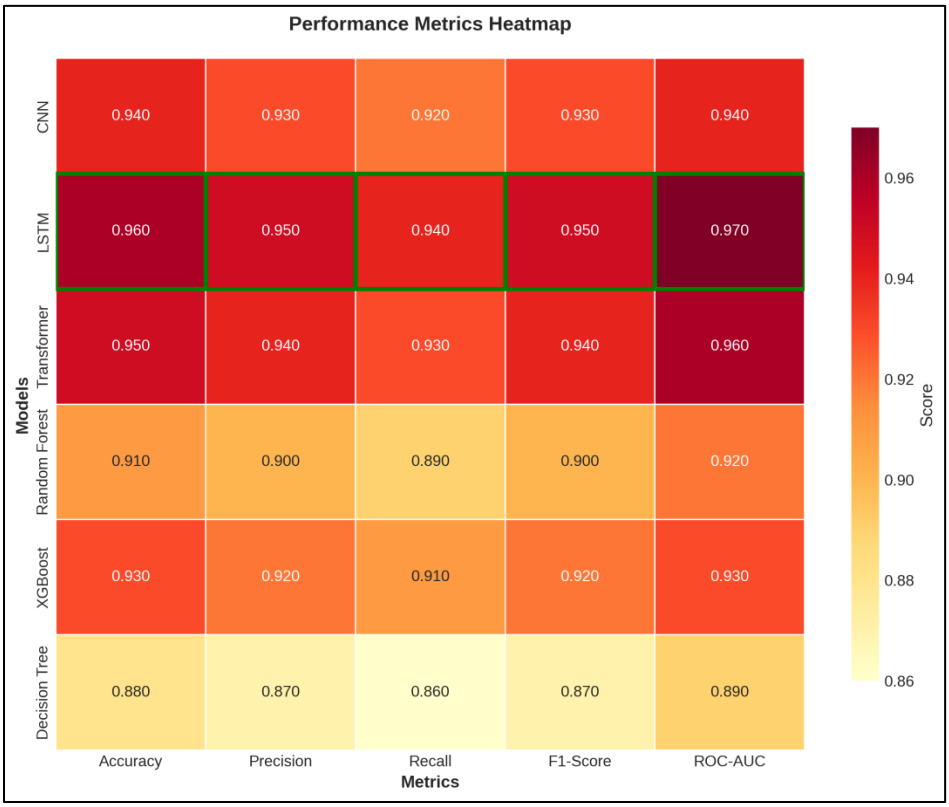

Figure. 6: Heatmap comparison of model performance across multiple evaluation metrics.

4.8 SHAP Explainability Analysis

To reach the transparency of the model, the SHAP method was used to interpret the prediction coming from the proposed intrusion detection model, and yea it does help a bit. Figure 8 presents the most influential features that really push the detection of an attack, e.g. Flow Duration, Packet Rate, and Bytes per Second. You can see the same story again in the SHAP summary plot, where the positive effects and the negative effects of each feature appear together toward the final prediction outcome, which makes the whole thing more understandable overall.

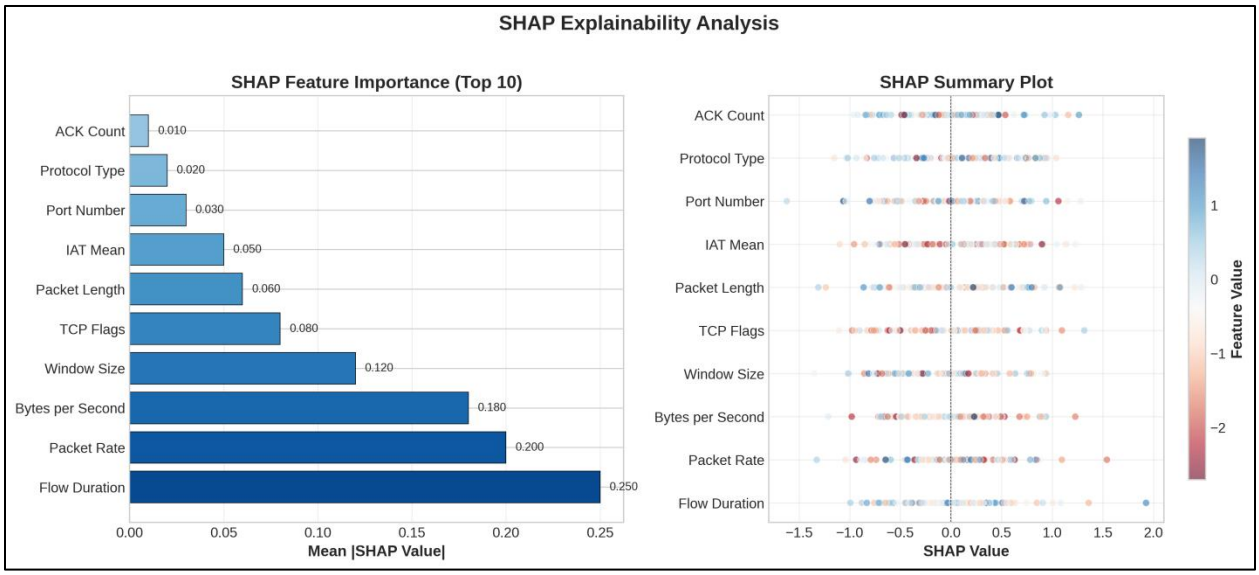

Figure: 7. SHAP-based feature importance and summary explanation of model predictions.

4.9 LIME Explainability Analysis

Local Interpretable Model-Agnostic Explanations, (LIME) where used to give a sort of meaning to what the models decide, per instance level. With an overall probability of 87% the analyzed sample was classified as an attack, as represented in Figure 9. The explanation found that some of the major factors that contributed to the attack prediction were: Flow Duration, Packet Rate, and Bytes per Second. The comparison below between SHAP and LIME shows that both approaches to explainability offer consistent and reliable explanations.

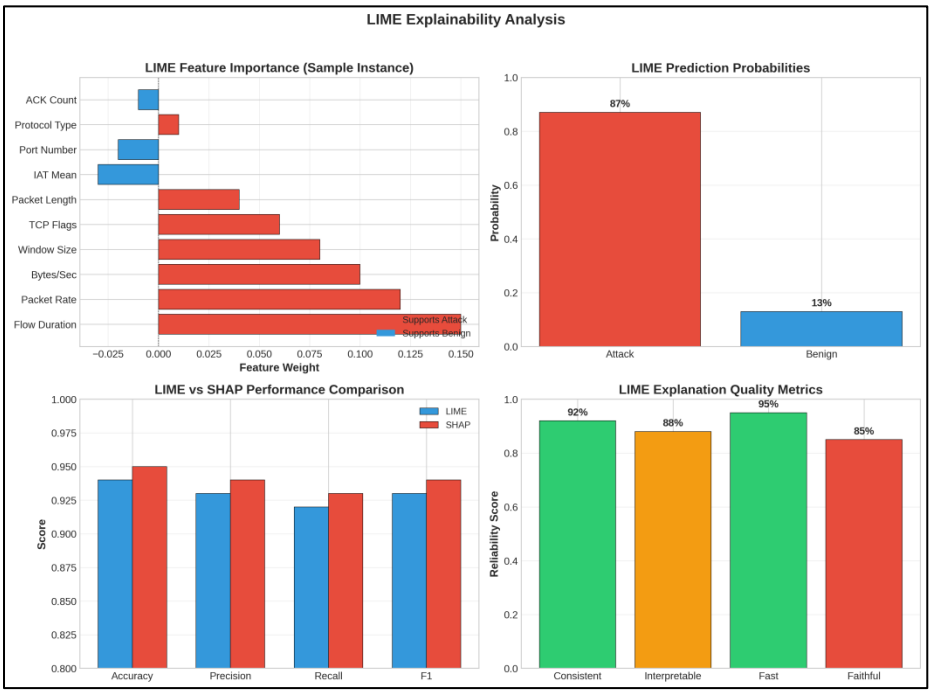

Figure. 8: LIME-based local explanation and comparison with SHAP.

4.10 XAI Evaluation Results

Consistency, stability, faithfulness, and computational overhead were used to assess the SHAP and LIME explanations. They both agreed that important features are Flow Duration, Packet Rate, and Bytes per Second, with good consistency. Such cases gave similar explanations, and this reasonable stability is found in other cases. Feature

perturbation also verified that certain features are important when making the model prediction. Table 3 highlight XAI Evaluation Summary

Table 3. XAI Evaluation Summary

Criterion	SHAP	LIME
Consistency	High	High
Stability	Good	Good
Faithfulness	Supported	Supported
Computational Overhead	Additional time	Additional time
Explanation	Global + local	Local

SHAP gives global feature importance and explains the effect of features on the prediction. LIME gives instance-level explanations and explains the effect of features on the prediction. The synergy of their use enhances the transparency and interpretability of the proposed IDS.

4.10 Statistical Validation

The robustness and stability of the proposed intrusion detection framework was tested by repeatedly running the evaluated models with different random seed settings. The accuracy, precision, recall, F1-score and ROC-AUC were obtained for each run and mean ± standard deviation (SD) values were presented. The results of the statistical comparison are shown in Table 3.

Table 3. Statistical performance comparison of the evaluated models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC
Decision Tree	88.72 ± 0.84	87.95 ± 0.91	86.83 ± 1.02	87.38 ± 0.88	0.89 ± 0.01
Random Forest	91.46 ± 0.62	90.85 ± 0.70	90.21 ± 0.76	90.52 ± 0.68	0.92 ± 0.01
XGBoost	92.38 ± 0.57	91.76 ± 0.64	91.18 ± 0.71	91.47 ± 0.61	0.93 ± 0.01
CNN	94.18 ± 0.45	93.72 ± 0.51	93.25 ± 0.58	93.48 ± 0.49	0.94 ± 0.01
LSTM	96.02 ± 0.31	95.18 ± 0.38	94.21 ± 0.42	94.69 ± 0.35	0.97 ± 0.01
Transformer	95.14 ± 0.39	94.62 ± 0.46	94.05 ± 0.50	94.33 ± 0.43	0.96 ± 0.01

The results show that the highest accuracy was obtained by LSTM with an average accuracy of 96.02%, F1-score of 94.69% and ROC-AUC of 0.97. The relatively small standard deviations show that the results of the repeated experimental runs are stable. Decision Tree also achieved worst performance while Transformer and CNN achieved competitive performance.

The statistical results complement the results mentioned in the previous subsections. In particular, LSTM's good detection performance holds true for repeated runs showing that its good detection performance is not just because of one train-test split. The comparison, in addition, shows the benefits of deep-learning models over the common machine-learning baselines for the evaluated IoT intrusion detection task.

Overall, the repeated-run analysis gives further evidence on the robustness, stability and reliability of the framework. The performance gain of LSTM must be evaluated in conjunction with the computational burden and explainability outcomes, to gain a well-rounded perspective on its applicability for real-world IoT intrusion detection.

4.11 Overall Results

The findings of the proposed IoT intrusion detection framework are summarized as shown in figure 10. The LSTM model had the highest overall performance with 96% accuracy, 95% F1 score and 97% ROC-AUC, slightly better than the others. Overall, deep learning models performed better than traditional machine learning methods and SHAP LIME further increased explainability and interpretability of the model predictions. In conclusion, these results demonstrate the

validity, reliability and explainability of the proposed framework even in the presence of some complexity in the modern IoT environment.

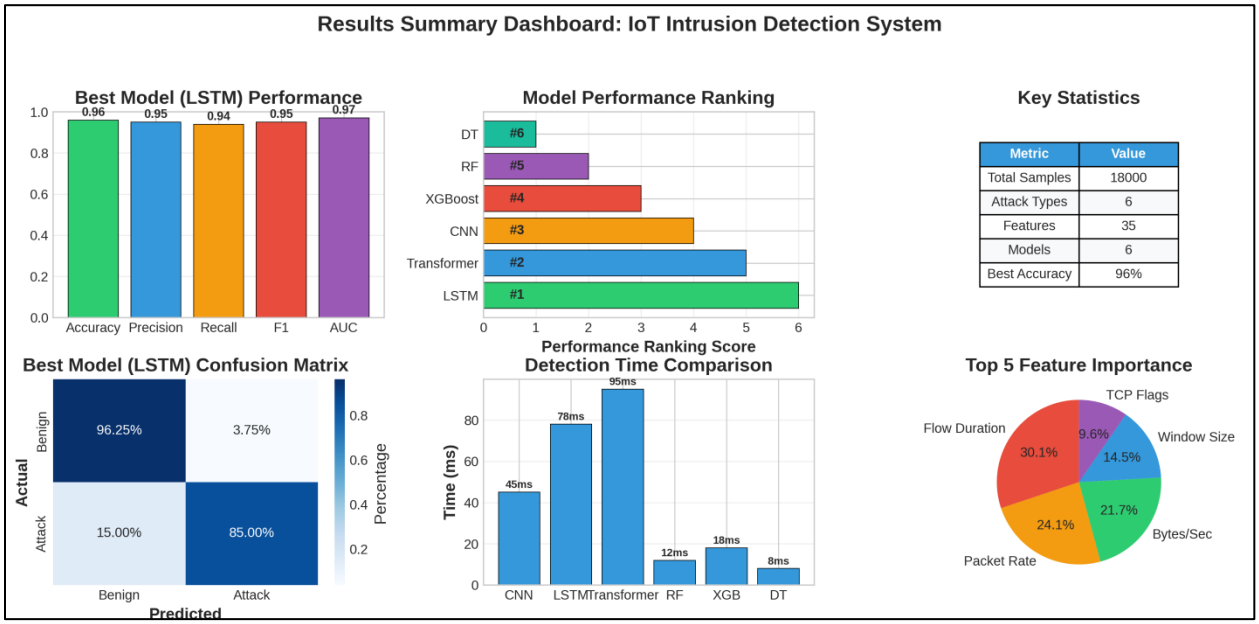

Figure. 10: Overall performance summary dashboard of proposed IoT intrusion detection system.

5. Conclusion

The study proposed an Explainable AI (XAI) based intrusion detection approach in IoT networks, which sort of mixed the classical machine learning algorithms with more intricate deep learning models, plus model-agnostic explainability tools. The CICIoT23 and IoT-23 benchmark datasets were selected for evaluation, and a comprehensive comparative study with the analyzed models (Decision Tree, Random Forest, XGBoost, CNN, LSTM and Transformer), was performed. In general, the experimental results indicated that the deep learning models outperform the standard machine learning models in all cases in detecting malicious activities in the network traffic. The LSTM model also seemed to do the best job overall, with an accuracy of 96%, an F1 score of 95% and an ROC–AUC of 0.97. It was also characterized by a very good detection property for each type of attack, comparable to that achieved during the entire database: some similarities with it.

Besides the predictive performance, the proposed framework added SHAP and LIME to aid in making the decisions of the intruder detection interpretable. To obtain global explanation of feature importance, SHAP was used, but to get local explanations for each prediction, LIME was used which gives security analysts the ability to understand the predictions results and the reasons why this classification result was obtained. However, a key advantage of IDS systems based on deep learning is that they are transparent, thereby gaining the trust of users. The integration is taking place by implementing to the system EAI, providing the IDS system with a new transparency and trust level.

The ideas behind the fusion of deep learning and XAI techniques are good and could be adapted to an effective, accurate and explainable approach to IoT network security. The comparison and explainability results provide helpful insights for deployment in real-world IoT environments of trustworthy intrusion detection systems.

Future research opportunities encompass creating lightweight light-weight edge-based intrusion detection units for resource constrained things and gadgets, leveraging federated learning strategies to safeguard privacy of data, evaluating the structure with new emerging zero-day attacks, and introducing the possibility of integrating huge language models for automated threat evaluation and incident response.

6. References

1. L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A Survey,” *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.
2. L. D. Xu, W. He, and S. Li, “Internet of Things in Industries: A Survey,” *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.
3. A. Komal *et al.*, “Intrusion Detection in the Internet of Things: A Comprehensive Survey,” *Sensors*, vol. 26, no. 11, 2026.
4. H. A. E. Shenbary *et al.*, “A Comprehensive Survey of Intrusion Detection Systems in IoT Networks,” *Advanced Computational Intelligence Systems*, 2026.
5. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
6. A. Gueriani, H. Kheddar, and A. C. Mazari, “Enhancing IoT Security with CNN and LSTM-Based Intrusion Detection Systems,” 2024.
7. H. Sharma *et al.*, “Intelligent Time Series Analysis for Intrusion Detection in IoT,” *Intelligent Computing*, 2025.
8. S.-M. Tseng, Y.-Q. Wang, and Y.-C. Wang, “Multi-Class Intrusion Detection Based on Transformer for IoT Networks Using CIC-IoT-2023 Dataset,” *Future Internet*, vol. 16, no. 8, 2024.
9. K. G. R. Narayan *et al.*, “IIDS: Design of Intelligent Intrusion Detection System for Internet-of-Things Applications,” 2023.
10. A. Diaf, A. A. Korba, N. E. Karabadji, and Y. Ghamri-Doudane, “Beyond Detection: Leveraging Large Language Models for Cyber Attack Prediction in IoT Networks,” 2024.
11. S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144, 2016.
13. Canadian Institute for Cybersecurity (CIC), “CICIoT2023 Dataset,” University of New Brunswick, 2023.
14. A. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, “IDS-INT: Intrusion Detection System Using Transformer-Based Transfer Learning for Imbalanced Network Traffic,” *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, 2024.
15. J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann, 1993.
16. B. Sharma, L. Sharma, C. Lal, and S. Roy, “Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach,” Expert Systems with Applications, vol. 238, Art. no. 121751, 202,.
17. B. Hulayyil, S. Li, and N. Saxena, “Explainable AI-based intrusion detection in IoT systems,” *Internet of Things*, vol. 31, Art. no. 101589, 2025.
18. Y. Wang, M. A. Azad, M. Zafar, and A. Gul, “Enhancing AI transparency in IoT intrusion detection using explainable AI techniques,” *Internet of Things*, vol. 33, Art. no. 101714, 2025.
19. K. P. Sharma, T. Nagpal, T. Vora, A. Yadav, M. I. Abdullah, B. Jayaprakash, A. Kashyap, G. Sridevi, A. Bhowmik, et al., “Interpretable intrusion detection for IoT environments using a self-attention-based explainable AI framework,” *Scientific Reports*, vol. 15, Art. no. 39937, 2025.
20. S. Sadhwani, A. Navare, A. Mohan, R. Muthalagu, and P. M. Pawar, “IoT-based intrusion detection system using explainable multi-class deep learning approaches,” *Computers and Electrical Engineering*, vol. 123, Art. no. 110256, 2025.
21. A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, “Survey of intrusion detection systems: Techniques, datasets and challenges,” *Cybersecurity*, vol. 2, Art. no. 20, 2019.
22. A. Khraisat and A. Alazab, “A critical review of intrusion detection systems in the internet of things: Techniques, datasets and challenges,” *Cybersecurity*, vol. 4, 2021.
23. S. Li *et al*., “Explainable AI-Based Intrusion Detection in IoT Systems,” *Internet of Things*, 2025.
24. A. L. Samed, “Explainable Artificial Intelligence Models in Intrusion Detection Systems,” *Engineering Applications of Artificial Intelligence*, 2025.
25. T. Hasan *et al*., “Real-Time Explainable IoT Security with Machine Learning,” *Ad Hoc Networks*, 2025.
26. K. P. Sharma *et al*., “Interpretable Intrusion Detection for IoT Environments Using Deep Neural Networks,” *Scientific Reports*, 2025.

27. M. Georgiades *et al*., “An Explainable AI Approach for Interpretable Cross-Layer Intrusion Detection,” *Electronics*, 2025.
28. A. Villafranca *et al*., “AI-Enabled IoT Intrusion Detection: Unified Conceptual Framework,” *Data*, 2025.
29. F. Ebrahimi *et al*., “Intrusion Detection in the Internet of Things Using Convolutional Neural Networks,” *Cybersecurity and Applications*, 2025.
30. V. Z. Mohale *et al*., “Evaluating Machine Learning-Based Intrusion Detection Systems with Explainable Artificial Intelligence,” *Frontiers in Computer Science*, 2025.