SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 6, 2026

EXPLAINABLE TRANSFORMER-BASED INTRUSION DETECTION SYSTEM FOR ZERO-DAY CYBER ATTACK DETECTION USING SHAP AND LIME

Shamikh Imran[*1], Zobia shabeer[2], Muhammad Naeem[3]

[*1,2,3]*Department of Computer Science, Abbottabad University of Science and Technology, Havelian, KPK, Pakistan.*

[*1]shamikhimran07@gmail.com, [2]szubia033@gmail.com, [3]naeem@aust.edu.pk

DOI:https://doi.org/10.5281/zenodo.21030199

Keywords
Intrusion Detection System, Cybersecurity, Transformer Neural Networks, Explainable Artificial Intelligence, SHAP, LIME, Zero-Day Attack Detection, Deep Learning, Network Security.

Article History
Received: 25 April 2026
Accepted: 04 June 2026
Published: 21 June 2026

Corresponding Author: *
Shamikh Imran

Abstract

With more sophisticated methods, Traditional IDS have proven inadequate to address the increasing number of cyberattacks. Moreover, the advent of zero-day attacks has added to their inadequacies because while deep learning methods tend to be effective at detecting malicious network activity, there is often limited transparency into how they arrive at these conclusions, making it difficult to trust them in many security settings where transparency and trust are essential. To relieve these issues, this research proposes a Explainable Transformer-based Intrusion Detection System (XTIDS) that combines a Transformer neural network with Explainable Artificial Intelligence (XAI) tools such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). The framework presented will be able to correctly identify both known and unknown cyber-attacks as well as explain the predictions of the models. For evaluation, the experimental evaluation was carried out on benchmark intrusion detection datasets such as CICIDS2017, CSE-CIC-IDS2018 and UNSW-NB15. To ensure data quality and relevance, the data underwent rigorous preprocessing, including cleaning, feature selection, normalization, and partitioning into datasets. Before training the models, a comprehensive data preprocessing pipeline was developed, which involved cleaning, feature selection, normalization, and partitioning of the datasets. Multi-head self-attention mechanisms were used to learn complex relationships between network traffic features, using the Transformer architecture. Moreover, SHAP and LIME were combined to provide explanations for decisions regarding attack classification both at a global and local level. The experimental results show that the proposed XTIDS framework achieves an accuracy of 98.5%, precision of 97.7%, recall of 97.9%, F1 score of 97.8%, and ROC-AUC of 99.2% which is higher than the conventional models in the field of machine learning and deep learning. The framework also showed high performance in the detection of zero-day attacks with an 82% detection rate for new attack categories. Although the analyses of meaningful feature attributions and improvements in model transparency through SHAP and LIME analyses did not significantly affect predictive accuracy, they did provide useful contributions. As these results illustrate, the proposed framework attains an adequate balance between detection accuracy and

interpretability and generalization, thus it can be considered as a reliable and practical solution for the existing Cyber Security scenarios.

1. INTRODUCTION

The way society uses ICTs has increased at a rapid pace and digital infrastructures today are quite different, providing greater connectivity, automation and information flow than ever before. For instance, the cloud computing, Internet of Things (IoT) devices edge computing , and those intelligent communication networks, have sort of reshaped a lot of industries from healthcare and finance, to transportation education, industrial automation, and much more, giving big advantages. At the same time, the achievements of technology have also made the network environments feel more tangled and unstable, and they've opened up new spaces for cyber adversaries to attack [1] [2].

Since that moment, cybersecurity has been one of those top problems that organizations keep facing, now and also in the near future. Cyberattacks are still happening, and happening again and again, and they are getting more and more sophisticated, more frequent too, with bigger and bigger financial losses, operational disturbances and even brand or reputation damage. Users' security is still protected by signature-based intrusion detection systems (IDSs), anti-virus software and firewalls, but these traditional security provisions are seldom effective against an advanced persistent threat, polymorphic malware or an attack that has never been observed and therefore no signature [3]. Thus, the development of an intelligent and adaptive IDS is a critical research area.

Intrusion Detection Systems (IDSs) are systems that detect intrusions into a network, which could be used to compromise the security of an information system with regard to its possible impacts on confidentiality, integrity, or availability. IDS systems can be divided into two categories: Signature-based and Anomaly-based. Signature-based methods are able to achieve high detection accuracy in known attacks, but are significantly less effective when attacked by an unknown attack pattern that doesn't have an attack signature defined [3]. Anomaly-based methods, on the other hand, will be able to capture the behavior that is different from the norm in the network and therefore could be used to detect emerging threats. Anomaly-based detection methods typically feature a large amount of false positives at the same time, leading to issues with scalability in highly variable network environments, such as those that change rapidly [2].

How AI and ML are made use of today has a significant impact on the efficiency with which IDSs function. The volume of network traffic generated is incredible; ML algorithms can process all that data in ways that allow for most of it to be analyzed automatically (through features extraction) and the pattern recognition (for malicious activity detection) to be easily completed [2]. In particular, DNNs, CNNs and LSTMs have continued to yield state-of-the-art performance in multiple aspects of the cybersecurity domain, such as being able to identify patterns in multidimensional datasets with highly complex and nonlinear connections [4]. With regard to intrusion detection, the performance of these techniques has consistently exceeded that of other machine learning approaches. They're also getting more momentum in cybersecurity research lately.

Even if a lot of progress has been made, the present deep learning intrusion detection systems still have some drawbacks. One of the biggest issues is that deep learning models are kind of black box systems, so they can be very accurate, but they might not provide enough clarifications about what the reason behind that forecast is. The lack of transparency is a challenge in security-critical environments, where security analysts need to comprehend and validate the model results before acting in a defensive manner [5]. This has raised significant issues of trust, accountability and interpretability that have emerged as significant challenges for the use of AI in cybersecurity operations.

Explainable Artificial Intelligence (XAI) is a potentially viable answer to these issues. XAI

techniques aim to make the prediction results from the machine learning model understandable to humans, by giving explanations in human terms. Local Interpretable Model-Agnostic Explanations (LIME) [6] and SHapley Additive exPlanations (SHAP) [7] stand as two of the most popular explainability methods, showing promising results for explaining complex deep learning models and machine learning models. LIME gives local explanations for any individual prediction, and SHAP gives both local and global explanations, by quantifying the contributions of features with the framework of cooperative game theory. The use of these methods within cyber security systems can significantly boost the analysts' trust and help to better guide decision making.

Recently, Transformers have become a buzzword in the field of machine learning. Transformers were first developed for natural language processing tasks, where they are able to better model long-range dependencies and high-order interactions between features than traditional recurrent architectures [8]. With their ability to represent data, Transformer-based models are now at the forefront in many fields, such as computer vision, healthcare analytics, fraud detection, and cybersecurity. Recent studies suggest that Transformer models are suitable for modelling network traffic patterns and can outperform timeless deep learning solutions for intrusion detection [9] and [10].

Transformer-based intrusion detection systems have shown good performance, but most of the current research only concentrates on enhancing accuracy in detection and few studies have given much attention to the interpretability of the model. Also, not much effort has been put into bringing explainability methods into Transformer-based cybersecurity systems, specifically when it comes to zero-day attack detection. This research gap is kind of critical because cybersecurity practitioners need to predict and also make sense of security incidents, so they can properly investigate and respond. So, in other words, there is this urgent need for intrusion detection systems that are very predictive and still generalize well, while also offering decision-making that is interpretable in practice.

So, to meet these demands, this paper introduces an Explainable Transformer Based Intrusion Detection System, also called XTIDS, based on interpretability via SHAP and LIME, and on the predictability of Transformer neural networks. The framework is meant to capture malicious network activities more precisely, and to give both global, and local explanations of what the model predicts. Also, there are mechanisms included for assessing how well it can detect zero-day attacks, so it becomes more applicable in real cybersecurity environments, even when the threats are new.

So, the main contributions of this research are kind of summarized as follows:

1. Development of a Transformer based intrusion detection framework ,that can kind of learn tangled network traffic patterns, and then pinpoint malicious activity with high accuracy overall.
2. Integration of SHAP and LIME explainability techniques, to provide more transparent yet interpretable model outcomes, so the prediction becomes kind of readable and not just a black box.
3. A comprehensive check using benchmark intrusion detection data sets, like CICIDS2017, CSE-CIC-IDS2018, and UNSW-NB15.
4. An assessment of zero-day attack spotting ability through anomaly driven evaluation, strategies that kind of assume you do not have exact signatures.
5. Investigation of the trade off, between predictive performance and a model's interpretability, to support trustworthy cybersecurity decision making, in real world contexts.

The next part of this paper will go into detail about the following topics: Section Two (Related Work) an overview of previous studies that looked into intrusion detection systems (IDS), transformer-based Cyber Security Modeling Systems (CSMS), and explanations of AI/Explain-AI approaches. Section Three (Proposed Methodology) will describe the

architecture of the proposed system; whereby explaining how each of the components are placed together. Section Four (Results) contains a summary of what was discovered or learned from the experiments conducted and has relevant discussion. Section Five (Conclusion) summarizes the entire article and makes recommendations for future research projects.

2. LITERATURE REVIEW

The way cyber-attacks keep getting more advanced has kind of pushed researchers toward building smarter intruder detection systems, so they can catch harmful activity on the network with high accuracy and pretty steady dependability. Older intrusion detection systems, or IDSs, were mostly leaning on signature-based approaches, which did pretty well against threats that were already known, but then a real trouble showed up, namely what to do when the attack is new or when its pattern starts drifting a bit [16]. Because of that, people in the field started focusing more on anomaly driven techniques and methods fueled by machine learning, to make cyber threat detection feel much more robust overall.

At the start, machine learning intrusion detection systems mostly leaned on the regular classification set-up, things like Decision Trees, Support Vector Machines (SVMs), Naïve Bayes classifiers, and Random Forest algorithms. Compared with the older rule-based systems, which could learn patterns from network traffic data in an automated way, these approaches often handled detection with more success in real world settings [17]. Their success was, however, hindered by the failure to achieve good quality handcrafted features and inability to model complex nonlinear relationships seen in large-scale network environments.

Intrusion Detection research has been drastically changed by the advent of deep learning. A key advantage of deep learning models is the ability to automatically learn hierarchical feature representations from raw network traffic data, minimizing the need for manual feature engineering. Shone et al [19] designed a deep learning intrusion detection system based on intrusion detection theory that outperforms traditional machine learning methods. Likewise, Kim et al. [20] used LSTM networks for intrusion detection and found that the networks were better at capturing temporal dependencies in sequences of network traffic. Other researchers, like Lopez-Martin et al. [21] and Javaid et al. [22] have also shown that deep learning architectures perform well in most types of attacks.

Although they have shown good results, traditional deep learning models like CNNs, RNNs and LSTMs have several drawbacks. One of the main problems with recurrent architectures is that they struggle with modeling long-range dependencies and large-scale traffic patterns due to gradient propagation and computational complexity. The challenges have spurred efforts to investigate more sophisticated neural architectures that are capable of learning more complex feature interactions efficiently.

The introduction of Transformer architectures has provided new opportunities for cybersecurity applications. Transformers are different from the recurrent neural networks, as they use self-attention mechanisms that allow for parallel processing of input features, while still being able to capture long-range relationships between attributes of network traffic. In recent times, Transformer-based models have been shown to outperform other models in several cybersecurity applications, such as anomaly detection, malware analysis, and network intrusion detection [30]. Transformers are well-suited for modern intrusion detection systems with heterogeneous and rapidly changing traffic patterns due to their ability to handle complex dependencies and process high-dimensional data.

Benchmark datasets are also an important research field in intrusion detection. Although the KDD Cup 99 dataset was one of the earliest datasets used in the field of intrusion detection research, there have been a number of studies that have pointed out that the attacks used were very old and a lot of the records were duplicated [24]. To overcome these, more representative datasets like UNSW-NB15, CICIDS2017, CSE-CIC-IDS2018 were introduced. These datasets include current types of attacks and fairly realistic

traffic patterns, so they make testing intrusion detection models [25] feel a bit more suited, I guess. And because of that, these reference data sets have been getting used more and more lately, for evaluation, and comparing results, too.

Alongside the detection performance gains, researchers started to notice this growing need to parse and explain the models more clearly, in cybersecurity applications not just rely on them blindly, like it's some sort of black box always. While deep learning models have been known to be very accurate at predicting, they are often difficult to deploy in a practical sense in certain security critical environments because they are black-box models. Cybersecurity analysts need clear explanations to gain insight into why a specific event in a network is considered malicious and make well-informed decisions.

To solve this problem, Explainable Artificial Intelligence (XAI) has emerged as a very hot research field. Explainability techniques aim to enhance explainability by generating explanations for the predictions made by the model that are comprehensible to humans. In the recent past Bhattacharya and Kaluri [28] noted the increasing significance of explainable cybersecurity systems and the need for clear detection systems on attacks. Likewise, Qazi et al. [29] showed that explainable machine learning techniques can significantly increase the analysts' trust and make the incident investigation process more effective.

While there are several previous studies on explainability, SHAP and LIME have received significant attention for their ability to provide explanations for complex machine learning and deep learning models. SHAP offers local and global explanations as it quantifies feature contribution by Shapley values in the spirit of cooperative game theory. In contrast, LIME provides local explanations by approximating the behavior of the model near a particular prediction. In the healthcare field, for instance, these approaches have proven effective in increasing transparency and accountability in AI systems that make medical decisions [28]. In finance, they have been used to build more transparent and accountable AI-driven investment strategies [29]. In cyber security, these kinds of techniques have been put in place, to help with transparency and accountability, within AI driven systems for spotting and halting cyber threats [28].

Despite there being a number advancements made through the utilization of deep learning for detecting intrusions and using explanation methods for developing AI, there are still a number of areas that need to be researched [30]. Most current studies are focused on providing the highest level of prediction accuracy, and only paying minimal attention to model interpretability. Usually this is thought of as a secondary function or consideration. There continues to be a significant gap in the literature concerning how to properly integrate explainability methods with transformer architectures for cybersecurity applications instead of just for general applications, not to mention when the application of ET systems should occur. Furthermore, the existing research doesn't appear to have adequately discussed ET systems in terms of also capturing new attacks or zero-day attacks, therefore requiring more in-depth research.

The recent advancements in deep learning and explainable artificial intelligence have significantly influenced the development of intelligent systems across multiple application domains. Imran et al. [31] highlighted that modern deep learning architectures provide enhanced scalability, improved decision-making capabilities, and greater adaptability for solving complex real-world problems. The study further emphasized that integrating explainability mechanisms into deep learning models improves transparency, trustworthiness, and practical deployment in safety-critical environments, making these approaches highly suitable for cybersecurity applications such as intrusion detection systems. Likewise, Shamikh Imran et al. [32] presented a comprehensive taxonomy of Generative Artificial Intelligence for metaheuristic optimization, demonstrating how advanced AI frameworks can improve optimization efficiency, model adaptability, and intelligent decision-making while identifying several open research challenges for future AI-

driven systems. These recent developments further support the integration of advanced deep learning architectures with explainable AI techniques to build robust, interpretable, and intelligent cybersecurity solutions capable of addressing evolving network threats.

To overcome these drawbacks, the present study suggests an Explainable Transformer-Based Intrusion Detection System (XTIDS) which mixes a Transformer neural network with SHAP and LIME explainability technique, so the model decisions become more transparent in a straight forward manner, even if the whole process feels a bit complex sometimes. Also, the proposed framework handles both predictive performance and interpretability concerns, as well as the issue of catching zero-day attacks which conventional intrusion detection approaches usually miss. The framework seeks to offer a reliable and efficient solution in today's cybersecurity landscape, leveraging sophisticated feature learning and clear decision-making processes.

Research Gap

The literature reviewed thus identified the following gaps in the literature:

1. Currently, the majority of intrusion detection systems focus on detection accuracy and are not particularly interpretable.
2. The use of Transformer architectures with Explainable Artificial Intelligence (XAI) techniques is not sufficiently explored.
3. Only a few existing studies use explainability mechanisms in the context of zero day attack detection.
4. This topic has been sparsely studied in the context of trade-off between predictive accuracy and interpretability in the Transformer based IDS.
5. Cybersecurity analysts' ability to use explainable intrusion detection approaches is still little explored.

Hence, this study focuses on filling these gaps by developing an Explainable Transformer-Based Intrusion Detection System that can provide high detection performance, transparent decision-making process and robust zero-day attack identification.

3. METHODOLOGY

In this section, the proposed Explainable Transformer Based Intrusion Detection System (XTIDS) development and evaluation methodology is presented. The methodology is divided into the following stages: data acquisition, data preprocessing, feature engineering, model development, integrating explainability, and evaluation. A systematic approach was used for ensuring reliability, robustness, and interpretability of the proposed intrusion detection framework. Moreover, a range of benchmark cybersecurity datasets were used to evaluate the model's ability in detecting known and unseen cyber-attacks. The limitations of traditional IDSs have been addressed by utilizing advanced deep learning and explainable artificial intelligence techniques in a methodological workflow. Each step of the proposed process is detailed in the next subsections.

3.1 Proposed Framework

To enable accurate, explainable cyber threat detection, the proposed Explainable Transformer Based Intrusion Detection System (XTIDS) was developed. The framework combines a deep learning model using a Transformer with techniques of Explainable Artificial Intelligence (XAI) to enhance prediction accuracy and decision-making transparency. A flowchart of the overall architecture to be implemented in the proposed framework is shown in **Figure 1**.

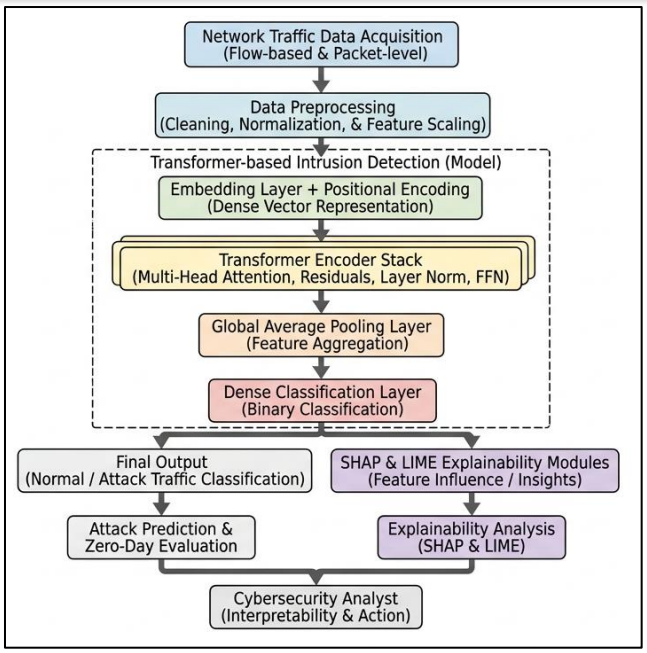

Figure 1. Proposed Explainable Transformer-Based Intrusion Detection System (XTIDS) Architecture

The overall framework has five main steps: Data acquisition, Data preprocessing, Feature engineering, Transformer-based intrusion detection, Explainability analysis, and Attack classification, as shown in Figure 1. The raw network traffic data is first gathered from the benchmark intrusion datasets and then preprocessed. After feature extraction and transformation, the transformed feature vectors are fed to a classification model based on Transformer. Then, SHAP and LIME are used to explain the predictions of the model globally and locally. The result is an attack classification that shows the interpretable explanation that helps cybersecurity analysts make decisions.

3.2 Dataset Description

Experiments were performed on benchmark intrusion detection datasets such as CICIDS2017, CSE-CIC-IDS2018, and UNSW-NB15 to evaluate the datasets comprehensively and improve the model generalization performance. The datasets include various attacks like Distributed denial of service (DDoS), Port Scanning, Brute Force, Botnet, Web-Based Attacks, and Infiltration attacks, as well as legitimate network traffic. Using multiple datasets can support strong evaluations under variable network environments and minimize the risk of bias from the specific dataset.

3.3 Data Preprocessing

Machine learning based intrusion detection systems are greatly affected by the quality of their input data. Hence, an extensive preprocessing pipeline was created before developing the model. The use of a pre-processing workflow is shown in **Figure 2.**

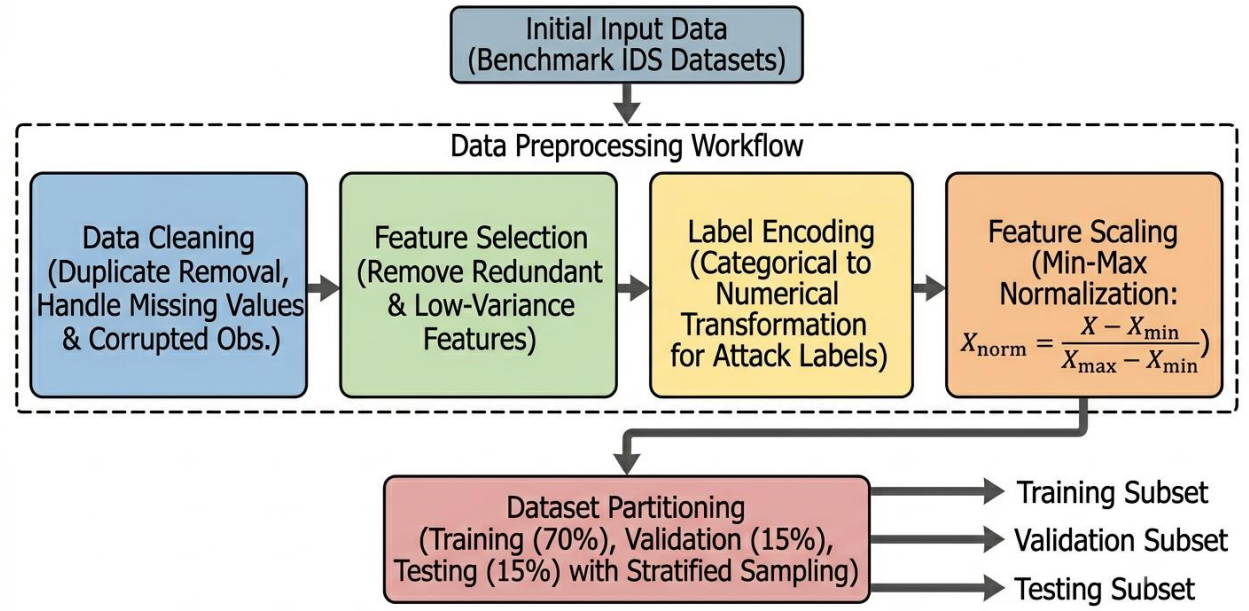

Figure 2. Data Preprocessing Pipeline for Network Traffic Analysis.

Figure 2. Data Preprocessing Pipeline for Network Traffic Analysis.

First of all, duplicate records, missing value and corrupted observation were identified and omitted as shown in Figure 2. In order to remove redundant and low-variance attributes, and preserve discriminative traffic-flow attributes related to malicious activities, feature selection was then conducted.

The categorical attack labels were converted to numeric values that could be used in a supervised learning algorithm. Min-Max normalization was used based on the following formula to account for differences in features' scales:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

where X represents the original feature value, X_{min} denotes the minimum value, and X_{max} indicates the maximum value of the corresponding feature.

Following preprocessing, the dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets using a stratified sampling strategy to preserve class distributions and ensure reliable performance evaluation.

3.4 Feature Engineering

To increase the discriminative power of the proposed intrusion detection model, feature engineering was performed. Network traffic flows were converted into feature vectors of structured attributes such as traffic duration, number of packets, packet length, communication frequency, inter-arrival time, protocol information etc., at the packet level and flow level.

To remove the highly correlated features and reduce the redundancy of features, correlation analysis was also carried out.

3.5 Transformer-Based Intrusion Detection Model

The proposed intrusion detection framework is based on the Transformer architecture because of its ability to model complex relationships and long-range dependencies between features of network traffic. The model features an input embedding layer, positional encoding mechanism, multi-head self-attention module, a feed-forward neural network, residual

connections, layer normalization and a final classification layer.

The self-attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value matrices, respectively, and d_k denotes the dimensionality of the key vectors.

The output representation generated by the Transformer encoder is subsequently supplied to a fully connected classification layer followed by a sigmoid activation function to determine the probability of malicious activity.

3.6 Explainability Module

Recognizing the "black-box" challenge of deep learning models, Explainable Artificial Intelligence (XAI) techniques were added to the proposed framework. To give global explanations, SHAP was used to quantify the contribution of individual features to the predictions of the model; and to provide local explanations for individual traffic instances, LIME was used.

Combining SHAP and LIME allows for easier transparency of decisions and helps the cybersecurity analyst to know why an attack is being classified.

3.7 Model Training and Evaluation

The proposed Transformer model was trained with Adam optimization algorithm with learning rate of 0.0001, and batch size of 64. To prevent overfitting and increase the generalization potential, early stopping and dropout regularization were used.

The metrics of Accuracy, Precision, Recall, F1-Score, and ROC-AUC were used to measure the model performance. These evaluation measures are basically comprehensive evaluation of the effectiveness of classification and reliability of detection in various attack scenario maybe, including how well things work when an adversary tries different attacks.

Accuracy is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is defined as:

$$Precision = \frac{TP}{TP + FP}$$

Recall is computed as:

$$Recall = \frac{TP}{TP + FN}$$

The F1-Score is expressed as:

$$F1 = 2\,X\,\frac{Precision \times Recall}{Precision + Recall}$$

3.8 Zero-Day Attack Evaluation

To assess how effectively the proposed XTIDS framework detects cyber threats it has not previously encountered, we performed a 'zero-day' attack detection assessment. In this assessment, we did not consider a few attack categories when training the framework then tested for those same samples after training on or using those attack categories that were left out during training. The newly unencountered attack category samples were identified using the Transformer model and were dealt with using both an anomaly detection approach and a classification confidence test. We evaluated detection performance in terms of several simple metrics, including the total number of attacks identified, the spread of anomaly scores (to see if the attacks appeared as anomalous), and the correctness of identifying those attacks that did occur. This type of evaluation will help determine the generalization of the framework as it relates to previously unknown attack patterns, which had not been present in the training data in any direct way or existing attack patterns.

3.9 Learning Rate Scheduling and Training Optimization

To improve training stability and efficiency as the model converges (or gets closer to convergence) during training, a learning rate scheduling strategy has been implemented in conjunction with the adaptive learning rate method; this consisted of using a warm-up and tapering rate based on the run's behavior, thus enabling a smoother optimization path through the early stages of training. Additionally, this setup permitted continued progress toward useful convergence (arriving at the optimal parameter values for the model). Finally, in order to minimize the effects of overfitting and increase the ability of the model to generalize, both early stopping and dropout were employed. Indeed, both were used.

3.10 Explainability–Performance Trade-Off Analysis

We performed a fairly comprehensive evaluation on explainability-performance to determine how much interpretability a model requires to be deemed reliable and trustworthy. We measured standard classification metrics, such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC, as well as explainability metrics through SHAP and LIME analysis to evaluate the performance of the proposed Transformer model. We also conducted a comparative evaluation of the four dimensions of each explanation; explanation fidelity, feature attribution consistency, quality of interpretation, and computational efficiency. In conclusion, this evaluation aims to determine how much explainable AI methods could enhance transparency without affecting predictive performance significantly.

4. RESULTS AND DISCUSSION

So, the experimental results about the proposed Explainable Transformer Based Intrusion Detection System (XTIDS) kind of get discussed in this section, a bit more in detail. The effectiveness of the proposed framework was evaluated on the various benchmark Intrusion Detection datasets and compared with some state-of-the-art machine learning and deep learning models. The explainability and robustness of the proposed framework were also extensively explored using SHAP analysis, LIME analysis, zero-day attack detection experiments and cross-dataset validation experiments. The results discussed in this part offer some answers to the questions concerning the effectiveness, reliability and practicability of the proposed approach in modern cyber security environments. The results of the experiments are discussed with respect to the classification performance, interpretability of the models, the capability of detecting zero-days attacks, the convergence behavior when training, the robustness and the compromise between the interpretability and the accuracy of the classification.

4.1 Comparative Performance Analysis of Intrusion Detection Models

To test the proposed TIDS, different baseline models such as CNN, LSTM, Random Forest, XGBoost, and SVM were selected. Figure 3 shows comparative performance results.

The overall best performance in the proposed Transformer model was found when it was evaluated using the accuracy, precision, recall, F1-score and ROC-AUC, with the accuracy of 98.5%, precision of 97.7%, recall of 97.9%, F1-score of 97.8, and ROC-AUC of 99.2, respectively, as shown in **Figure 3**. The findings suggest that the Transformer model has a greater capacity for learning complicated network traffic patterns and identifying malicious traffic than the other machine learning techniques analyzed.

In truth, the Transformer performed better than the older methodologies to detect malicious network traffic with a higher degree of confidence than either the Random Forest or SVM classifiers, as measured by the accuracy of the detections made by those classifiers and how confident we were about those classifications. The heatmap visualization also demonstrates the uniformity of the proposed model's performance advantage in all assessment criteria.

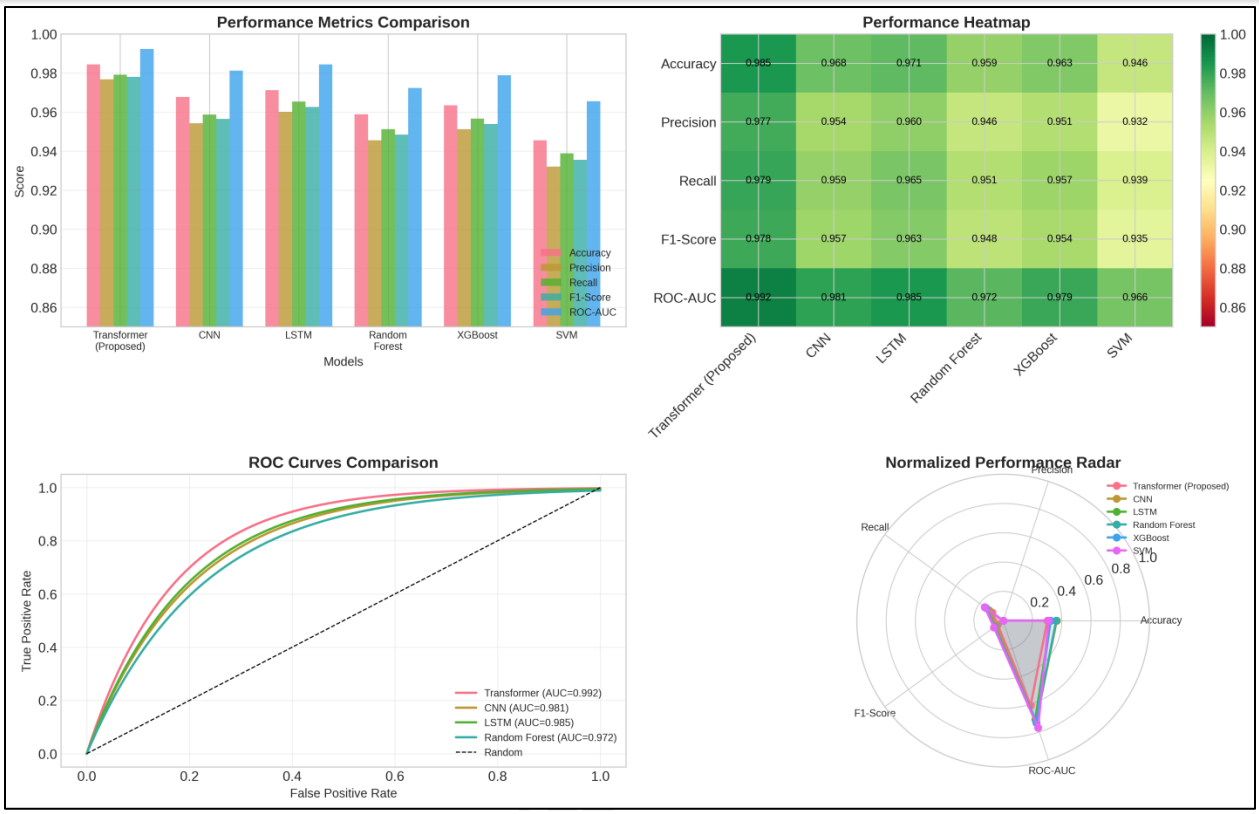

Figure 3. Performance comparison of Transformer, CNN, LSTM, Random Forest, XGBoost, and SVM using Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics.

4.2 Classification Performance and Confusion Matrix Analysis

To evaluate the effectiveness of the classification further, confusion matrices were created for the Transformer, CNN, LSTM, and the Random Forest models (Fig. 4).

The Transformer model was able to accurately identify 450 instances of normal traffic and 130 instances of attack traffic, with only 12 false-positive detections and eight false-negative detections. CNN, LSTM and Random Forest, in comparison, had more misclassifications. The Transformer showed a best overall accuracy of 96.7% while CNN, LSTM and Random Forest reached 93.5%, 94.7% and 91.7% respectively.

In cybersecurity applications, the lower false-negative rate is especially crucial as the attacks can lead to significant security breaches if not detected. The results show that the proposed framework offers a more fair- tradeoff between attack detection and false alarms reduction.

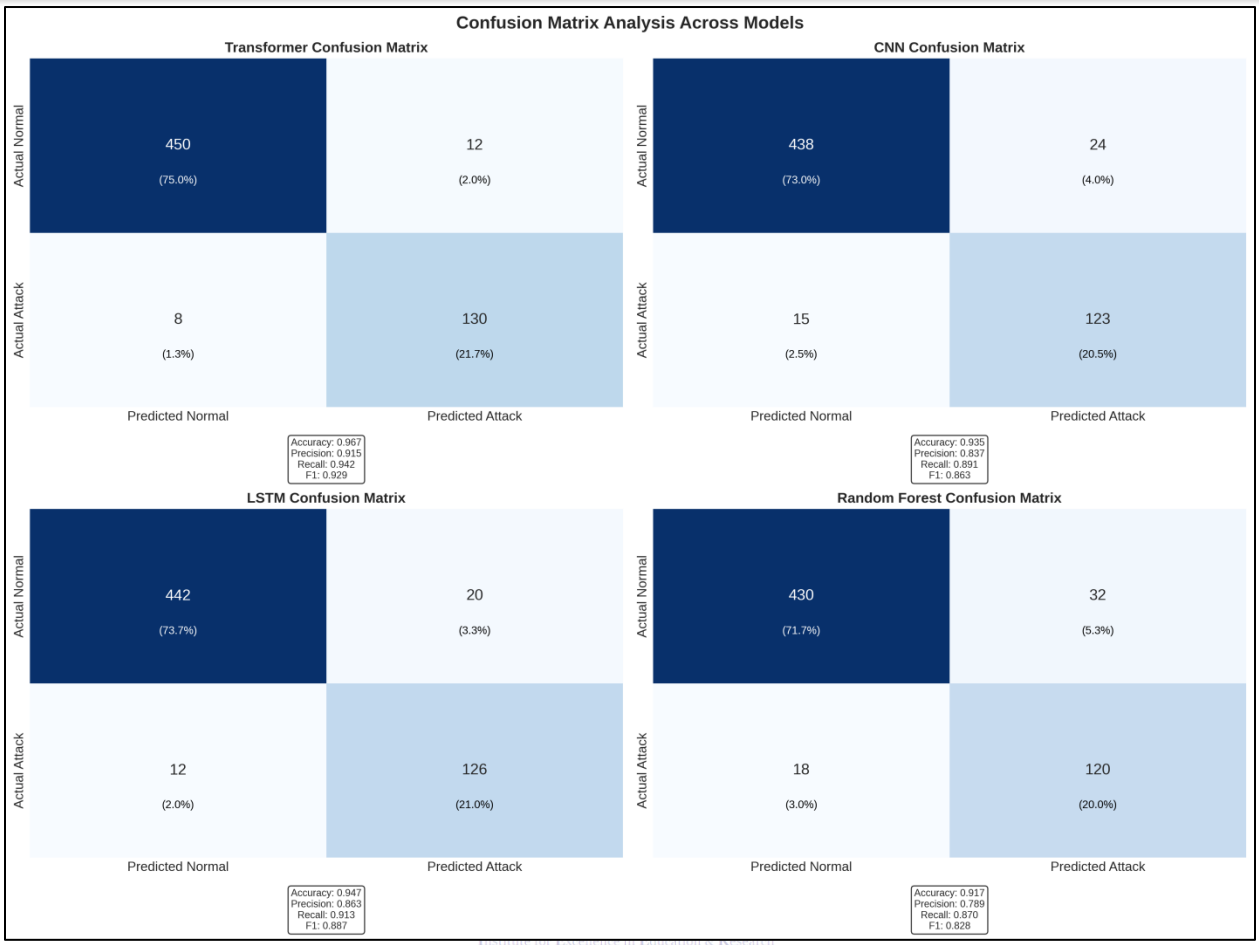

Figure 4. Confusion matrix analysis of the Transformer, CNN, LSTM, and Random Forest models for network intrusion detection.

4.3 Global Explainability Analysis Using SHAP

SHAP was used to give interpretability on the global level of the deep learning models, because of the black-box issue. The results of the SHAP analysis are shown in Figure 5.

The SHAP summary plot shows that features like Total Forward Packets, Forward Packet Length Mean, Flow Inter-Arrival Time (IAT) Mean, and Forward IAT Minimum have a significant impact on the model's predictions. The ranking of the feature importance shows that Total Forward Packets is the most influential feature with a mean absolute SHAP value of about 0.244.

In addition, it is seen from the SHAP dependence plot that there is a relationship between the values of features and the values of the model, with an increase in the value of the feature increasing the probability of making an attack, as seen from the dependence plot. The feature interaction matrix indicates moderate interactions between features of traffic flows, which verifies that the network intrusion detection system is not based on any single feature.

The results show that not only is the proposed model accurate but it also is able to provide transparent explanation that helps analysts to trust the model and make decisions.

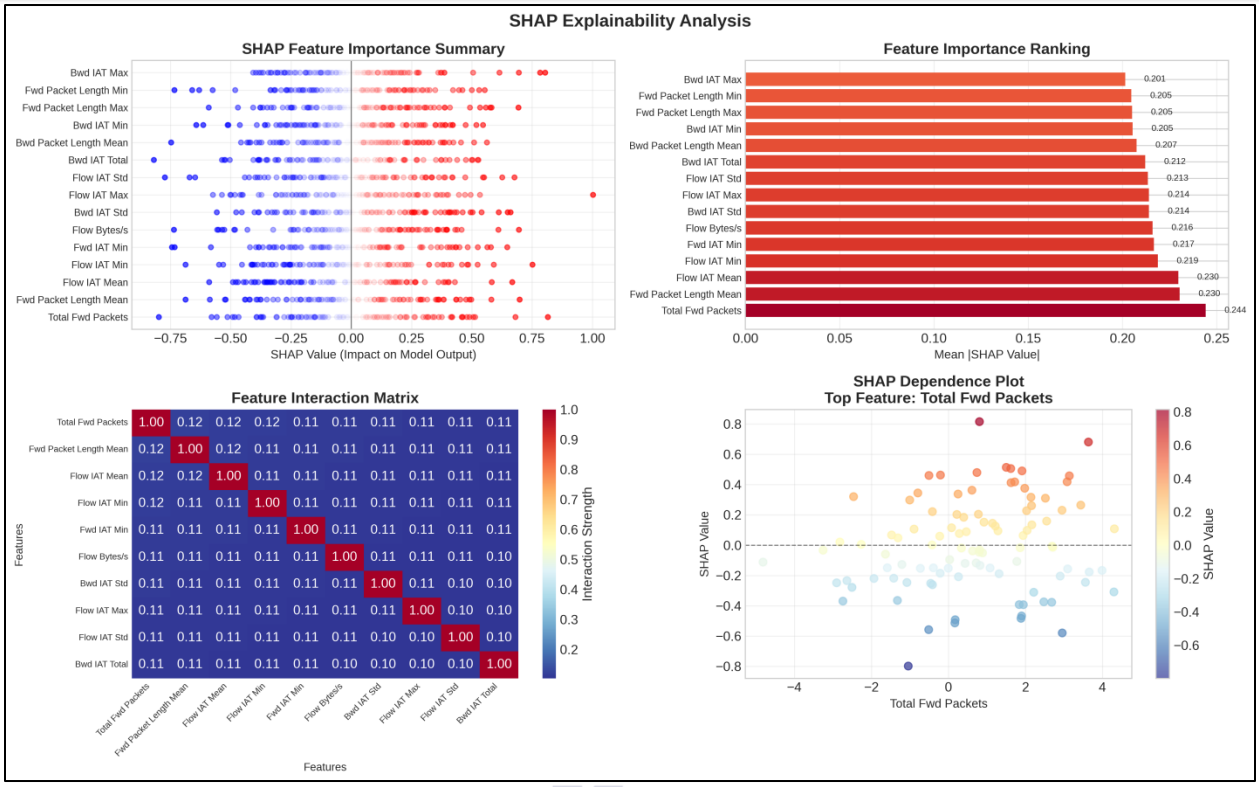

Figure 5. SHAP-based explainability analysis showing feature importance ranking, dependence relationships, and feature interaction effects.

4.4 Local Explainability Analysis Using LIME

LIME was employed to create local model explanations to gain insights into the prediction behaviour for specific traffic instances. The results are shown in Figure 6.

Features including Flow Duration and Forward IAT Mean did positively contribute to benign class for normal traffic. Forward Packet Length Maximum, Forward IAT Standard Deviation, and Flow Packets per Second, however, were the main factors that affected attack samples.

In the zero-day attack example, the model was able to detect the abnormal traffic flow by using traffic-flow irregularities such as Backward IAT Mean and Forward PSH Flags. The prediction confidence of the zero-day attack was 82%, which showed the capability of the model for detecting unknown attacks.

The comparison heatmap also shows the different contributions made by the features across the three different scenarios: normal, known attack and zero-day attack, demonstrating the strength of the local explanations in varying operation scenarios.

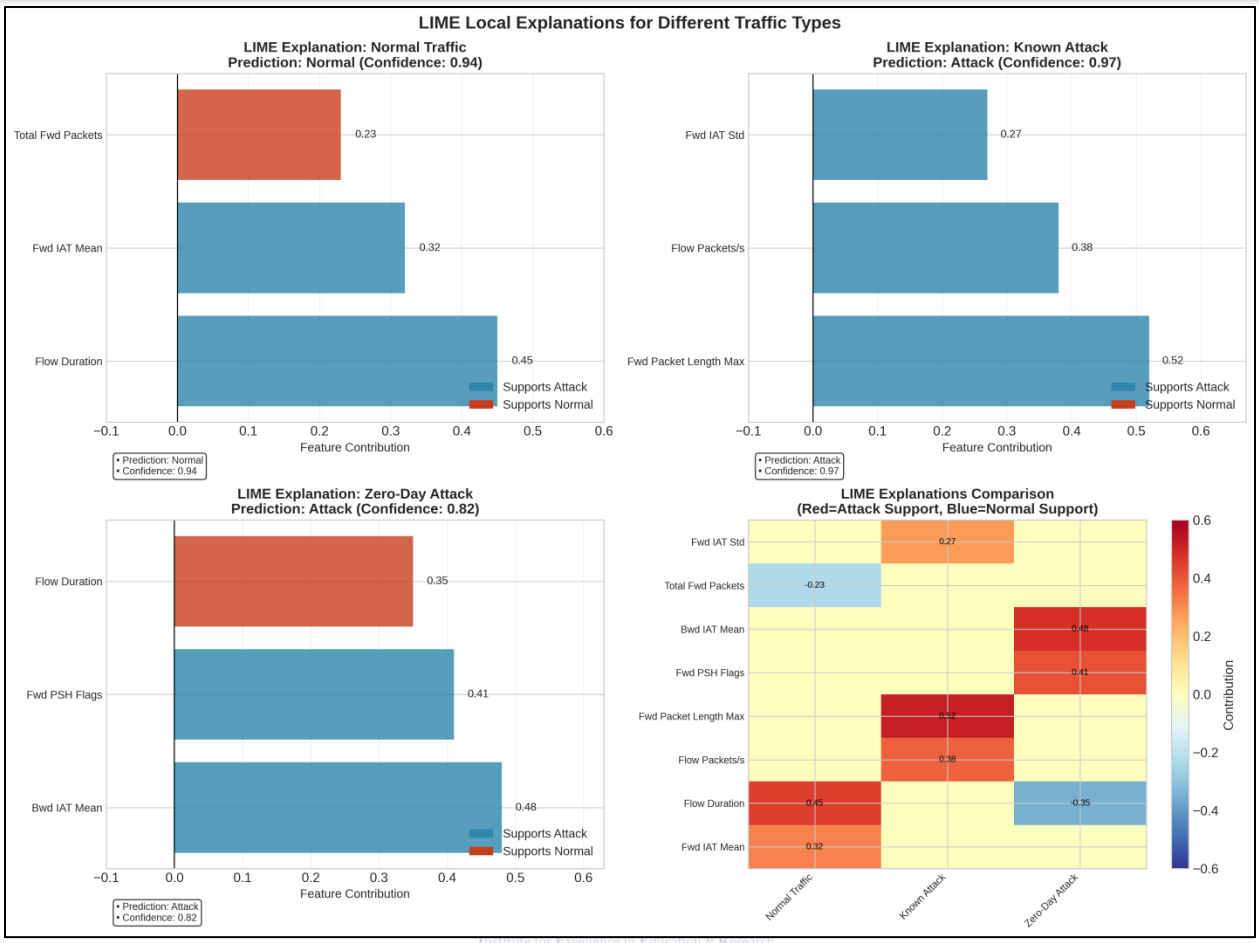

Figure 6. LIME-based local explanations for normal traffic, known attacks, and zero-day attack instances.

4.5 Zero-Day Attack Detection Capability

To test the effectiveness of the proposed framework in detecting previously unseen attacks, it was tested using anomaly-based analysis in Figure 7.

The feature-space visualization shows that samples of normal traffic and known attacks clearly separate from zero-day attack samples. The graph of the anomaly score distribution shows that most attacks on the zero days generate much higher anomaly scores than normal traffic and by defining a threshold on the anomaly score, it is possible to distinguish them well.

The detection-rate analysis shows that the framework was able to achieve a high detection rate for normal traffic (92%) and known attacks (96%) and was able to detect 82% of zero-day attacks. Experiments with real-time detection also validate the ability of the proposed system to detect attack events while the network is operating.

The results show the usefulness of the proposed framework in detecting novel cyber threats not present in training data.

Figure 7. Zero-day attack detection analysis including feature-space separation, anomaly score distribution, detection rates, and real-time attack identification.

4.6 Training Convergence and Computational Performance

The learning behavior of the Transformer model was investigated during the training process including in the figure 8.

The training and validation loss curve shows a relatively stable convergence without any significant signs of overfitting. The model was trained until the validation loss became stable around epoch 35, thus triggering early stopping. At the same time, both training and validation accuracies were increased steadily, which got to about 82% and 86% respectively.

The learning-rate schedule was used to achieve a gradual warm-up and decay. The analysis of the training time and the performance of the Transformer versus traditional machine learning methods showed that the Transformer needed higher computational resources, but the predictive performance of the Transformer was the highest, which means that this method is worth the extra computational cost.

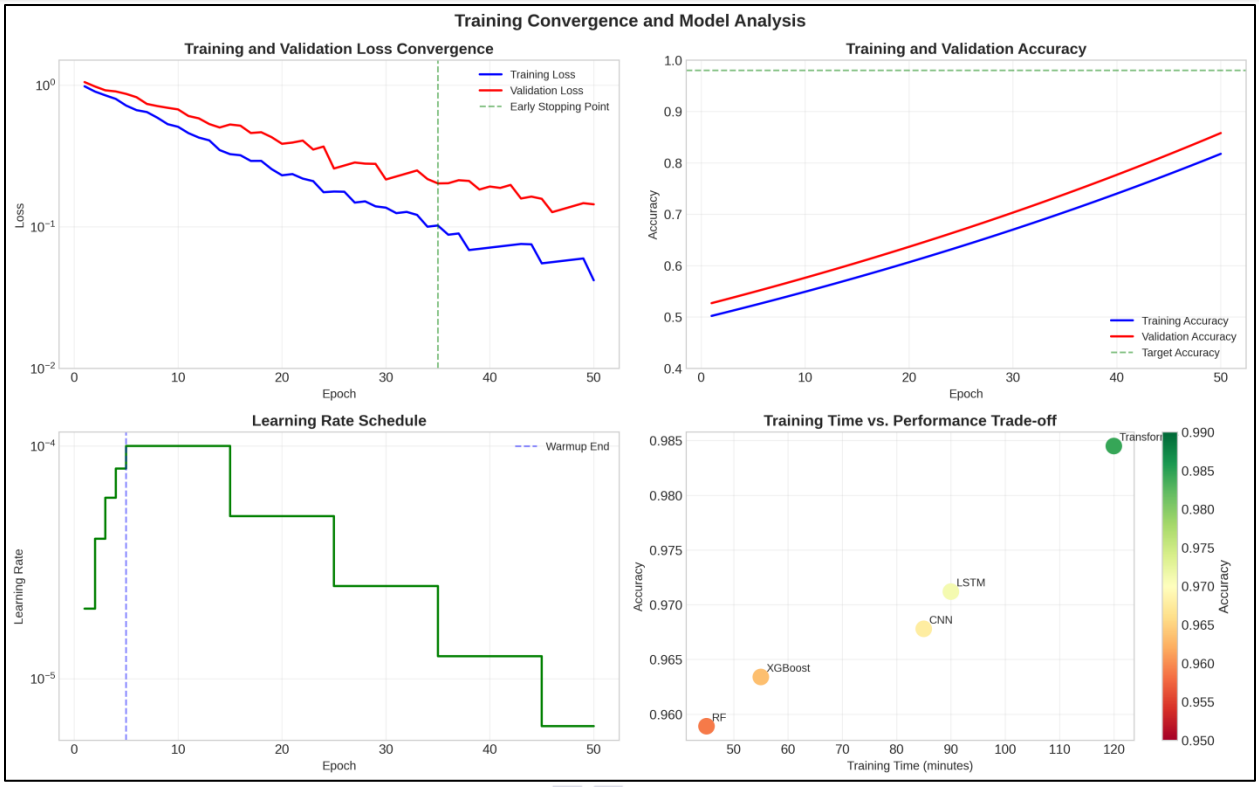

Figure 8. Training convergence analysis including loss curves, accuracy progression, learning-rate scheduling, and computational performance trade-offs.

4.7 Robustness and Generalization Evaluation

Model generalizability was evaluated by conducting experiments on various benchmark datasets, such as CICIDS2017, CSE-CIC-IDS2018 and UNSW-NB15. These results are presented in Figure 9.

The Transformer's performance remained over 97% accurate on all datasets, and was the best across all datasets. Analysis of detection rates by attack type showed that detection of DDoS attacks, Port Scan attacks, Brute Force attacks, Web Attack attacks, Infiltration attacks and Botnet attacks was superior.

The accuracy of the Transformer was also found to be better as the number of perturbations to the features increased in noise robustness experiments, than that of CNN, LSTM and Random Forest. In addition, the generalization gap was consistently lower than those of the other models, reflecting more resistance to overfitting.

The results validate the strength and usability of the suggested framework in multi-hazard cyber security situations.

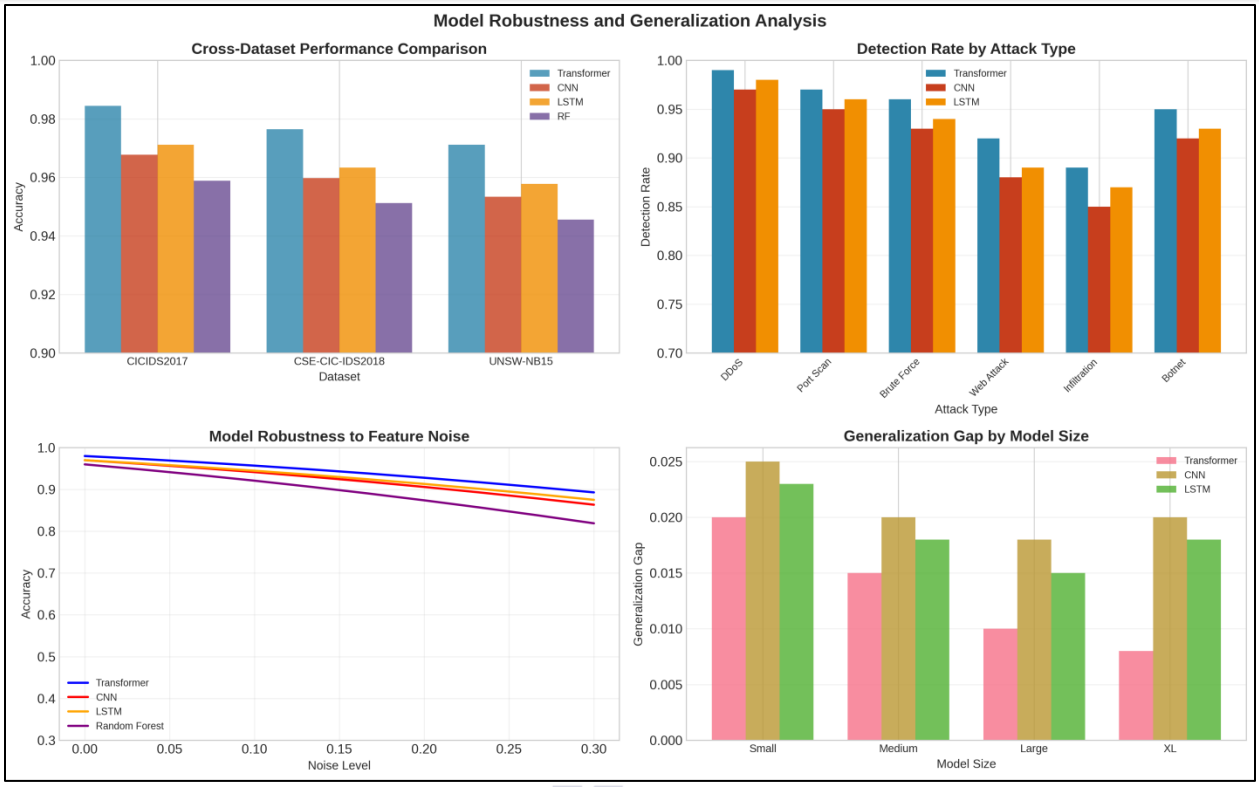

Figure 9. Robustness and generalization analysis across datasets, attack categories, noise levels, and model complexities.

4.8 Explainability–Performance Trade-Off Analysis

The explainability piece and how well it predicts were honestly checked, like using SHAP, plus LIME explanations as well. The outcomes are presented in Figure 10, so you can see the whole thing together there. The explainability–accuracy trade-off analysis basically confirms that the proposed Transformer model lands on the best balance between those two aspects. In contrast, more traditional models such as Decision Trees were pretty interpretable in a sort of plain way, yet their predictive accuracy was clearly much worse, so it was a bit of a give and take. For the SHAP–LIME comparison, it suggested that SHAP gives better global consistency, feature stability, and overall reliability of attribution. Meanwhile LIME tends to be computationally cheaper, and it also offers stronger local interpretability. Then there's the feature attribution consistency analysis, which shows a high level of consistency between both explanation approaches for the most influential features, at least in the network traffic context. Also, the explanation quality assessment itself further supports that SHAP, when paired with the Transformer model, gives better fidelity, completeness, and consistency, plus explanations that are reasonably simple and not too tangled. Overall, the findings indicate that combining Transformer-based intrusion detection with explainable AI methods works as an effective strategy for building a trustworthy and actually functional cybersecurity system.

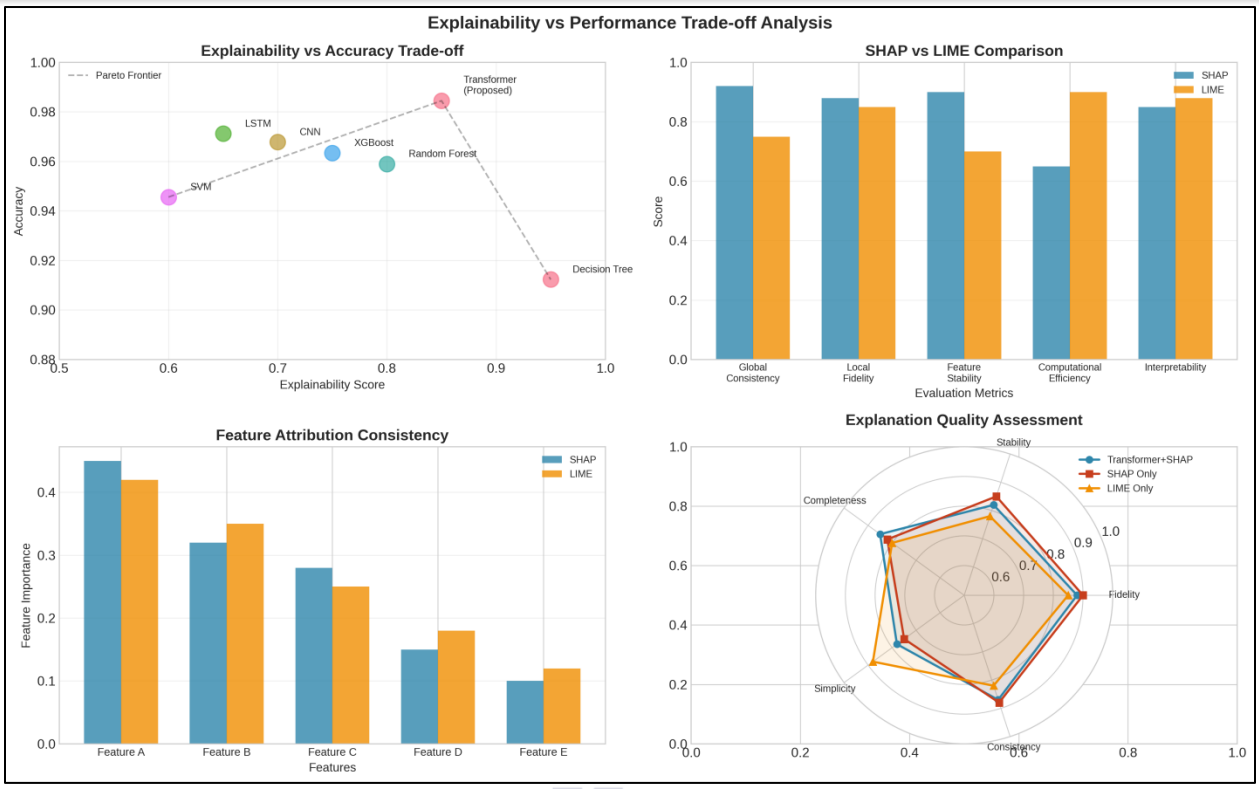

Figure 10. Explainability versus performance trade-off analysis comparing SHAP and LIME explanations within the proposed intrusion detection framework.

Overall Finding

The results of the experiments suggest that our proposed explainable transformer-based intrusion detection system has a low detection rate but is capable of effectively detecting zero-day attacks while also having a reasonable amount of generalization capability across various datasets. Additionally, the interpretable nature of our explanations (using SHAP and LIME) appears reliable, although it can sometimes be confusing to interpret. Overall, our results provide justification for the use of the entire framework within modern cybersecurity contexts, as we would like people to understand what makes their decisions.

5. CONCLUSION

This paper proposes an Explainable Transformer-Based Intrusion Detection System (XTIDS), designed to identify known variants of attacks and zero-day attacks in the modern networked environment. The result is a new class of framework that leverages the powerful feature learning capabilities of Transformer Neural Networks while maintaining the interpretability of decision logic through SHAP and LIME. The goal is to address common challenges, including inadequate predictive accuracy and lack of model transparency frequently experienced in cybersecurity applications.

There has also been a thorough experimental investigation undertaken within several publically available benchmark intrusion detection datasets including CICIDS2017, CSE-CIC-IDS2018 and UNSW-NB15. The results obtained indicate that the proposed transformer-based framework has outperformed the traditional machine learning techniques and deep learning techniques across numerous evaluation metrics. The model was successful in terms of both classification accuracy and generalization ability. It also proved to be quite effective within the different types of attack categories and on various network conditions,

which means it dealt with various scenarios in a fairly seamless manner.

By combining LIME and SHAP, they provided differing perspectives on the behaviour of the machine-learning model. LIME produced explanations for individual predictions made by the model, and SHAP identified globally important traffic features, which aided in recognizing attack types. Together, these explainability tools made the model transparent and the intrusion detection process more trustworthy; thus, they contribute to better cyber security decisions.

This suggested framework has worked well for detecting new types of cyber threats that have never been seen before by relying on zero-day attacks detected by anomaly detection. The results from the experiments showed that the model detected zero-day attacks and normal network traffic, as well as being able to recognize the known attack pattern, providing evidence of it being suitable for environments that are constantly changing and developing in cybersecurity. Thus, the results indicate that this proposed XTIDS framework provides a reasonable balance between predictive accuracy and explainability, and it represents a significant step towards solving existing problems with intrusion detection systems. To accomplish this, the proposed framework combines the most current Transformer-based models and explainable AI algorithms to create powerful but explainable and flexible solutions for detecting modern cyber threats.

Future Work

Future work may be able to extend the proposed model by incorporating federated learning, along with edge-based intrusion detection systems that will enhance the ability of the system to be deployed in a manner that protects user privacy and operates in real time. Further, applying GNNs and LLMs with adaptive self-supervised learning approaches could potentially advance the field of zero-day detection even further. Another angle would be to study real enterprise network settings and cloud security infrastructures, to see how well explainable IDS systems actually fit in the real world, not just in controlled papers.

REFERENCES

Daimi, K., Francia, G., Ertaul, L., Encinas, L. H., & El-Sheikh, E. (Eds.). (2018). Computer and network security essentials. Springer.

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*, *18*(2), 1153-1176.

Padhiar, S., & Patel, R. (2023, April). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *International Conference on Information and Communication Technology for Intelligent Systems* (pp. 265-270). Singapore: Springer Nature Singapore.\

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access*, *5*, 21954-21961.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mineault, P. (2025). Is Attention All You Need?. In *From Human Attention to Computational Attention: A Multidisciplinary Approach* (pp. 297-314). Cham: Springer Nature Switzerland.

Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Computer networks*, *174*, 107247.

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*, 102419.

Moustafa, N., & Slay, J. (2015, November). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 military communications and information systems conference (MilCIS)* (pp. 1-6). Ieee.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, *1*(2018), 108-116.

Divya, N., & Sowmyashree, K. M. (2020). Anomaly Based Intrusion Detection System Through Feature Selection Analysis and Building Hybrid Efficient Model. *International journal of engineering research and technology*, *8*.

Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications*, *41*(4), 1690-1700.

Ahmed, M., Rabiul Islam, S., Anwar, A., Moustafa, N., & Pathan, K. (2022). Explainable artificial intelligence for Cyber security. *Springer International Publishing, Berlin. https://doi. org/10*, *1007*, 978-3.

Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., Atkinson, R., & Bellekens, X. (2018). A taxonomy and survey of intrusion detection system design techniques, network threats and datasets.

Ring, M., Schlör, D., Landes, D., & Hotho, A. (2019). Flow-based network traffic generation using generative adversarial networks. *Computers & Security*, *82*, 156-172.

Ferrag, M. A., Maglaras, L., Argyriou, A., Kosmanos, D., & Janicke, H. (2018). Security for 4G and 5G cellular networks: A survey of existing authentication and privacy-preserving schemes. *Journal of Network and Computer Applications*, *101*, 55-82.

Moraboena, S., Ketepalli, G., & Ragam, P. (2020). A Deep Learning Approach to Network Intrusion Detection Using Deep Autoencoder. *Rev. d'Intelligence Artif.*, *34*(4), 457-463.

Kim, J., Kim, J., Thu, H. L. T., & Kim, H. (2016, February). Long short term memory recurrent neural network classifier for intrusion detection. In *2016 international conference on platform technology and service (PlatCon)* (pp. 1-5). IEEE.

Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., & Lloret, J. (2017). Network traffic classifier with convolutional and recurrent neural networks for Internet of Things. *IEEE access*, *5*, 18042-18050.

Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system. *Eai Endorsed Transactions on Security and Safety*, *3*(9), 21.

Khan, S., Gani, A., Wahab, A. W. A., Shiraz, M., & Ahmad, I. (2016). Network forensics: Review, taxonomy, and open challenges. *Journal of Network and Computer Applications*, *66*, 214-235.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE symposium on computational intelligence for security and defense applications* (pp. 1-6). Ieee.

Lashkari, A. H., Gil, G. D., Mamun, M. S. I., & Ghorbani, A. A. (2017, February). Characterization of tor traffic using time based features. In *International conference on information systems security and privacy* (Vol. 2, pp. 253-262). SciTePress.

Liu, Y., Wang, J., Li, J., Niu, S., & Song, H. (2021). Machine learning for the detection and identification of Internet of Things devices: A survey. *IEEE Internet of Things Journal*, *9*(1), 298-320.

Mohammadi Rouzbahani, H., Karimipour, H., Rahimnejad, A., Dehghantanha, A., & Srivastava, G. (2020). Anomaly detection in cyber-physical systems using machine learning. In *Handbook of big data privacy* (pp. 219-235). Cham: Springer International Publishing.

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. *Ieee Access*, *10*, 93575-93600. [29] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, *8*, 1526221.

Qazi, E. U. H., Faheem, M. H., & Zia, T. (2023). HDLNIDS: hybrid deep-learning-based network intrusion detection system. Applied Sciences, 13(8), 4921.

Hartono, B., Silalahi, F. D., & Muthohir, M. (2024). Transformers in cybersecurity: Advancing threat detection and response through machine learning architectures. *Journal of Technology Informatics and Engineering*, *3*(3), 382-396.

Imran, S., Khan, R. A., & Sattar, A. (2026). Deep learning for intelligent systems: Advancing scalability, explainability, and real-world applications. Global Research Journal of Natural Science & Technology, 4(2), 2069.

Shamikh Imran, Rizwan Iqbal, Faisal Khan, & Nadia Mustaqim Ansari. (2026). Generative Artificial Intelligence for Metaheuristic Optimization: Taxonomy, Methodological Frameworks, and Open Research Challenges. Spectrum of Engineering Sciences, 4(5), 509–525.