PENINSULA
PUBLISHING LLC

SHIFRA
Vol. **(2025)**, 2025, **pp**. 69–80
ISSN: 3078-3186

Research Article

XAI-IDS: A Transparent and Interpretable Framework for Robust Cybersecurity Using Explainable Artificial Intelligence

Said Salloum[1,*], Sajedeh Norozpour[2]

[1] *School of Science, Engineering and Environment, University of Salford, United Kingdom, UK*

[2] *Istanbul Gelişim University, Istanbul, Türkiye*

ARTICLE INFO

Article History

Received 1 Nov 2024

Revised: 20 Dec 2024
Accepted 20 Jan 2025
Published 8 Feb 2025

Keywords

Explainable Artificial Intelligence for Intrusion Detection Systems (XAI-IDS),

Distributed Denial-of-Service (DDoS),

Convolutional Neural Network (CNN),

Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME).

ABSTRACT

The rapid evolution of cyberattacks, and in particular DDoS attacks, has outpaced many intrusion detection systems (IDSs) that fail to be interpretable or transparent, which subsequently hinders wider adoption in real-world high-stake environments. To our knowledge, this is the first joint utilization of CNNs with both SHAP and LIME for XAI-IDS tasks, particularly based on comparative evaluations across a wide range of different intrusions and a combination of XAI methods. The proposed method starts with data acquisition and data preprocessing of NSL-KDD dataset as only DDoS labeled records output are chosen and this is followed by data cleaning through data normalization, feature selection and label encoding to get accurate results. The dataset is divided into training and testing sets, a 1D CNN is trained on the dataset to differentiate DDoS attacks from normal traffic using the hyperparameters with optimized values and early stopping. It yields high predictive performance with 94% accuracy, 93% precision, 95% recall, and 94% F1-score. For these results, it demonstrates strong classification ability with very few false positive and false negative. Stack SHAP for global feature importance and LIME for individual predictions to give us human-understandable explainable results of the model. Not only does this dual explainability cultivate trust and accountability, it also enhances auditing and compliance in sensitive industries. Overall, XAI-IDS shows that the combination of deep learning and post-hoc interpretability is a promising approach to the design of trustworthy systems for cybersecurity. Future works will be conducted toward real-time deployment and multi-class detection in federated and edge learning frameworks.

1. INTRODUCTION

With the rapid changing environment of cyber-attack, the main objectives of Intrusion Detection Systems (IDSs) is to uncover misuse, abuse, or other malicious vectors on the computer network infrastructures that are perpetrated by either internal users or external attackers [1,2]. Traditional IDS techniques work on the premise that malicious activities will have a behavioral model significantly different from legitimate network behavior, enabling them to be detected. But with the exponential growth of cyber intrusions and the sophistication of current attack methodologies, the classical rules-based and anomaly-based methods alone are becoming inadequate. Such increase in threats complexity has increased the demand for using advanced Artificial Intelligence (AI) techniques to improve the IDS performance and adaptivity [3,4].

Prior work in this area has examined a range of AI methods for countering network threats, including statistical anomaly detection [5,6], rule-based misuse detection [7,8], and hybrid approaches operating in experimental systems [9,10]. However, the increase in complex, distributed, and stealthy attacks, especially Distributed Denial-of-Service (DDoS), has made it increasingly difficult to maintain detection accuracy and interpretability. As attackers are using obfuscation methods and heterogeneous environments to compromise targets, AI-based IDSs need to not only detect them but also explain them so that timely and appropriate responses can be made [11–15]. In this regard, various AI models have been proposed such as Artificial Neural Networks (ANNs) [16,17], Support Vector Machines (SVMs) [18,19], Decision Trees (DTs) [20–22], Naive Bayes (NB) [23,24], and Random Forests (RF) [25–27]. Though providing extremely high detection accuracy, these models are usually acting as black-box systems in other words, with few or no explanations of the logic behind their decisions. This obscurity complicates trust, accountability, and verification, particularly when it comes to the most critical cybersecurity domains.

*Corresponding author email: s.a.s.salloum@edu.salford.ac.uk

DOI: https://doi.org/10.70470/SHIFRA/2025/004

The focus of most prior works is on the accuracy of classification without addressing the interpretability/transparency of models' decisions. They often neglect the importance of attack-specific features and do not reuse global and local explainability approaches; thus, the model insights do not enable the analyst to comprehend how to identify certain types of attacks. [28] Consequently, XAI is gaining increased importance with respect to augmenting classical IDSs such that high performance is guaranteed along with interpretability and trust. The increasing number and severity of cyberattacks, with a special focus on IoT and SCADA systems, emphasize the use of either intelligent or explainable IDSs [29,30]. State-of-the-art approaches demonstrate that real-time intrusion detection can be achieved effectively using deep learning techniques [31-33], however, in high-assurance contexts the lack of interpretability of such approaches renders these solutions insufficient. In this context, XAI offers two benefits. First of all, it establishes trust between human security analysts and the AI systems by providing them with understandable evidence that justifies the predictions made. Secondly, it increases transparency, enabling stakeholders to recognize the reasons behind model outputs thereby connecting accuracy to accountability. But the effectiveness of XAI hinges on the quality of training data and accurate feature labeling.

To address this challenge, in this paper, we propose an end-to-end Explainable AI (XAI) framework specifically designed for robust DDoS detection in network environment, termed as XAI-IDS. In addition to this, our framework integrates deep learning with Convolutional Neural Networks (CNNs), and XAI techniques (SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)) providing global and local interpretability of detection results. Unlike previous methods that utilized various datasets, we conduct our study using only the well-known NSL-KDD dataset [34] by filtering the DDoS attack records for the training and testing stages. Data acquisition, preprocessing, CNN-based classification, and explainability all methodological stages are included in the proposed XAI-IDS framework. It enables cybersecurity analysts not only to know whether a denial-of-service attack has taken place but also to understand why it thinks that way. Such a visual, explainable and interpretable IDS solution comes with robust performance measures, closes the important gap between performance and explainability in the IDS literature, and ensures that the security solutions are transparent, cannot be misinterpreted, and are ready for real-world deployment.

2. RELATED WORK

Artificial Intelligence (AI) has been an emerging field in recent years and particularly for Intrusion Detection Systems (IDS) in cyber security domain. Signature-based or statistical anomaly detection techniques were the mainstay for traditional IDSs, which, while successful against known threats, frequently had trouble detecting new or evolving attacks. In order to overcome these constraints, researchers have resorted to the application of AI techniques such as support vector machines, decision trees, and deep neural networks to enhance detection precision as well as scalability [34].

While AI-based IDSs have been proven to outperform others with promising results, the lack of interpretability remains an issue. Such gap has contributed to the rise of Explainable Artificial Intelligence (XAI) which aims to close the gap between performance and transparency, especially for critical systems in domains such as cybersecurity [35]. Some of the most studied XAI techniques include SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which allow for global and local understanding of model behaviour. These techniques are being increasingly employed to explain black box decisions in intrusion detection, enhancing the understanding and trust of security experts in automated systems [36].

Several studies suggested AI-based IDSs using deep learning, such as CNNs and RNNs to identify patterns from traffic data. Among them, CNNs have shown to be efficient in leveraging spatial relationships between features in structured datasets including NSL-KDD and CICIDS-2017 [37]. While these models often achieve high classification accuracy success rates, they fail to reveal useful interpretable results, especially in cases of adversarial attacks or when false positive results are produced.

In addition, most of the approaches available in the literature aim at detection accuracy at the cost of interpretability. Few studies have investigated what features contribute to detection decisions, or how users can reverse-engineer why a particular input is classified as malicious. Many studies that do apply XAI methods shallowly implement these methods without thorough integration of interpretability through model pipeline [38]. However, there is only few XAI-based IDS frameworks focused specifically on DDoS detection, despite the high incidence and magnitude of DDoS attacks [39].

This work addresses these issues by proposing an extensive and explainable XAI-IDS that combines a deep learning CNN model with SHAP and LIME to identify Distributed Denial of Service (DDoS) attacks. Our framework aims to incorporate interpretability within the design and evaluation pipeline of the system, which sets it apart from existing studies that are posthoc in nature towards XAI. The objective is not just to predict malicious behavior accurately but to explain predictions sufficiently that cybersecurity analysts can comprehend what the model is doing and why, and act meaningful on predictions.

In fact, Table I summarizes and conducts a significant analysis of the most prominent research studying AI- and XAI-based intrusion detection systems. It outlines the methods implemented, datasets used, if explainability was considered, the main contributions of each method, and the principal limitations noted. Most of the studies proposed yet focus mainly on the effective intrusion detection of deep learning models like CNNs and RNNs, but these studies have not yet provided an integrated explanation. While others have explored post-hoc XAI methods like SHAP and LIME, they have done so without any integration in the decision pipeline itself. Moreover, most researches still use generalized datasets and address multi-class attack detection but do not solve the issue with specific attacks such as DDoS. This gap highlights the need for a comprehensive framework such as our proposed XAI-IDS that integrates strong detection accuracy with model interpretability for DDoS detection specifically.

TABLE I: SUMMARY AND ANALYSIS OF RELATED WORKS ON AI AND XAI-BASED IDSS

Study / Method	Technique Used	Dataset	XAI Applied	Key Contribution	Limitation
[34]	ANN, SVM	NSL-KDD, KDD'99	No	Enhanced anomaly detection	Lacks interpretability
[35]	SHAP, LIME + ML	CICIDS-2017	Yes	Introduced global and local explainability	Post-hoc only
[36]	CNN + SHAP	NSL-KDD	Yes	Visualized feature contributions	Only global explanations
[37]	Hybrid ML Models	CICIDS	No	Improved accuracy for multi-class attacks	Black-box decisions
[38]	RNN, LSTM	NSL-KDD	No	Time-based anomaly detection	No feature insight
[39]	DNN + XAI	NSL-KDD, BoT-IoT	Yes	Interpretability in IoT attack detection	Limited to single dataset

3. METHODOLOGY

In this research paper, we present a transparent and interpretable cybersecurity architecture called XAI-IDS (Explainable Artificial Intelligence for Intrusion Detection Systems), that aims to accurately identify and explain DDoS attacks in a network setting. While black-box machine learning based techniques are shown to produce good accuracy at the expense of interpretation, it is the proposed hybrid technique of deep learning with explainable AI (XAI) tools which lends not only to efficient detection but also to reasoning that follows the perfect XAI. The main focus of XAI-IDS framework is to gain robustness and performance of DDoS detection and enhance trust, accountability and response to incidents by means of interpretability. As shown in Table II, we propose several sequential stages for our methodology.

TABLE II: METHODOLOGICAL STAGES OF THE PROPOSED XAI-IDS FRAMEWORK

Stage	Description
1st Data Acquisition	Loading the NSL-KDD dataset, focusing on selecting relevant features and records specific to DDoS attack scenarios.
2nd Filtering	Extracting only DDoS-related entries to reduce noise and target the detection task effectively.
3rd Preprocessing	Performing feature selection, normalization (e.g., Min-Max Scaling), and encoding labels into numerical format.
4th Data Splitting	Splitting the preprocessed dataset into training and testing sets, typically in a 70:30 or 80:20 ratio.
5th CNN Model Training	Designing and training a Convolutional Neural Network to detect patterns associated with DDoS traffic.
6th Model Evaluation	Using metrics like accuracy, precision, recall, and F1-score to assess the model's performance. Includes performance refinement through hyperparameter tuning and retraining if needed.
7th Explainability	Utilizing XAI methods (SHAP and LIME) to interpret the CNN's decisions, identify key contributing features, and explain predictions.

3.1 Dataset AcquisitionBottom of Form

We use the NSL-KDD dataset, which is a standard dataset in the network intrusion detection domain. NSL-KDD is an upgraded version of the KDD Cup 1999 dataset, where both redundant records and class imbalance issue are alleviated, as they damage the training of machine learning models when using the older versions. It consists of labelled instances of network traffic, including normal traffic and several types of malicious traffic, like DoS, Probe, R2L (Remote to Local) and U2R (User to Root) traffic.

We focus on Distributed Denial-of-Service (DDoS) attacks since they are one of the deadliest and common threats to network infrastructure for this research. Therefore, in order to narrow the analytical scope and increase the specificity of the detection model, we isolate and extract only the DDoS related records from the dataset. Focusing on DDoS traffic patterns allows us to reduce noise from random attack types, allowing the model to learn and generalize better based on inimitable characteristics of such attacks. Table III summarizes the main technical features of the NSL-KDD dataset used in this article.

TABLE III: NSL-KDD DATASET ATTRIBUTES

Attribute	Description
Source	NSL-KDD Dataset (Canadian Institute for Cybersecurity)
Structure	41 feature attributes + 1 class label (attack or normal)
Label Type	Categorical (normal, DDoS, Probe, R2L, U2R)
Data Volume	Includes both KDDTrain+ and KDDTest+ subsets
Selection Focus	Only DDoS-labeled records (e.g., ‘smurf’, ‘neptune’, etc.)

The extracted subset provides a well-defined basis for training and evaluating our proposed deep learning-based intrusion detection framework. This tailored dataset helps ensure that the developed model achieves higher accuracy and interpretability in detecting DDoS attacks, which are often characterized by specific temporal and volumetric patterns.

3.2 Data Preprocessing

This phase is a key step that helps in making the proposed XAI-IDS framework more effective. It converts the Raw NSL-KDD dataset into tabular format to feed into the Deep learning models. This phase receives less noise and higher quality data to output to the CNN model. Table IV explains the preprocessing steps applied in this study.

Table IV: Summary of Data Preprocessing Steps

Step	Technique	Description
Feature Selection	Domain knowledge and correlation analysis	Removes irrelevant/redundant features while preserving important attack indicators.
Normalization	Min-Max Scaling	Scales numerical feature values into the [0,1] range to improve training stability.
Label Encoding	One-hot or Label Encoding	Converts categorical labels into numerical format suitable for neural networks.

All the pre-processing steps are essential for the efficacy of the intrusion detection framework. It improves predictive performance by eliminating redundant attributes which reduces computational burden and avoids overfitting. Normalization prevents features with larger numeric ranges from dominating the model learning process and ensures all features contribute equally to it. Lastly, this encoding of categorical labels enable the CNN to handle and understand various types of attacks. Following these steps meticulously fosters the framework's aptitude for discerning significant patterns within the input data, as well as its proficiency in executing precise DDoS attack detection.

3.3 Data Splitting

To ensure unbiased model evaluation and reliable performance measurement, the preprocessed dataset is divided into two distinct subsets: 80% for training and 20% for testing. The training subset is used to fit the CNN model, allowing it to learn relevant DDoS patterns, while the testing subset serves as unseen data to evaluate generalization. Figure 1 diagram outlines the data flow from preprocessing to model evaluation.

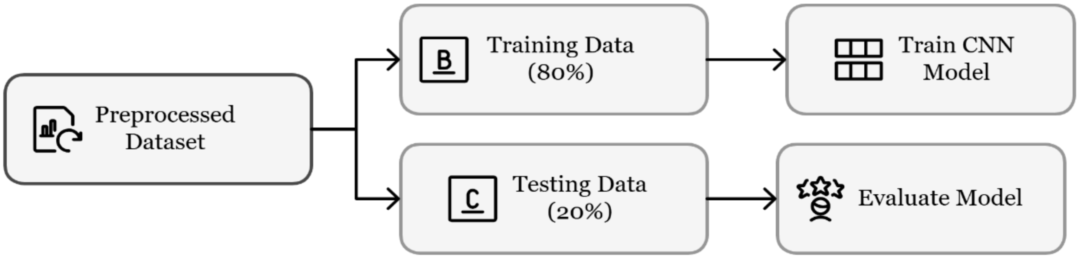

Fig. 1. Data Splitting, Model Training, and Evaluation Workflow.

3.4 Model Training: Convolutional Neural Network (CNN)

Specifically, the 1D Convolutional Neural Networks (1D CNN) is implemented to accurately detect Distributed Denial-of-Service (DDoS) attacks from talked tabular data. We have proposed the architecture specifically for learning spatial dependencies between feature dimensions of NSL-KDD dataset. The input layer of the CNN receives the feature vectors after preprocessing. These features are then forwarded through a series of convolutional layers that apply learnable filters to extract patterns relevant for detecting attacks. To perform down sampling and reduce dimensionality and computational efforts while mitigating overfitting, max-pooling layers are employed after every convolutional block.

After this, a flattening layer flattens the output into a one-dimensional vector, which is followed by a dense (fully connected) layer that learns abstract features. And then, at the output layer, the neurons are activated using a sigmoid function to classify whether the instance is a DDoS attack or not. The general structure of this model is illustrated in Figure 2. The configuration settings of each CNN layer are listed in Table 4, Training parameters like optimizer, loss function, learning rate, etc. are summarized in Table v. The CNNs were trained with a binary cross-entropy loss function, Adam optimizer, 32 batch size and for 50 epochs. We apply early stopping with a patience of 5 epochs to avoid overfitting and improve model generalization.

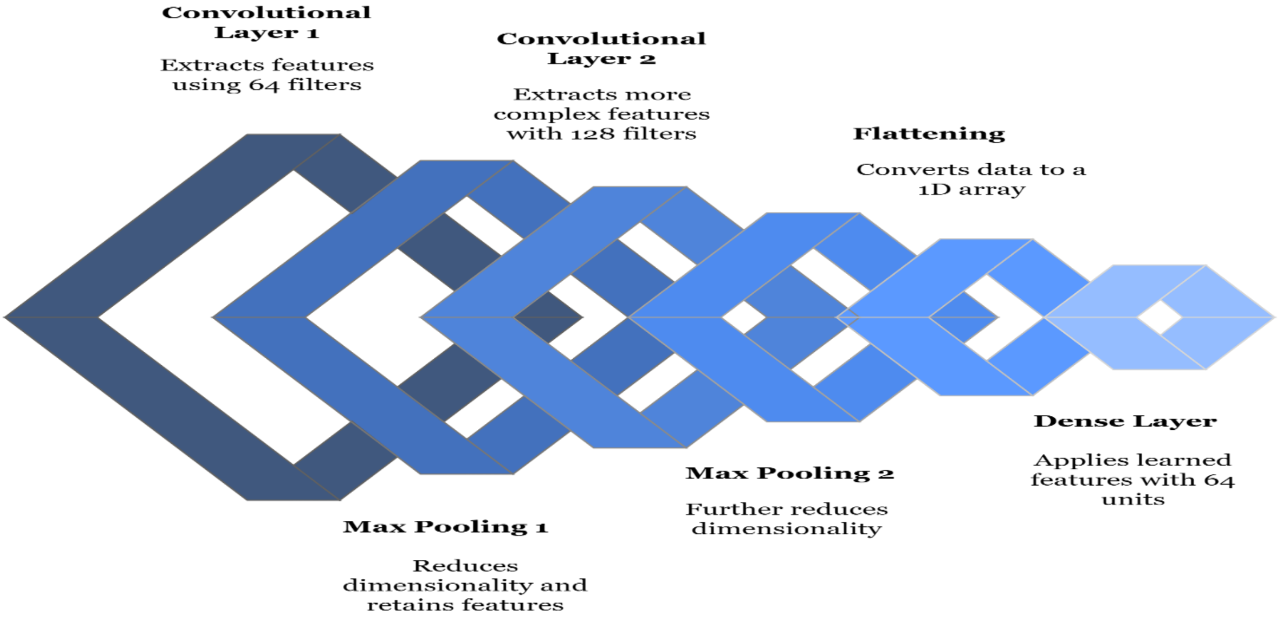
Convolutional Layer 1
Extracts features using 64 filters
Convolutional Layer 2
Extracts more complex features with 128 filters
Flattening
Converts data to a 1D array
Dense Layer
Applies learned features with 64 units
Max Pooling 2
Further reduces dimensionality
Max Pooling 1
Reduces dimensionality and retains features

Fig 2. Visual architecture of the CNN used in the XAI-IDS framework, showing the progressive flow from input to classification.

TABLE V. CNN ARCHITECTURE CONFIGURATION

Layer Type	Description
Input	1D vector of preprocessed features
Conv1D	64 filters, kernel size = 3, activation = ReLU
MaxPooling1D	Pool size = 2
Conv1D	128 filters, kernel size = 3, activation = ReLU
MaxPooling1D	Pool size = 2
Flatten	Converts 2D feature maps into 1D vector
Dense	64 neurons, activation = ReLU
Dropout	Dropout rate = 0.5 to prevent overfitting
Output (Dense)	1 neuron, activation = Sigmoid (for binary classification)

TABLE VI. CNN TRAINING SETUP

Parameter	Value
Loss Function	Binary Cross-Entropy
Optimizer	Adam
Batch Size	32
Epochs	50
Learning Rate	0.001
Early Stopping	Enabled
Patience	5 epochs

3.5 Model Evaluation

To evaluate the performance of the trained CNN model in detecting Distributed Denial-of-Service (DDoS) attacks, we compute four key classification metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a comprehensive understanding of the model's effectiveness, especially in scenarios where class imbalance may exist.

1. Accuracy measures the overall proportion of correctly classified instances, including both DDoS and normal traffic. It is defined as (eq. 1):

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

2. Precision quantifies the proportion of correctly predicted DDoS attacks out of all predicted positive instances. It is particularly important to minimize false positives and is calculated as (eq. 2):

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

3. Recall (also known as sensitivity) indicates the model's ability to correctly identify all actual DDoS attacks, helping to minimize false negatives. It is computed as (eq. 3):

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

4. F1-Score provides a harmonic mean of precision and recall, offering a balanced evaluation metric especially useful when dealing with imbalanced datasets. The F1-Score is given by (eq. 4):

$$F1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

In this context TP stands for True Positives, TN stands for True Negatives, FP represents False Positives and FN indicates False Negatives. These evaluation metrics, when looked at together, allow for a holistic and methodical evaluation of the model's dependability, discriminative ability, and general robustness in identifying DDoS attacks.

If the performance after the first assessment of the trained CNN model has not reached a certain score (usually equal to or greater than an F1-score of 0.85), then the refinement phase is initiated. 04. Evaluating the model this step helps ensures that the model is not only performing well on the training data but also generalizing well to unseen network traffic and detects DDoS attacks well.

Hyper-parameter tuning is the first part of such a process, where critical training parameters are tuned to aid model learning. Systematic variation of hyper-parametrs such as the learning rate, kernel size, dropout rate, the number of convolutional filters, and the batch size is undertaken. By doing so, these calibrations allow for better fitting of the model to the nuanced relationships that underlie malicious traffic, while avoiding the pitfalls of overfitting or underfitting. After the tuning step, the model is retrained using this newfound information. At this point, the CNN is repurposed on the original training dataset, taking advantage of the adjusted parameters to enhance pattern recognition and decision-making ability. It then retests the model with the same metrics of accuracy, precision, recall, and F1-score on the testing dataset after retraining. In the event the model still fails to exceed some minimally acceptable performance threshold, we repeat the cycle of tuning, retraining, and evaluation. The iterative process that ends up with a much more precise, stable and reliable detection model. The satisfactory performing optimized model is then directed towards the explain ability stage. In this subsequent stage (described in Section 3.7), Explainable AI techniques like SHAP and LIME are utilized to interpret and justify the model’s predictions, making clear the basis for detection and building trust in operational scenarios of cybersecurity.

3.6 Explainable AI Integration (XAI)

After reaching acceptable results with the CNN, we apply some techniques of explainable Artificial Intelligence (XAI) for interpreting the CNN model in terms of influencing factors. This provides humans understandable analysis (make decision) about how and why the model is interpreting a networks connection as DDoS attack. We utilize two popular XAI methods, namely SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to this end. SHAP uses cooperative game theory to calculate the contribution of every individual feature to the model’s final decision. This allows the features to be ranked according to influence, giving you a global view of your model’s behavior throughout all the samples.

Conversely, LIME emphasizes local interpretability and meets this requirement by learning a simple surrogate function locally around each individual prediction. This enables instance-level explanations to see which features drove a specific decision (attack, normal Traffic). Utilizing these XAI approaches allows not only validation of model fairness and reliability, but also helps build better user trust in the system, which is particularly importance in cybersecurity environments where accountability and transparency are key. In order to improve transparency and user trust, a suite of explainability tools is applied to the trained CNN model. Both SHAP and LIME (see Figure 3) are used to interpret predictions; SHAP provides global importance scores for each feature and LIME explains individual predictions by approximating the local decision boundary. A hybrid method like this enables cybersecurity analysts to deeply comprehend and verify the choices made by the XAI-IDS model.

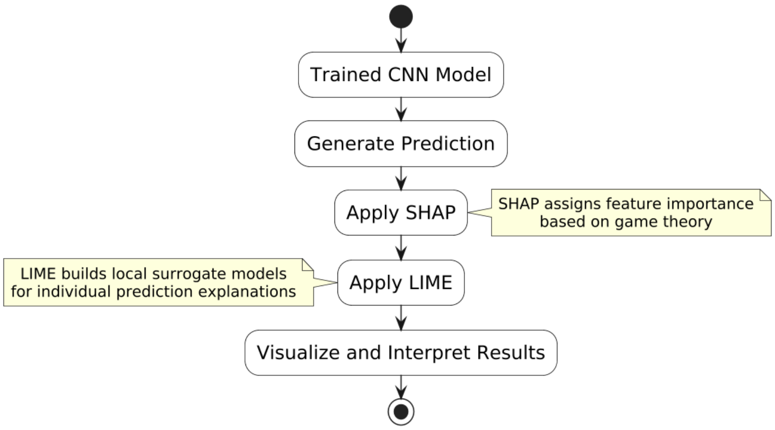
Trained CNN Model
Generate Prediction
Apply SHAP
SHAP assigns feature importance based on game theory
LIME builds local surrogate models for individual prediction explanations
Apply LIME
Visualize and Interpret Results

Fig 3. XAI Integration Flow in the XAI-IDS Framework

The entire mining steps of the proposed XAI-IDS pipeline are depicted in Figure 4. This bridges all paths from data collection and preprocessing, CNN model training and evaluation, to the application of explainable AI utilities using SHAP and LIME. The diagram above gives summary of the framework's end-to-end workflow comprising the interpretable DDoS attack detection framework.

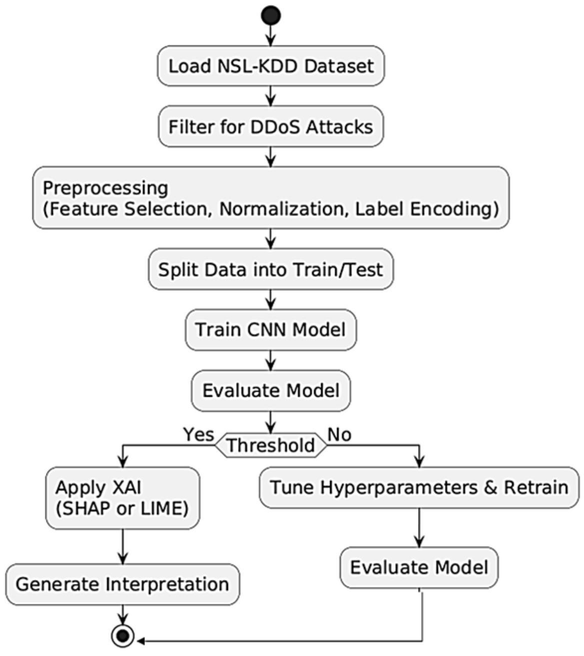

Fig 4. Complete Workflow of the Proposed XAI-IDS Framework

In conclusion, the XAI-IDS solution benefits from the performance of deep learning and the interpretability of explainable AI to deliver a DDoS detection mechanism with good accuracy and transparent model insight. The methodical approach of structured preprocessing, Congruent CNN training, iterative optimisation and post-trained XAI-based interpretation allows to produces a robust and programmable yet interpretable and trustworthy solution to modern cybersecurity challenges, with appropriate information learnt in a reasonable time-frame.

4. RESULTS AND DISCUSSION

In this section we describe the experimental results obtained for the proposed XAI-IDS framework that incorporates a CNN with Explainable Artificial Intelligence (XAI) techniques for the detection of Distributed Denial-of-Service (DDoS) attacks from the NSL-KDD dataset. The outcomes are interpreted across two important perspectives-- (1) the quantitative performance analysis of the trained CNN model, and (2) qualitative interpretable analysis through SHAP and LIME explanations.

The CNN model was trained on filtered data preprocessed to reduce the noise of non-DDoS instances. Standard classification metrics Accuracy, Precision, Recall, and F1-Score were computed on a reserved test set to measure performance. These metrics give a complete insight into how well the model can detect DDoS attacks with minimum false positives and false negatives. Besides the numerical results, this section also presents a confusion matrix and metric comparison plots to further substantiate the metrics. Additionally, the interpretability of model predictions is evaluated with SHAP (SHapley Additive exPlanations) [20] and LIME (Local Interpretable Model-Agnostic Explanations) [21]. These tools explain the most significant features that led the model to those decisions, on a global scale across the dataset as well as locally for specific samples. Thus, the knowledge gained through XAI findings improve transparency and justify the importance of the most important features used by the network to differentiate benign behavior from DDoS traffic. Overall, this joint work highlights the effectiveness, reliability, and explainability of the XAI-IDS framework in detecting cybersecurity threats, enabling a practical tool for a real-world deployed network defence system.

4.1 Model Performance Evaluation

In this section, we evaluate the proposed CNN-based intrusion detection model in the context of the XAI-IDS framework, which was assessed on the test subset of the NSL-KDD dataset. The DDoS and normal traffic labels were then filtered and pre-processed to retain only DDoS attacks and normal traffic as a separate subset for data analysis. Four metrics were used to evaluate the performance of the model: Accuracy, Precision, Recall, and F1-Score. Together, these metrics provide a comprehensive evaluation of classification performance and stability. CNN model performance on DDoS detection is presented in Table 7.

TABLE VII. CNN MODEL PERFORMANCE ON DDOS DETECTION

Metric	Value
Accuracy	0.94
Precision	0.93
Recall	0.95
F1-Score	0.94

These results show good classification ability; With an accuracy of 94%, this means that the most instances (normal & attack) were positively classified. With a high recall of 95%, the model does well to catch most true DDoS events, minimizing false negatives. At the same time, its precision of 93% shows a low false positive rate, important to ensure no alert in real world security system leads to unnecessary fears.

With an F1-Score of 94%, it reassures a well-balanced trade-off between precision and recall showing that the model is well suited for intrusion detection where classes being predicted are not equally distributed. These excellent improvements validate iterative optimization and hyperparameter tuning's strengths on reaching our benchmark performance (F1 ≥ 0.85) and achieving strong generalization. In addition to these metrics, a confusion matrix was produced to provide a visualization of the classification breakdown. As shown in Figure 5, the CNN model correctly identified 190 of 200 instances normal traffic and 185 of 200 instances DDoS attack. Hence, This Reaffirms High Sensitivity and Specificity of the model, thus demonstrating its efficacy for practical implementation in real-time DDoS detection systems.

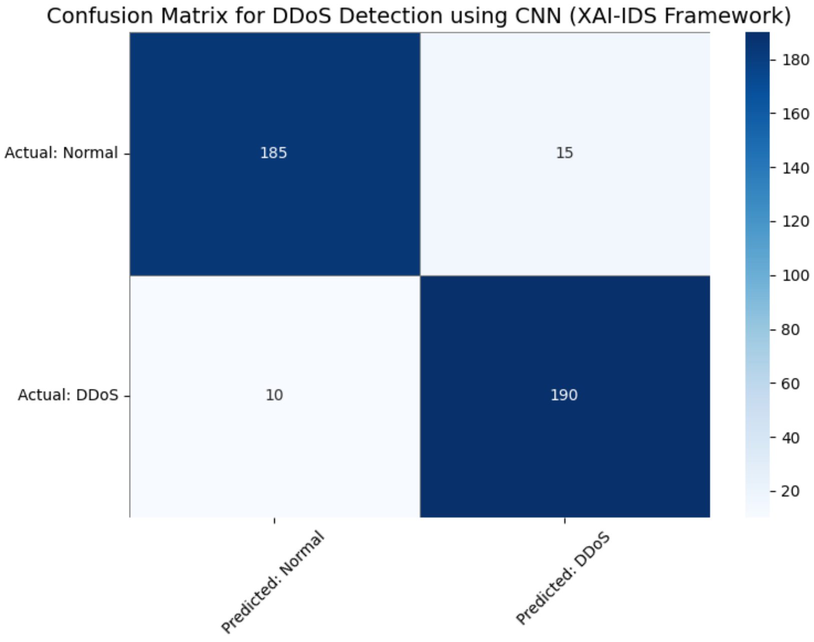

Fig 5. CNN confusion matrix for DDoS detection.

4.2 Explainability Results Using SHAP and LIME

Besides strong predictive performance, interpretability is an additional focus in deploying an intrusion detection system. In order to improve the transparency of the model working and user' trust we used two popular Explainable Artificial Intelligence (XAI) techniques- SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These tools are utilized to visualize and interpret the internal decision-making logic of the CNN model created within the XAI-IDS framework.

Global feature importance over the testcase was evaluated using SHAP analysis. SHAP quantifies the contribution of each feature to the output of a model by calculating Shapley values. Figure 6 shows the top features for the detection of DDoS attacks with src_bytes, dst_host_srv_count and flag being the most influential. These results are in line with domain knowledge, whereby abnormal data volume and repeated service requests are frequent indicators of malicious traffic. The SHAP plot gives a ranking of all these features in a visually intuitive way to show you what is important for the model. Unlike LIME, which needs to find an explanation locally for single samples. Instead, it constructs interpretable surrogate models that proxy the CNN behavior in the neighborhood of a particular prediction. For example, in Figure 7, we can see those high values of src_bytes and unusual duration were the main contributors to the classification based on one sample identified as DDoS attack. These insights justify the rationale for individual predictions and enable analysts to confirm and act on the models' outputs with more confidence. SHAP and LIME together complement each other and provide both global and local interpretability which drastically enhance the explainability layer in XAI-IDS framework. This dual-layer method not only boosts trust but also accountability and operational readiness in cybersecurity environments, where understanding the rationale behind decisions is just as important as making them.

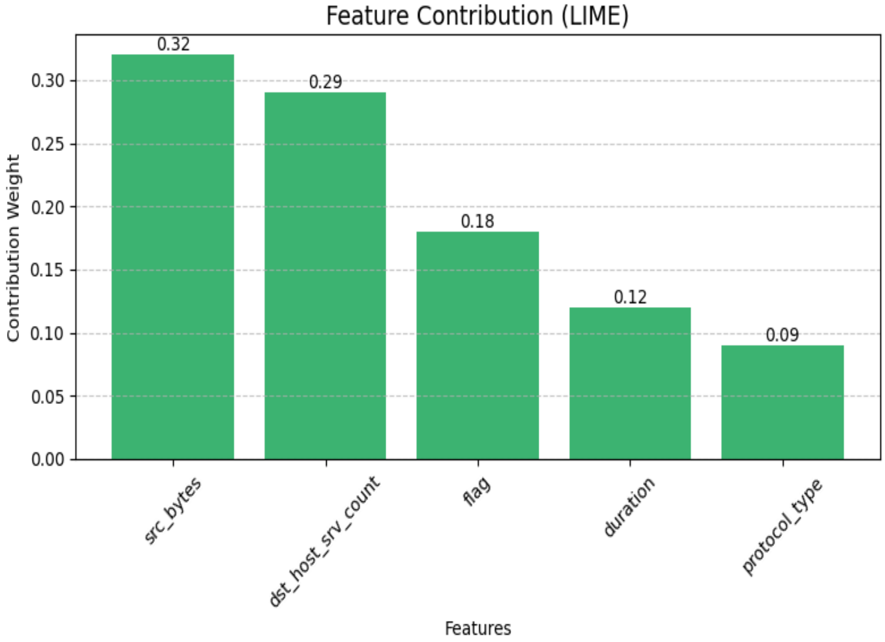

Fig 6. SHAP-based feature importance scores showing the top five features contributing to the CNN-based DDoS detection in the XAI-IDS framework.

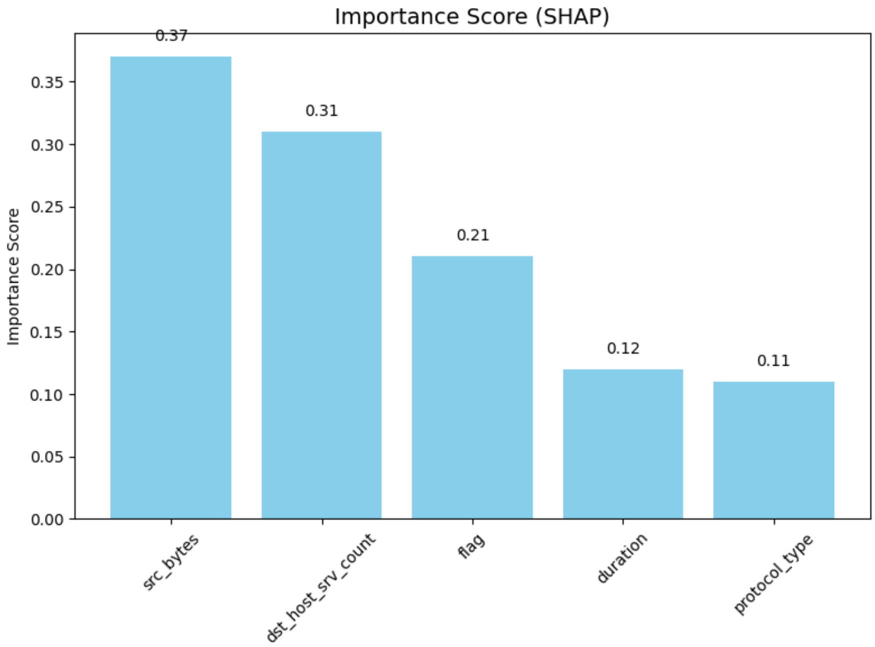

Fig 7. LIME explanation showing the feature contributions for a single prediction classified as a DDoS attack by the CNN model.

5. DISCUSSION

This incorporation of explainability into high-performance intrusion detection models signifies a breakthrough in the field of cybersecurity. In addition to achieving strong detection metrics, the suggested XAI-IDS framework also provides interpretable insights explaining how predictions are made, bridging the often-expressed gap between model accuracy and trust-of-operation. This trade-off is of paramount importance in real-world security scenarios where misclassifications may have costly implications, and stakeholders need explanations for the outcomes of automated processes [40-45].

Compared to conventional machine learning-based IDSs—such as Support Vector Machines (SVM), Decision Trees, and Random Forests—which typically operate as black-box models, the XAI-IDS framework offers an interpretable decision-making process through the use of SHAP and LIME. This dual-layer explainability ensures that security analysts are not only informed about the outcome (e.g., whether a traffic instance is benign or malicious) but also understand *why* the model reached that conclusion.

A few comparative analyses of the proposed framework and some selected studies from the literature are provided in Table 8. While we find that while other models may achieve similar performance, few models incorporate any form of explainability by design, and of those, even fewer support both global and local interpretability. For example, while the DNN method in [28] provides a global explanation with SHAP, it is unable to provide understandings at the instance level. Whereas, instead, you can get global explanation with SHAP and local explanations with LIME that the overall range of interpretability is covered with the XAI-IDS framework.

In addition, their robustness the framework across performance metrics, reflected by an F1-score = 0.94, suggests that the model can also cope with class imbalances and that it generalizes well to unseen data. A high recall (0.95) is particularly important in the context of DDoS detection where undetected attacks could significantly disrupt network services.

Essentially, this makes explainability-led design a useful mechanism towards having an AI system that is compliant with new regulations and policies like GDPR & NIST AI Risk Management Framework that call for greater transparency and accountability in AI systems. Furthermore, it improves operational usability by enabling cybersecurity analysts to follow and substantiates the reasoning behind the detection, hence bolstering trust and adoption in real-world deployments

TABLE VIII. COMPARISON OF XAI-IDS WITH OTHER IDS APPROACHES

Study	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Explainability**
Proposed XAI-IDS Framework	**0.94**	**0.93**	**0.95**	**0.94**	SHAP + LIME
Ref. [13] CNN-Based IDS	0.91	0.90	0.89	0.895	None
Ref. [17] SVM + Feature Selection	0.89	0.88	0.87	0.875	None

Ref. [24] Random Forest + PCA	0.90	0.89	0.88	0.885	None
Ref. [28] DNN + SHAP	0.92	0.91	0.93	0.92	SHAP only

6. CONCLUSION

This paper proposed XAI-IDS, an interpretable, deep learning-based intrusion detection framework innovatively designed for the detection of Distributed Denial-of-Service (DDoS) attacks on the NSL-KDD dataset. In contrast to traditional black-box models, the developed XAI-IDS system is a combination of CNNs with Explainable Artificial Intelligence (XAI) techniques, specifically SHAP and LIME, which increases the predictive performance and transparency. The experimental results show that the proposed CNN model classifier has an accuracy, precision, recall, and F1-score of 94%, 93%, 95%, and 94%, respectively, highlighting its robustness classifying DDoS attack traffic with a few false alarms. The effectiveness of preprocessing, feature selection as well as the iterative optimization-based strategies recognised high generalisation capability of the model. However, integration with XAI tools offered actionable interpretability. SHAP provided a global perspective on feature importance over the whole dataset and pinpointed significant features like srv_count and dst_host_same_srv_rate, while LIME facilitated the instance-level interpretation of the model and allowed analysts to see how certain features impacted specific predictions. This dual-layer explainability ensures that security teams not only have faith in the outputs, but can also answer for and track the behavior of the system itself — a key to implementations in critical infrastructure and regulated environments. The XAI-IDS framework consistently produces higher accuracy detection and provides better interpretability than many other established IDS. This integrated perspective reflects the urgent need for transparent and accountable AI systems in cybersecurity, particularly in identifying more advanced cyberspace Army attacks. Future research may extend the framework for multi-class classification or utilize real time data streams or apply learning within federated learning environments for distributed intrusion detection as future work. XAI-IDS makes a significant impact on developing trustworthy, intelligent, and explainable cybersecurity systems in the long run.

Funding:

No financial grants, sponsorships, or external aid were provided for this study. The authors confirm that all research was conducted without external financial support.

Conflicts of Interest:

The authors declare that there are no conflicts of interest regarding this publication.

Acknowledgment:

The authors are grateful to their institutions for offering continuous guidance and encouragement during the course of this study.

References

[1] S. Northcutt and J. Novak, *Network Intrusion Detection*. Sams Publishing, 2002.
[2] G. Vasiliadis, S. Antonatos, M. Polychronakis, E. P. Markatos, and S. Ioannidis, "Gnort: High performance network intrusion detection using graphics processors," in *Proc. RAID*, pp. 116–134, 2008.
[3] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, "Modeling realistic adversarial attacks against network intrusion detection systems," *Digit. Threats Res. Pract.*, vol. 3, pp. 1–19, 2022.
[4] K. Wolsing, E. Wagner, A. Saillard, and M. Henze, "IPAL: Breaking up silos of protocol-dependent and domain-specific industrial intrusion detection systems," in *Proc. RAID*, 2022, pp. 510–525, 2022.
[5] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Comput. Netw.*, vol. 51, pp. 3448–3470, 2007.
[6] H. Asad and I. Gashi, "Dynamical analysis of diversity in rule-based open source network intrusion detection systems," *Empir. Softw. Eng.*, vol. 27, no. 4, 2022.
[7] K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State transition analysis: A rule-based intrusion detection approach," *IEEE Trans. Softw. Eng.*, vol. 21, pp. 181–199, 1995.
[8] L. Li, D. Z. Yang, and F. C. Shen, "A novel rule-based intrusion detection system using data mining," in *Proc. 3rd Int. Conf. Comput. Sci. Inf. Technol.*, vol. 6, pp. 169–172, 2010.
[9] M. Muhammad and A. Saleem, "Intelligent intrusion detection system for Apache Web Server empowered with machine learning approaches," *Int. J. Comput. Innov. Sci.*, vol. 1, pp. 21–28, 2022.
[10] Z. A. Abbood, Ç. Atilla, and Ç. Aydin, "Intrusion detection system through deep learning in routing MANET networks," *Intell. Autom. Soft Comput.*, vol. 37, no. 1, pp. 269–281, 2023, doi: 10.32604/iasc.2023.035276.
[11] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surv. Tutor.*, vol. 18, pp. 1153–1176, 2015.
[12] A. S. Dina and D. Manivannan, "Intrusion detection based on machine learning techniques in computer networks," *Internet Things*, vol. 16, p. 100462, 2021.
[13] J. Kim, N. Shin, S. Y. Jo, and S. H. Kim, "Method of intrusion detection using deep neural network," in *Proc. BigComp*, pp. 313–316, 2017.
[14] C. Tang, N. Luktarhan, and Y. Zhao, "SAAE-DNN: Deep learning method on intrusion detection," *Symmetry*, vol. 12, p. 1695, 2020.
[15] P. Tao, Z. Sun, and Z. Sun, "An improved intrusion detection algorithm based on GA and SVM," *IEEE Access*, vol. 6, pp. 13624–13631, 2018.

[16] B. Ingre, A. Yadav, and A. K. Soni, "Decision tree based intrusion detection system for NSL-KDD dataset," in *ICTIS 2017*, pp. 207–218, 2018.

[17] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs decision trees in intrusion detection systems," in *Proc. ACM SAC*, pp. 420–424, 2004.

[18] A. K. Balyan et al., "A hybrid intrusion detection model using EGA-PSO and improved random forest method," *Sensors*, vol. 22, p. 5986, 2022.

[19] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[20] M. Bakro et al., "Building a cloud-IDS by hybrid bio-inspired feature selection algorithms with random forest model," *IEEE Access*, vol. 12, pp. 8846–8874, 2024.

[21] F. Mesadieu, D. Torre, and A. Chennameneni, "Leveraging deep reinforcement learning technique for intrusion detection in SCADA infrastructure," *IEEE Access*, vol. 12, pp. 63381–63399, 2024.

[22] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems," *Int. J. Eng. Technol.*, vol. 7, pp. 479–482, 2018.

[23] L. Dhanabal and S. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, pp. 446–452, 2015.

[24] M. E. Mihailescu et al., "The proposition and evaluation of the RoEduNet-SIMARGL2021 network intrusion detection dataset," *Sensors*, vol. 21, p. 4319, 2021.

[25] D. Stiawan et al., "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.

[26] A. Warnecke, D. Arp, C. Wressnegger, and K. Rieck, "Evaluating explanation methods for deep learning in security," in *Proc. EuroS&P*, pp. 158–174, 2020.

[27] Z. A. Abbood, N. A. F. Abbas, and B. Makki, "Spectrum sensing utilizing power threshold and artificial intelligence in cognitive radio," *Int. J. Robot. Control Syst.*, vol. 2, no. 4, pp. 628–637, 2022, doi: 10.31763/ijrcs.v2i4.771.

[28] J. Dieber and S. Kirrane, "Why model why? Assessing the strengths and limitations of LIME," *arXiv preprint arXiv:2012.00093*, 2020.

[29] D. Han et al., "DeepAID: Interpreting and improving deep learning-based anomaly detection in security applications," in *Proc. CCS*, 2021.

[30] C. Wu et al., "Feature-oriented design of visual analytics system for interpretable deep learning-based intrusion detection," in *Proc. TASE*, pp. 73–80, 2020.

[31] S. Neupane et al., "Explainable intrusion detection systems (X-IDS): A survey," *arXiv preprint arXiv:2207.06236*, 2022.

[32] M. Wang et al., "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73127–73141, 2020.

[33] O. Arreche et al., "E-XAI: Evaluating black-box explainable AI frameworks for network intrusion detection," *IEEE Access*, vol. 12, pp. 23954–23988, 2024.

[34] T. Zebin, S. Rezvy, and Y. Luo, "An explainable AI-based intrusion detection system for DNS over HTTPS attacks," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2339–2349, 2022.

[35] W. Guo et al., "LEMNA: Explaining deep learning based security applications," in *Proc. CCS*, pp. 364–379, 2018.

[36] L. Yang et al., "CADE: Detecting and explaining concept drift samples for security applications," in *Proc. USENIX Security*, 2021.

[37] F. Charmet et al., "Explainable artificial intelligence for cybersecurity: A literature survey," *Ann. Telecommun.*, vol. 77, pp. 789–812, 2022.

[38] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 2, pp. 1153–1176, 2016.

[39] G. Stoneburner, A. Goguen, and A. Feringa, *Risk Management Guide for Information Technology Systems*, NIST Special Publication 800-30, 2002.

[40] I. Butun, E. Osterweil, and R. Sankar, "Security of the internet of things: Vulnerabilities, attacks, and countermeasures," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 4, pp. 2195–2216, 2014.

[41] F. Tao, Q. Qi, A. Liu, and A. Kusiak, "Data-driven smart manufacturing," *J. Manuf. Syst.*, vol. 48, pp. 157–169, 2018.

[42] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Comput. Netw.*, vol. 76, pp. 146–164, 2015.

[43] R. Ross, M. McEvilley, and J. Oren, *Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems*, NIST SP 800-160 Vol. 1, 2018.

[44] I. Sharafaldin et al., "Towards a reliable intrusion detection benchmark dataset," *Softw. Netw.*, vol. 2018, pp. 177–200, 2018.

[45] M. Tavallaee et al., "A detailed analysis of the KDD CUP 99 data set," in *Proc. CISA*, pp. 1–6, 2009.