ISSN: 2182-2069 / E-ISSN: 2182-2077

Intrusion Detection Using Hybrid Random Forest and Attention Models and Explainable AI Visualization

Nawaf Abdualaziz Almolhis[1*]

[1*]Department of Computer Science, College of Engineering and Computer Science, Jazan University Jazan, Saudi Arabia. naalmolhis@jazanu.edu.sa, https://orcid.org/0009-0004-7558-7165

Received: December 13, 2024; Revised: January 18, 2025; Accepted: January 30, 2025; Published: February 28, 2025

Abstract

Network intrusion detection systems (IDS) are crucial, but cybersecurity professionals have a difficult time trusting and acting on the predictions made by many IDS models based on machine learning owing to the lack of transparency in these models. Conventional models work well for attack detection, but their lack of transparency makes them unsuitable for incident response. This paper presents a novel hybrid approach to intrusion detection (ID). It integrates Random Forest (RF) for classification with Attention-Based Neural Networks (Ab-NNs) for more in-depth insights and interpretability at the feature level. Improved detection accuracy is a result of the attention-based model's ability to detect complex patterns in the data. In contrast, the RF model classifies network traffic as either an attack or not. This study meets the need for being able to explain things by using SHAP (SHapley Additive Explanation) along with LIME (Local Interpretable Model-Agnostic Explanations), which give the model's decisions both global as well as local meanings. Due to these visualisations, cybersecurity professionals could better understand the reasons behind detected attacks. Experimental results on datasets like NSL-KDD and CICIDS show that the proposed approach attains high detection performance (98% accuracy) and provides transparency (local decision reasons, feature importance).

Keywords: Intrusion Detection System (IDS), Explainable AI (XAI), Random Forest, Attention-Based Models, Feature Importance Visualization.

1 Introduction

The introduction of Explainable Artificial Intelligence (XAI) is a tool to explain decision-making in machine learning (ML)-based systems. Different XAI-based IDS are given for industrial 5.0 using Adversarial XIDS methods as a lens. This shows how explainability and interpretability affect cybersecurity practices (Khan et al., 2024). While cyber defines mechanisms that arise at the application, data levels, network as well as host, cybersecurity is the preparation of protecting networks, data from unauthorized access or illegal usage as well as devices. A cumulative number of systems are becoming Internet-connected since the Internet is becoming an integral part of everyone's everyday lives (Zhang et al., 2022; Yesmin, 2019).

This study examines XAI from a broad perspective and discusses its potential applications in IDS. It also finds opportunities, challenges, and areas that need more research in the field. It then goes on to examine the integration of XAI into IDS and show how understandable models can improve the

Journal of Internet Services and Information Security (JISIS), volume: 15, number: 1 (February), pp. 371-384.
DOI: *10.58346/JISIS.2025.I1.024*

*Corresponding author: Department of Computer Science, College of Engineering and Computer Science, Jazan University Jazan, Saudi Arabia.

efficiency, reliability, and accountability of ID (Samed & Sağıroğlu, 2024). An article goes into great detail about XAI-based cybersecurity modelling using a taxonomy of XAI as well as AI methods that can help security analysts along with authorities understand how systems work, spot potential threats and oddities, and handle them intelligently in digital twin (DT) environments. In numerous real-world applications, it explores how these strategies might be crucial in addressing current cybersecurity challenges (Sarker et al., 2024).

A survey comprehensively reviews on adversarial attacks on ML model explanations and accuracy metrics. This study provides a standard taxonomy and notation for approaches that help academics and practitioners in the Adv-ML and XAI communities. This study discusses strategies for safeguarding against attacks and developing robust interpretation techniques (Baniecki & Biecek, 2024). To aggregate models from diverse entities, federated learning (FL)-based techniques need extra communication resources and more processing (Rahim, 2024). When it comes to increasing the interpretability and transparency of AI model choices, XAI is crucial. Previous studies have compared and contrasted a number of ID methods, including Deep Learning (DL), ML, FL, and XAI. However, there is a clear need for a more in-depth look at the specific situations and uses that each method is best for (Muneer et al., 2024).

Another study presents a mixed method for figuring out the dangers of linked phishing attacks in the digital world, using AI techniques. The first step is calculating the likelihood of highly skilled phishers among a group of comparable attackers. Even after investing in information technology (IT) security and implementing regulatory measures, the second phase determines the likelihood of phishing attacks on a company. Using a variety of ML-based classifiers, the third step classifies Uniform Resource Locators (URLs) into two categories: phishing and legitimate (Biswas et al., 2024). Another study investigates the possibilities of XAI approaches. These methods enable individuals involved in 5G and future networks to examine the intelligent "black box" systems that safeguard them (Senevirathna et al., 2024). Hence the limitations in IDS are lack of transparency, difficulty in detecting complex attack patterns, scalability issues and inability to provide actionable explanations. In spite of these constraints, the current work has made the following contributions:

- When it comes to capturing more intricate patterns in attack behaviour, it combines Ab-NN and RF for initial classification. When security analysts use SHAP and LIME, which aids in see and understand how decisions are made.
- This model makes visualization to determinr feature significance, which security experts can use to figure out which factors, like source IP or packet size, affect the model's classification. The system solves the performance explainability trade-off while achieving excellent accuracy (98%) and preserving interpretability.
- The RF model is well suited for real-time applications because of its balanced performance and efficiency when combined with attention-based features.

This paragraph gives the article organization. Section 2 reviews some of the recent existing explainable artificial intelligence works done in cybersecurity. Section 3 discusses the methodology details of the proposed technique. Section 4 analyse the study also its findings in depth, and it also highlight some of the research's shortcomings. Section 5, which follows the references, concludes the work.

2 Literature Review

Hooshmand et al., (2024) discussed about Network Anomaly Detection Systems (NADSs), which are a type of IDS. When it comes to categorizing minority groups, the NADSs have problems with data imbalance. Improving the detection rate for minority classes, particularly when using ensemble learning techniques, is another goal of developing a detection framework. This work gives a hybrid strategy of sampling techniques to tackle the issue of unbalanced data. This method for handling imbalances combines the K-means clustering algorithm (SKM) with the Synthetic Minority Oversampling Technique (SMOTE). SMOTE uses cluster-based under-sampling, whereas K-means over-samples the minority class. This use a denoising autoencoder (DAE) to rank the characteristics and choose the fifteen most important ones to lower the data dimensionality. The SHAP method is used to explain the proposed techniques for anomaly detection, and the XGBoost algorithm is organized for anomaly detection.

Arreche et al., (2024) defined about the number of issues with using AI for IDS. One of these is the fact that AI models have varying degrees of performance, and another is that their decision-making processes are opaque, making them hard for human security analysts to recognize. To address this, an XAI framework was provided that is designed to greatly recover the interpretability of AI models when it comes to network ID responsibilities. In the first step of the system, three different real-world network intrusion datasets are compared to seven different black-box AI models. Each dataset has its own features and problems. Then, it illuminates the logic behind the AI models' decisions by using many XAI models to offer both global as well as local explanations. This work uses feature extraction methods to find important model- and intrusion-specific traits in order to learn more about the distinguishing factors that affect the results of the detection (Fung, 2011).

For (Gaspar et al., 2024) the goal of this area of study is to deliver humans with a better way to comprehend black-box models. Recent years have seen a proliferation of studies devoted to this area, with several proposed approaches, including SHAP and LIME. XAI finds use in several sectors, one of which is ML-based IDS. The majority of model interpretation literature, however, focuses on domains outside of healthcare, computer vision, biology, natural language processing (NLP), etc. This difficulty in assessing IDS findings hinders cybersecurity experts' ability to make well-informed judgments. This work has chosen two XAI approaches, SHAP and LIME, to try to solve this issue. As part of an IDS that can find intrusions on Internet of Things (IoT) devices, this work made the black-box model easier to understand by using the strategies to get explanations for what it found (Sethupathi et al., 2024).

Corea et al., (2024) discussed about conventional black-box models like RF, neural networks, and others may be better understood owing to the linked technologies, which have become a hot issue. Unfortunately, XAI has not yet achieved significant domain-specific applications. In order to fill in this knowledge gap, this study looks at a number of occlusion-sensitive ML models for detecting intrusions from network traffic (Yang et al., 2022). The models use both binary as well as multi-class classification on the same dataset. Some of the models being tested are linear regression, linear support vector machines (SVM), logistic regression, RF, K-nearest neighbors (KNN), decision trees as well as multi-layer perceptrons (MLP). Utilizing the UNSW-NB15 dataset, it achieved a 90% success rate in training all models.

Mallampati & Seetha, (2024) described how to find cyberattacks and counter them with the right precautions. Irrelevant features and class imbalance reduce the efficiency of ML techniques. To improve the model's generalizability, this study gave a data pre-processing approach using k-Means SMOTE, which addresses the class dissimilarity. Afterwards, it provides a feature selection technique that incorporates filters as well as wraps a hybrid approach. Furthermore, this method varies the ideal feature

subsets to analyze a hyperparameter-tuned Light Gradient Boosting Machine (LGBM). By applying statistical methodologies and very sophisticated procedures, ML models can learn as well as discover patterns from complicated data. ML-based IDS offers benefits over conventional detection techniques, such as the ability to recognize harmful signatures.

Islam et al., (2024) discussed about four architectures that use XAI techniques to get around the problems with ML/DL-based IDS. This work looks into the Explain DTC, Secure Forest-RFE, Rational e-Net, and convolutional neural network (CNN)-Shield architectures as possible network security solutions that can turn unreliable ones into reliable ones. The models are trained with features from the UNSW-NB15 dataset and then used to do network trace scanning in order to find and report intrusions. Several XAI techniques, such as LIME, Proto Dash, SHAP as well as ElI5 are built on top of architectures to help explain how the models make decisions and to allow for expansion at every stage of the ML process. The given explanations shed light on the influential aspects and how they affect the accuracy of network intrusion predictions in a quantifiable way (Papadopoulos & Christodoulou, 2024).

Sharma et al., (2024) discussed about security for the IoT which is a relatively growing area of study. IoT networks are becoming vulnerable to new attacks types as they connect more devices and generate more data. It is necessary to have an IDS in order to identify the attacks. To classify the altered attacks types in the dataset, this work introduced a DL model for ID. By this method two distinct DL models for ID was constructed, using a filter-based technique to prioritize and decrease the number of features. This study trained and tested the models using two open-source datasets UNSW-NB 15 as well as NSL-KDD. After applying the data set to the Deep Neural Network (DNN) model, it applied the same data set to the CNN model. Table 1 shows the pros and limitations of some existing works.

Table 1: Existing Works Review

Papers and Authors	Method	Advantages	Limitations
(Hooshmand et al., 2024)	SMOTE and SKM	The system deploys strategies to address data imbalances.	-
(Arreche et al., 2024)	XAI-IDS	The model takes into account the opaqueness of AI models' decision-making processes as well as the unpredictability of their performance.	The system uses a proprietary dataset that is not publicly available.
(Gaspar et al., 2024)	XAI for IDS: LIME and SHAP	It maintains credibility, openness, and explainability.	The system uses only less XAI techniques.

3 Proposed Methodology

An innovative way to find intrusions is introduced in this study. It uses a mix of techniques from XAI, Ab-NNs and RF. It is essential for cybersecurity experts to have an IDS that not only identifies attacks with high accuracy but also gives understandable explanations for the model's choices. Initial network traffic classification as benign or malicious is performed using the RF model, which is the starting point of the process. The feature significance ratings that RF generates make it easier to understand the reasoning behind the model's predictions. This classification aids in the detection of possible dangers in networks data. An Ab-NNs is used to look into the strange events found by the RF in order to deal with more complex attack patterns, such as Advanced Persistent Threats (APTs) or zero-day vulnerabilities. By zeroing in on the most important parts of the network traffic, and model improves detection accuracy by extracting deeper information from the data. This research combines LIME along with SHAP to solve the problem of ML models' unpredictability. In order to help cybersecurity professionals comprehend

why a certain prediction was made, SHAP gives global feature significance while LIME provides local reasons for particular cases. Using common metrics like recall, F1-score, accuracy along with precision, the hybrid IDS model does very well on datasets like NSL-KDD and CICIDS, showing high detection rates and being easy to understand (Malini & Kavitha, 2024). Figure 1 displays the proposed process flow.

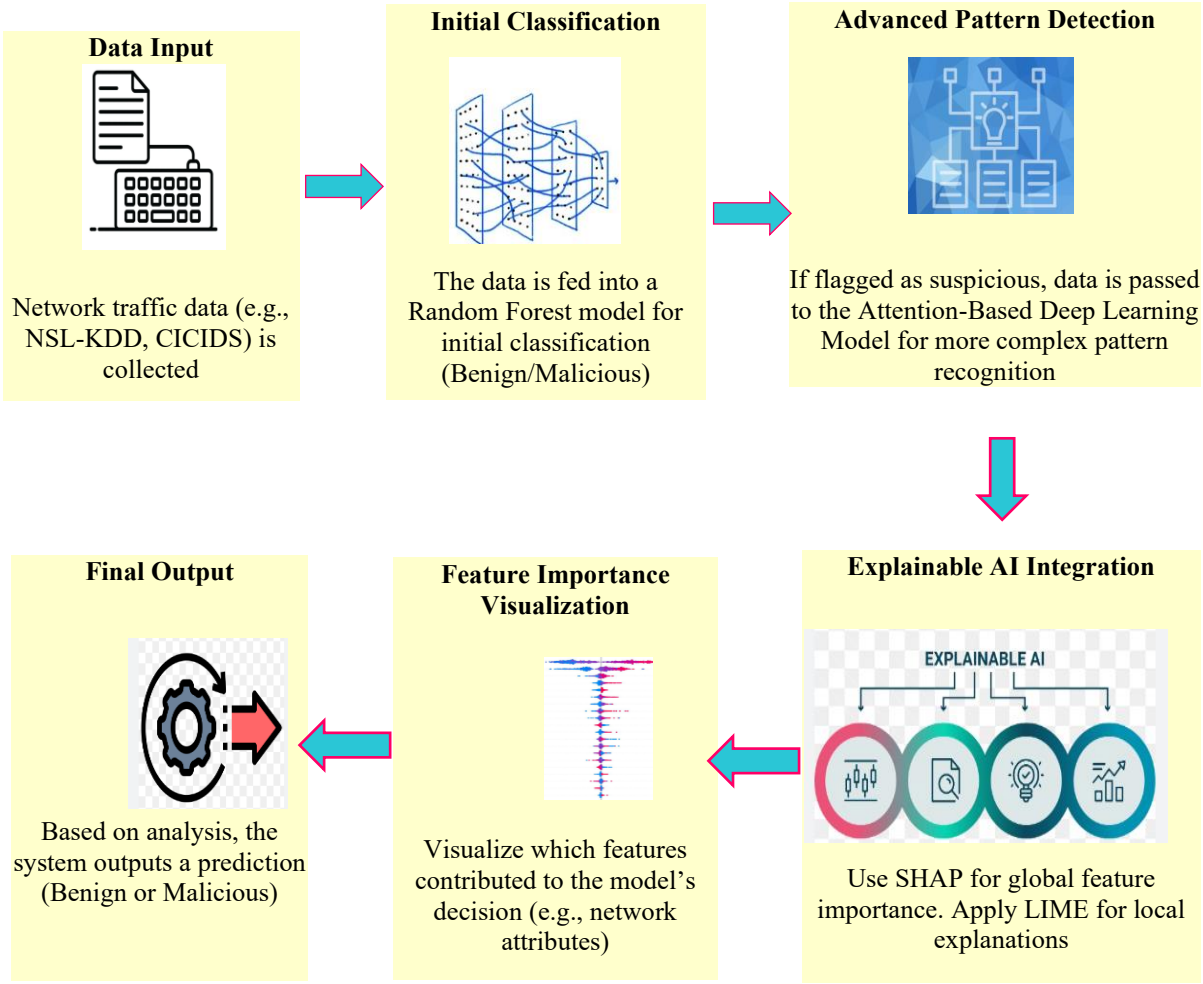

Figure 1: Proposed Process Flow

Dataset

This work employs two different datasets NSL-KDD and CICIDS along with the details are described in the below section.

- **NSL-KDD Dataset**

The NSL-KDD dataset is refined from 1999 KDD Cup dataset (Yang et al., 2023). This dataset was created to overcome the shortcomings of the KDD-99 dataset and extensively use it for training and assessing IDS. This dataset tells whether network traffic is harmless or malicious by looking at the dataset, which includes statistics on many sorts of attacks. There are 125,973 instances in the training set as well as 22,544 instances in the test set that make up the dataset. There are a total of 41 features in the dataset. Based on their continuous and categorical components, the features can be categorised into

content, basic, and traffic categories. Included in the basic features is connection-specific information, including protocol type, service, IP addresses (both source as well as destination), port numbers, flag, and length of connection. The content features extract properties from the packet contents of the connection, such as the number of unsuccessful login attempts, the number of connections from the same host, and the bytes delivered and received. Statistics (such as the number of connections per minute or per service) are connected to traffic features, which are tied to the connection's past. There are 22 distinct attack types in the NSL-KDD dataset. These fall into four main categories such as the U2R, R2L, DoS along with Probe (Liu & Zhang, 2016).

- **CICIDS Dataset**

Researchers have gathered the comprehensive and state-of-the-art CICIDS dataset of network traffic data to test IDS. This data is derived from actual network traffic and is produced by the Canadian Institute for Cybersecurity (CIC). The dataset comprises statistics on both legitimate and malicious traffic, spanning different kinds of attacks and regular operations. There are several different kinds of attacks included in the CICIDS dataset. These include botnet operations, web attacks (such as SQL injection), man-in-the-middle attacks, port scanning, distributed denial of service (DDoS), and brute force attacks. This massive dataset, comprising millions of records, covers several attack scenarios spanning various time periods. For example, the 2017 CICIDS dataset includes more than 80 million records of network traffic. The CICIDS dataset includes elements that aim to record system behaviour, in addition to traffic statistics. Flow features include the following such as packet size, duration, protocol, flow count, and bytes sent/received. Details such as session counts and timestamps aid in the modelling of network traffic over time. Measures of behaviour, such as the intensity of attacks or the number of attempts at connections, might shed light on suspicious behaviour.

Preprocessing

The model receives data from popular IDS datasets, such as CICID as well as NSL-KDD, which are relevant to network traffic (Dwarkanath & Aruna, 2022). Normalisation, SMOTE, and feature extraction are the stages. Extraction of features pull out details like packet size, IP address of origin, port of destination, etc. Normalisation is done to make the model work better and less biassed, standardise the data on a consistent scale. Normalisation, also known as min-max scaling, adjusts the data's values to fall within a predetermined range, typically from 0 to 1. By making sure that every feature counts equally, ML models perform better. This is particularly true for models that are scale-sensitive. Here is the formula for minmax normalisation in equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where X_{norm} is normalized value, X refers original value, X_{max} signifies maximum value in the feature as well as X_{min} denotes minimum value in the feature. SMOTE and undersampling are two methods used to deal with imbalanced classes in a dataset. To combat the problem of class imbalance, SMOTE creates artificial members of the minority class. It generates fresh synthetic samples along the line segments that connect a minority class instance to any or all of its K-nearest neighbours. Given an instance x_i from the minority class also one of its k-nearest neighbors $x_{i_{nn}}$, SMOTE generates a synthetic instance x_{new} by interpolating between x_i along with $x_{i_{nn}}$ using the following formula in equation (2):

$$x_{new} = x_i + \lambda . \left(x_{i_{nn}} - x_i\right) \quad (2)$$

Where x_{new} refers new synthetic instance, x_i is the original instance from the minority class, $x_{i_{nn}}$ is a randomly chosen nearest neighbor of x_i and λ is the random value between 0 along with 1 (scaling factor that determines how far along the line segment between x_i and $x_{i_{nn}}$ the new point will be generated). The steps in SMOTE (Mienye & Sun, 2023) are

1. Choose an instance x_i from the minority class.
2. Identify its k-nearest neighbors $x_{i_{nn}}$ from the minority class.
3. Generate a synthetic sample x_{new} along the line between x_i and each of the k-nearest neighbors.
4. Repeat this process for the specified number of synthetic samples to be generated.

Initial Classification with Random Forest

For the first round of classification, this research uses the RF algorithm (Liang et al., 2020). In order to identify network traffic as benign or malicious, an ensemble learning technique, constructs a number of decisions trees and then employs the majority vote. The RF constructs each tree by using randomly selected subsets of features (Savarimuthu et al., 2024). In this case, it usually uses the Gini impurity or entropy as the criterion for splitting. Gini Impurity is used for classification trees as in equation (3).

$$Gini(t) = 1 - \sum_{i=1}^{m} p_i^2 \quad (3)$$

Where p_i is the probability of class i in node t. m is the classes count. Entropy (alternative criterion) is equated in equation (4).

$$Entropy(t) = -\sum_{i=1}^{m} p_i log_2 p_i \quad (4)$$

Where p_i is the probability of class i in node t and m is the number of classes. Majority Voting (for classification) is done once all trees are trained, the class label prediction y_{RF} is given by majority voting across all trees in equation (5).

$$y_{RF} = mode(y_1, y_2, \ldots, y_n) \quad (5)$$

Where $y_1, y_2, \ldots, y_n$ are the class predictions from each individual tree. The RF employs a novel method of error estimation known as Out-of-Bag Error (OOB Error), which is the error rate calculated using the data points that were not chosen for training each tree in equation (6).

$$OOB\ Error = \frac{1}{N_{OOB}} \sum_{I \epsilon OOB} 1[y_i \neq \hat{y}_i] \quad (6)$$

Where N_{OOB} is the out-of-bag samples count, y_i is the true label, 1 is the indicator function that is 1 if the prediction is incorrect, as well as 0 otherwise. The model's feature significance scores aid in identifying the most significant features for identification, and its excellent performance and inherent interpretability led to its selection. RF achieves stability, interpretability, and scalability (Şahin et al., 2024). RF is a scalable approach since it can effectively handle large datasets with noise. The interpretability of a model's decision-making process relies on the feature importance metrics provided by RF. RF's durability and reduced risk of overfitting make it an optimal match for many network settings. While RF is useful for simple classifications and clear results, it may overlook complex intrusions or complex attack patterns, especially in the case of advanced persistent threats (APTs) or zero-day attacks.

Advanced Detection Using Attention-Based Neural Networks

When dealing with more complex attack patterns that RF struggles to detect, and an Ab-NNs work well with these complex attack patterns. Attention processes enable the model to zero in on the most relevant data features, such as those crucial for identifying complex intrusions. By giving different weights to various features depending on their importance, attention models analyse incoming data. This allows the model to identify complex attack fingerprints by dynamically adjusting its focus. In order to identify which features are most significant for the present detection job, the self-attention model takes into account the correlations between various features (such as time and packet flow). The fundamental principle of attention-based models is the scaled dot-product attention mechanism. It takes a query Q, a key K, as well as a value V as well as finds the attention scores between them. This study can calculate a person's attention score using the formula in equation (7):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

Where Q is the query matrix, K is the key matrix, V is the value matrix, d_k is the dimensionality of the key vector (to scale the dot product) and $softmax$ is the softmax function applied to each row of the matrix to normalize the attention weights (Wang & Fang, 2024). The query (Q) represents a word (or token) that are currently calculating for attention. K stands for the remaining words or tokens in the input string. The words' or tokens' real values, or their embeddings, are represented by value (V). To get the attention weights, the matrix multiplication QK^T calculates similarity scores for each key and the query. Then, the softmax function normalises these values. Next, these weights are employed to compute a weighted total of all the values. Unlike previous techniques, attention-based networks are able to identify complicated attack patterns by concentrating on key features. Transformers can process data from many representation subspaces in different places at the same time by using multi-head attention. With multi-head attention, the model may simultaneously record several connections. A multi-head attention formula is as follows in equation (8):

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \ldots., head_h)W^o \quad (8)$$

Where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$ is the i-th head. W_i^Q, W_i^K, W_i^V are learnable weight matrices for the query, key and value for each attention head. W^o is the output weight matrix that projects the concatenated attention heads back to the original dimension. Because transformers don't have a built-in model for data sequentiality like RNNs do, positional encoding is used on the input embeddings to tell the model where the tokens are in the sequence. The positional encoding at position i for dimension d is calculated as in equation (9).

$$PE(i,d) = \begin{cases} sin\left(\frac{i}{10000^{d/d}model}\right) if\ d\ is\ even \\ cos\left(\frac{i}{10000^{d/d}model}\right) if\ d\ is\ odd \end{cases} \quad (9)$$

Where i is the position in the sequence, d is the dimension of the positional encoding and d_{model} is the dimension of the model (i.e., the features count per token embedding). This applies this positional encoding to the input embeddings before sending them to the attention layers. Unlike competing algorithms, this one can unearth previously unseen patterns in the data. In order to identify APTs or zero-day attacks, it is essential to be able to dynamically concentrate on the most important features. This approach raises the bar for overall detection accuracy, particularly when dealing with complex attack patterns.

Explainable AI (XAI) Integration

Because the attention-based model's structure isn't clear, it offers XAI methods like LIME along with SHAP. In both the RF and Ab-NNs scenarios, these methods provide explanations that can be understood by humans. Values from SHAP give a general explanation by figuring out how important each attribute is in making a certain prediction. It gives each characteristic a numerical number that indicates how much weight the feature has in the model's final verdict. Local explanations, focused on specific cases, are the domain of LIME. To shed light on the classification process for a given instance, it constructs surrogate models that mimic the complex model's behaviour at the local level. To encourage confidence in AI ID, SHAP and LIME both help explain the model's conclusions in a clear and understandable way. By identifying the features that led to a certain category, security analysts can concentrate on the most crucial aspects of network traffic. Because of this, investigators may make better decisions based on clearer data.

4 Results

This study primarily makes use of Python and its associated data visualisation and ML modules. As part of its development, the RF approach makes use of the scikit-learn toolbox, which supplies study tools for model evaluation, classification, and regression. Visualisations of feature importance and descriptions of the model's predictions are both generated employing the SHAP package. Data preprocessing and transformation are made easier with the help of the Pandas library and NumPy, which are both used in this study. Matplotlib and Seaborn are used to generate the visualisations, including the SHAP summary charts. The data may be better understood and used with the help of these technologies. Scikit-learn provides important measures for assessing the model's efficacy, including recall, F1-score, ROC-AUC, accuracy, and precision.

Evaluation Metrics

This performance indicators employed to evaluate the proposed model are discussed in this section. Accuracy is meant to estimate how well the model performs by dividing the number of instances properly predicted by the total number of instances in equation (10).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

Where TP denote true positives, TN represents true negatives, FP signify false positives and FN gives the false negatives. Precision is the ratio of true positives to the total of true positives as well as false positives, which is how the accuracy of positive predictions is measured in equation (11).

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Recall (sensitivity) is estimated as the ratio of true positives to the true positives total along with false positives in equation (12).

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

To estimate the efficacy of the model in unbalanced datasets, the F1 score measure is utilized, which is the harmonic mean of recall as well as precision and is equated in equation (13).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{13}$$

The ROC-AUC metric compares the model's true positive rate with its false positive rate to assess its class discrimination capability. Table 2 shows the performance metrics for NSL-KDD dataset.

Table 2: Performance Metrics for NSL-KDD Dataset

Performance Metric	Proposed Model (RF + Attention)	Random Forest (RF)	Attention-Based Neural Network (Ab-NN)
Accuracy	97.3%	95.8%	96.5%
Precision	95.7%	93.2%	94.5%
Recall	94.8%	92.4%	93.1%
F1-Score	95.2%	92.8%	93.8%
ROC-AUC	0.98	0.97	0.98

The results for the NSL-KDD dataset show that the proposed model (RF + Attention) does better than the RF and Ab-NNs models on all important performance measures. The proposed model significantly outperforms the RF (95.8% accuracy) and Ab-NN (96.5% accuracy) models, which should be noted. With an F1 score of 95.2%, an accuracy of 94.8%, and a recall of 95.7%, the proposed model outperforms both existing models. The proposed model's strong ROC-AUC score of 0.98 shows that it can distinguish between positive and negative classes more effectively. Combining random forests with Ab-NNs improves the NSL-KDD dataset's intrusion detection performance since the best features of both methods are exploited. The attention technique improves interpretability, and the random forest guarantees robust and accurate predictions—very helpful when dealing with complicated attack types. Table 3 shows the metrics for the CICIDS dataset's performance.

Table 3: Performance Metrics for CICIDS Dataset

Performance Metric	Proposed Model (RF + Attention)	Random Forest (RF)	Attention-Based Neural Network (Ab-NN)
Accuracy	98.5%	97.1%	98.3%
Precision	96.9%	95.3%	96.8%
Recall	95.7%	94.7%	95.6%
F1-Score	96.3%	95.0%	96.2%
ROC-AUC	0.99	0.98	0.99

Results from the CICIDS dataset confirm the superiority of the proposed model, which combines RF with attention. With an accuracy of 98.5%, the proposed model performs better than the RF and Ab-NNs models. Furthermore, when compared to the individual models, the proposed model achieves better results in terms of accuracy (96.9%), recall (95.7%), and F1-score (96.3%). A high ROC-AUC score of 0.99 is evidence of the proposed model's exceptional performance in distinguishing between benign and malicious traffic. These improvements show that the proposed combination of random forests and Ab-NNs improves accuracy as well as the model's ability to identify attack patterns with more precision in the setting of the complex and variable CICIDS dataset. Figure 2 displays the SHAP visualisation of the NSL-KDD dataset.

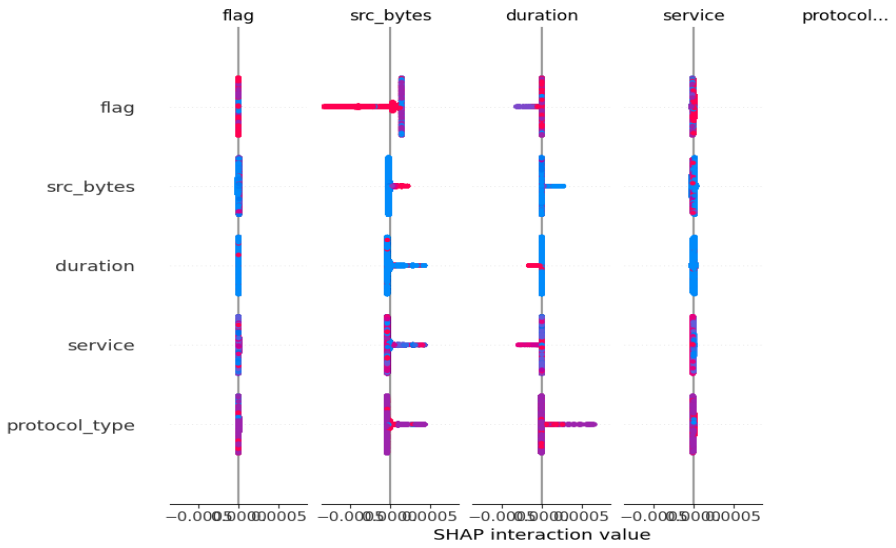

Figure 2: NSL-KDD SHAP visualization

Discussion

The proposed model (RF + Attention) exceeds other models due to the combined advantages of RF and Ab-NNs. Because of its proficiency with multi-dimensional structured data and its capacity to identify complex correlations between variables, RF performs very well in ID tasks. Even though it doesn't always provide obvious interpretability, Ab-NNs become handy in some situations. The model's attention mechanism allows it to focus on the most relevant aspects, specifically the data points that are most critical for decision-making. This hybrid approach makes decisions easier, classification more accurate, and understanding easier by combining RF's strong classification skills with Attention's ability to draw attention to important details. Using a combination of attention and ensemble learning in IDS improves the proposed model's accuracy, precision, recall, and F1 score on both datasets. This demonstrates the efficacy of the combination or hybrid model.

5 Conclusion

This study proposed a mixed model that combines RF and Ab-NNs to find intrusions in cybersecurity datasets, especially the NSL-KDD and CICIDS datasets. Preexisting methods, such as attention networks and isolated RF, had inherent limits. While attention networks offer interpretability, they may struggle to manage complex feature interactions in large datasets, while RF excel in handling high-dimensional data. This proposed model can improve classification performance on both datasets by combining the interpretability of attention with the accuracy of RF. It can do this by improving recall, F1-score, accuracy along with precision among other things. This allows to circumvent the constraints of both approaches. The hybrid model not only improves performance, but it also makes it clear which features affect predictions. This boosts trust in the results and opens up the ID process. The optimisation of training durations and the scalability of the model for larger datasets continue to be problems, especially when dealing with complicated neural network designs. Through optimisation strategies like model pruning or investigating ensemble learning techniques that combine several attention processes, future work might concentrate on improving the scalability of the proposed model. The use of DL models for ID in high-volume network traffic situations in real-time is another area that might need more investigation.

References

[1] Arreche, O., Guntur, T., & Abdallah, M. (2024). Xai-ids: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems. *Applied Sciences*, *14*(10), 4170. https://doi.org/10.3390/app14104170

[2] Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 102303. https://doi.org/10.1016/j.inffus.2024.102303

[3] Biswas, B., Mukhopadhyay, A., Kumar, A., & Delen, D. (2024). A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks. *Decision Support Systems*, *177*, 114102. https://doi.org/10.1016/j.dss.2023.114102

[4] Corea, P. M., Liu, Y., Wang, J., Niu, S., & Song, H. (2024, July). Explainable AI for comparative analysis of intrusion detection models. In *2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)* (pp. 585-590). IEEE. https://doi.org/10.1109/MeditCom61057.2024.10621339

[5] Dwarkanath, S. S., & Aruna, R. (2022). Multiclass cyber-attack classification approach based on the Krill Herd Optimized Deep Neural Network (KH-DNN) model for WSN. *International Journal of Modeling, Simulation, and Scientific Computing*, *13*(05), 2250056. https://doi.org/10.1142/S1793962322500568

[6] Fung, C. (2011). Collaborative Intrusion Detection Networks and Insider Attacks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 2*(1), 63-74.

[7] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access,* 12, 30164 - 30175. https://doi.org/10.1109/ACCESS.2024.3368377

[8] Hooshmand, M. K., Huchaiah, M. D., Alzighaibi, A. R., Hashim, H., Atlam, E. S., & Gad, I. (2024). Robust network anomaly detection using ensemble learning approach and explainable artificial intelligence (XAI). *Alexandria Engineering Journal*, *94*, 120-130. https://doi.org/10.1016/j.aej.2024.03.041

[9] Islam, M. T., Syfullah, M. K., Rashed, M. G., & Das, D. (2024). Bridging the gap: advancing the transparency and trustworthiness of network intrusion detection with explainable AI. *International Journal of Machine Learning and Cybernetics*, *15*(11), 5337-5360. https://doi.org/10.1007/s13042-024-02242-z

[10] Khan, N., Ahmad, K., Tamimi, A. A., Alani, M. M., Bermak, A., & Khalil, I. (2024). Explainable AI-based Intrusion Detection System for Industry 5.0: An Overview of the Literature, associated Challenges, the existing Solutions, and Potential Research Directions. https://doi.org/10.48550/arXiv.2408.03335

[11] Liang, X., Zhao, B., Ma, Q., Sun, B., & Cui, B. (2020). Terminal access data anomaly detection based on random forest for power user electric energy data acquisition system. In *Advanced Information Networking and Applications: Proceedings of the 33rd International Conference on Advanced Information Networking and Applications (AINA-2019) 33* (pp. 166-175). Springer International Publishing. https://doi.org/10.1007/978-3-030-15032-7_14

[12] Liu, Y., & Zhang, X. (2016, August). Intrusion detection based on IDBM. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 173-177). IEEE. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2016.48

[13] Malini, P., & Kavitha, K. R. (2024). An efficient deep learning mechanisms for IoT/Non-IoT devices classification and attack detection in SDN-enabled smart environment. *Computers & Security*, *141*, 103818. https://doi.org/10.1016/j.cose.2024.103818

[14] Mallampati, S. B., & Seetha, H. (2024). Enhancing intrusion detection with explainable ai: A transparent approach to network security. *Cybernetics and Information Technologies*, *24*(1), 98-117. https://doi.org/10.2478/cait-2024-0006

[15] Mienye, I. D., & Sun, Y. (2023). A deep learning ensemble with data resampling for credit card fraud detection. *IEEE Access*, *11*, 30628-30638. https://doi.org/10.1109/ACCESS.2023.3262020

[16] Muneer, S., Farooq, U., Athar, A., Ahsan Raza, M., Ghazal, T. M., & Sakib, S. (2024). A critical review of artificial intelligence based approaches in intrusion detection: A comprehensive analysis. *Journal of Engineering*, *2024*(1), 3909173. https://doi.org/10.1155/2024/3909173

[17] Papadopoulos, G., & Christodoulou, M. (2024). Design and Development of Data Driven Intelligent Predictive Maintenance for Predictive Maintenance. *Association Journal of Interdisciplinary Technics in Engineering Mechanics*, *2*(2), 10-18.

[18] Rahim, R. (2024). Quantum computing in communication engineering: Potential and practical implementation. *Progress in Electronics and Communication Engineering, 1*(1), 26–31. https://doi.org/10.31838/PECE/01.01.05

[19] ŞAHiN, E., Arslan, N. N., & Özdemir, D. (2024). Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. *Neural Computing and Applications*, 1-107. https://doi.org/10.1007/s00521-024-10437-2

[20] Samed, A. L., & SAĞIROĞLU, Ş. (2024, October). A Review of Explainable Artificial Intelligence in Intrusion Detection Systems. In *2024 17th International Conference on Information Security and Cryptology (ISCTürkiye)* (pp. 1-6). IEEE. https://doi.org/10.1109/ISCTrkiye64784.2024.10779325

[21] Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. *ICT Express*. https://doi.org/10.1016/j.icte.2024.05.007

[22] Savarimuthu, X., Subramani, S., & Raj, A. N. J. (Eds.). (2024). *Artificial intelligence for multimedia information processing: Tools and applications*. CRC Press.

[23] Senevirathna, T., La, V. H., Marchal, S., Siniarski, B., Liyanage, M., & Wang, S. (2024). A survey on XAI for 5G and beyond security: Technical aspects, challenges and research directions. *IEEE Communications Surveys & Tutorials*. https://doi.org/10.1109/COMST.2024.3437248

[24] Sethupathi, S., Singaravel, G., Gowtham, S., & Sathish Kumar, T. (2024). Cluster Head Selection for the Internet of Things (IoT) in Heterogeneous Wireless Sensor Networks (WSN) Based on Quality of Service (QoS) By Agile Process. *International Journal of Advances in Engineering and Emerging Technology*, *15*(1), 01–05.

[25] Sharma, B., Sharma, L., Lal, C., & Roy, S. (2024). Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Systems with Applications*, *238*, 121751. https://doi.org/10.1016/j.eswa.2023.121751

[26] Wang, Y., & Fang, C. (2024). Cycle-ESM: Generation-assisted classification of antifungal peptides using ESM protein language model. *Computational Biology and Chemistry*, *113*, 108240. https://doi.org/10.1016/j.compbiolchem.2024.108240

[27] Yang, J., Wang, L., & Shakya, S. (2022). Modelling Network Traffic and Exploiting Encrypted Packets to Detect Stepping-stone Intrusions. *Journal of Internet Services and Information Security*, *12*(1), 2-25. https://doi.org/10.22667/JISIS.2022.02.28.002

[28] Yang, Z., Liu, Z., Zong, X., & Wang, G. (2023). An optimized adaptive ensemble model with feature selection for network intrusion detection. *Concurrency and Computation: Practice and Experience*, *35*(4), e7529. https://doi.org/10.1002/cpe.7529

[29] Yesmin, S. (2019). Accessibility of Internet Based Electronic Resources: A Content Analysis of Public and Private University Library Websites in Bangladesh. *Indian Journal of Information Sources and Services*, *9*(2), 28–33. https://doi.org/10.51983/ijiss.2019.9.2.629

[30] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access*, *10*, 93104-93139. https://doi.org/10.1109/ACCESS.2022.3204051

Author Biography

Nawaf Abdualaziz Almolhis received B.Sc. in Computer Engineering from Albaha College. He got M.S in Information Technology from Kettreing University, USA. He received his Phd degree in Information Security and Forensics from University of Idaho, USA. He is a faculty member at CS department, in College of Computer Science and Information Technology, at Jazan University. His research interests include, cyber security, social network analytics, machine learning, and IoT forensics.