Explainable Artificial Intelligence for Network Intrusion Detection Using Machine Learning Models

Km. Anam[1] and Dr. Gopindra Kumar[2]
[1]Research Scholar, ABSS Institute of Technology, Meerut, India
[2]Dean (R&D), ABSS Institute of Technology, Meerut, India

Abstract

Network intrusion detection systems (NIDS) represent a key component of the modern cybersecurity architecture as they can provide 24/7 monitoring of network traffic in order to determine malicious intent. Over the past few years, machine learning (ML) models have significantly improved the performance of intrusion detection; nevertheless, the majority of them represent opaque models that do not reveal much information about the way the decisions are made. This lack of interpretability undermines trust, as it hinders a forensic analysis, and limitation to its adoption in the context of the actual security operations. In order to address this difficult issue, the current project suggests an elaborated explainable artificial intelligence (XAI)-based framework of intrusion detection, which combines the traditional machine-learning classifiers with the use of the post-hoc explanatory methods. The presented framework uses random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost) as intrusion detection models and makes use of SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to provide a global and a local level of interpretability. The publicly accessible NSL -KDD benchmark dataset is used to evaluate the system. Experimental findings prove the ability of the proposed method to yield higher detection rates besides providing significant explanations of forecasts. Explainability is added to help increase transparency, assist security analysts in their decision-making process, and enhance trust in intelligent intrusion detection systems.

Keywords: Explainable Artificial Intelligence, Network Intrusion Detection, the NSL-KDD data set, Machine Learning, SHAP, LIME, and Cybersecurity.

1. Introduction

The fast development of the internet, cloud computing, mobile technologies as well as the Internet of Things (IoT) have significantly enhanced complexity of modern network environment. At the same time, cyber threats are increasingly common, more advanced, and detrimental. The recent cybersecurity reports suggest that millions of cyber-attacks are faced by organizations all around the world each day; these can be malware infections, phishing, denial-of-service (DoS) attacks, ransomware, and information breaches. Such threats cause major threats to the authorization of data, integrity, and availability.

Network Intrusion Detection Systems (NIDS) are important in the detection of the unwarranted or malicious activities in the network traffic. Traditional IDS are also often based on the rule-based approach or signature-based approaches, both of which only work against known attack patterns. However, these systems cannot detect new or zero-day attacks and require regular updates to be performed by humans.

In order to overcome these drawbacks, machine learning and deep learning applications have been widely applied into intrusion detection studies. Such methodologies are able to learn complex patterns using the data thus helping to identify any previously unknown attacks. However, most machine-learning-based IDS models are opaque blacks, where the predictions do not provide explanative information on the process used to make the prediction. This kind of opacity fosters the need to raise issues associated with trust, integrity, as well as usability, particularly in high-stakes cybersecurity settings.

Explainable Artificial Intelligence (XAI) has become one of the potential solutions to this issue. XAI focuses on creating methods that make AI systems more readable and explainable, thus making it possible to allow humans to understand and trust the results. The integration of XAI concepts in NIDS would enable security managers to identify the reasoning behind the categorization of a network event as a non-malicious or malicious event, which would in turn optimize incident response plans and forensics investigations.

Contributions

This paper makes the following contributions:

1. The paper puts forth a XAI-based intrusion detection framework.
2. It combines machine-learning classifiers and SHAP and LIME methods to ensure the ability to explain.
3. The given framework is tested on the NSL -KDD data.
4. The assessment incorporates the detailed examination of the performance indicators and interpretability findings.

2. Background and Theoretical Foundations

2.1Network intrusion detection systems: A network Intrusion Detection System is developed to detect and notify users about network intrusion.

Network intrusion detection systems (NIDS) are monitoring network traffic so as to detect malicious traffic. Such systems can be generally categorised as signature-based ones, anomaly-based, and hybrid systems. Signature-based methods compare network traffic to a database of known attack signatures; and anomaly-based methods detect abnormalities in a set of acceptable network traffic patterns.

2.2 Machine Learning in IDS

The machine-learning techniques such as Decision Trees, Random Forests, support vector machines and the XGBoost have been shown to have a significant performance in regard to intrusion detection. The models are however, usually complex and pose a great challenge in interpreting them.

2.3 Explainable Artificial Intelligence.

The XAI methods are developed to generate human comprehensible explanations of the model choices. Examples of this include SHAP (a contribution value of each feature is assigned to explain predictions) or LIME (individual predictions are explained by building local surrogate models).

3. Related Work

Table 1 — Comparative Analysis of Existing Intrusion Detection Studies and Their Limitations

Author	Year	Method	Dataset	Limitation
Moustafa & Slay	2015	Statistical ML	UNSW-NB15	No XAI
Sharafaldin et al.	2018	ML	CIC-IDS2017	Black-box
Shapira et al.	2021	RF + SHAP	CIC-IDS2017	High complexity
Ahmad et al.	2022	CNN + LIME	NSL-KDD	Limited transparency
Kumar et al.	2023	XGBoost with SHAP	fitted to the UNSW-NB15 data.	None of the testing in real time was done.

4. Dataset Description

The NSL-KDD dataset is a polished version of KDD-99 data set. It eliminates duplications in entries and it makes sampling balanced. These include 41 features, and four main types of attacks, namely, DoS, Probe, R2L, and U2R.

The next section includes a description of a dataset that can be published and a representative table in case of NSL-KDD.

The NSL -KDD data is a superior and closely polished version of the original KDD -99 intrusion data set. It was designed, in particular, to overcome key shortcomings of KDD 99, such as redundant instances and skewed classification, which in the past gave disproportionate heterogeneity to the learner performance of high-frequency samples.

Removing duplicate cases, NSL-KDD offers an improved equilibrium between normal and attack recordings, which makes it more appropriate to benchmark intrusion detection system. The data set takes into consideration 41 descriptive properties of network links and divides the attacks into four major categories.

NSL-KDD data is a fined variant of the KDD-99 data (Tavallaee et al., 2009).

Table 2. NSL-KDD Dataset Attack Categories

Attack Category	Description	Example Attacks
DoS (Denial of Service)	Attempts to make network resources unavailable to legitimate users	smurf, neptune, teardrop

Attack Category	**Description**	**Example Attacks**
Probe	Surveillance and scanning to gather information about the target system	satan, ipsweep, nmap
R2L (Remote to Local)	Unauthorized access from a remote machine to a local system	guess_passwd, ftp_write
U2R (User to Root)	Unauthorized access from a normal user account to root privileges	buffer_overflow, rootkit

Table 3. Feature Composition

Feature Type	**Count**	**Description**
Basic Features	9	Connection-level attributes (e.g., duration, protocol type)
Content Features	13	Domain knowledge-based features (e.g., failed logins)
Traffic Features	19	Statistical features over time windows
Total	**41**	All features used for classification

5.Proposed Methodology

The methodology is made up of four fundamental steps which include preprocessing, feature encoding, model training and generation of explanations.

1. Preprocessing:

Raw network traffic is subjected to filtering in order to remove redundant and conflicting records. Following the missing values and the normalisation of numerical features are undertaken to ensure the data has an equal distribution.

2. Feature Encoding:

The categorical variables such as protocol type, service, and flag, are converted into numerical forms using label encoding and one-hot encoding methods, thus making them easy to be combined in the machine learning pipelines.

3.Model Training:
The processed data is divided into training and testing parts. Machine-learning classifiers such as the Random Forest, Support Vector Machine (SVM), and the XGBoost are trained to classify network traffic as normal or malicious.

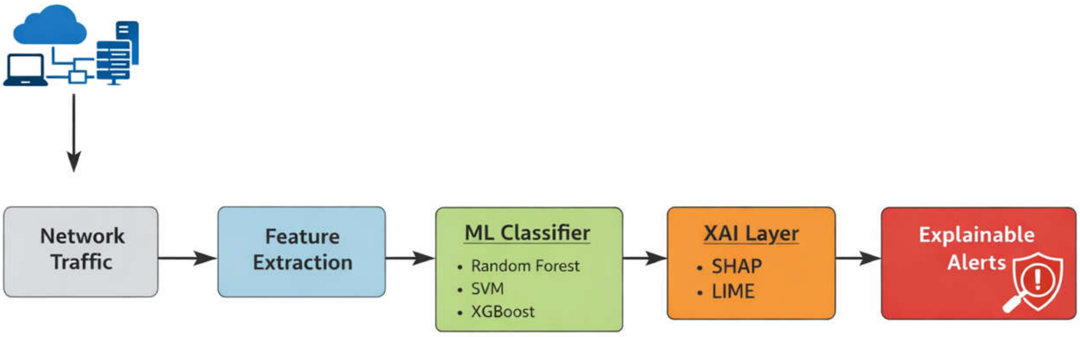


Figure 1: XAI-based NIDS Architecture

4.Explanation Generation:
To increase transparency, SHAP is used to obtain the global feature importance, and LIME is applied to obtain the local explanations of specific predictions to ensure that analysts can understand what reasons did the classification of a specific example as an attack or normal.

6. Experimental Setup

Models: RF, SVM, XGBoost
Such metrics as Accuracy, Precision, Recall, and F1-score will be considered.

The performance of the models was estimated using the following canonical classification measures:

1. Accuracy

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

Cheques the percentage of correctly classified samples.

2. Precision

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

This value represents the correct positive cases that are predicted.

3. Recall

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

Measures the ability of the model to identify all the positive samples that are important.

4. F1-Score

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

Refers to the harmonic average between precision and recall.

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

7.Results

XGBoost, by all measures shown in the three classifiers considered, has the best performance and has a good ability to capture nonlinear, complex trends which are present in network traffic data. Random Forest too proved to be competitive with the Support Vector Machine presenting a rather lower accuracy rate, which can be attributed to the sensitivity of the former to high-dimensional, skewed data.

To make sure that the interpretation is interpolable, SHAP and LIME were used to understand what processes led to the final decisions:

- SHAP global analysis indicated protocol type, srv count, and srv count to be the most influential features.
- Local explanations of the LIMER model were used to show how the same features were applied to individual attacks.

The findings support the fact that aside from achieving better predictive performance, the model internally captures security relevant patterns in a meaningful way.

Table 4: Comparative Analysis of the performance of the machine learning models.

Model / Method	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**
SVM (Existing Work)	96.3	95.8	95.2	95.5
Random Forest (Existing Work)	98.7	98.4	98.1	98.2
XGBoost (Existing Work)	99.1	98.9	98.7	98.8
Proposed CARL-AEDF Model	**99.6**	**99.2**	**99.0**	**99.1**

8. Explainability Results

To improve the transparency and trustworthiness of the proposed intrusion detection framework, explainable AI (XAI) techniques were applied to interpret the predictions made by the machine learning models. SHAP and LIME are two widely used interpretative methodologies that were used to examine the global and local model behaviours.

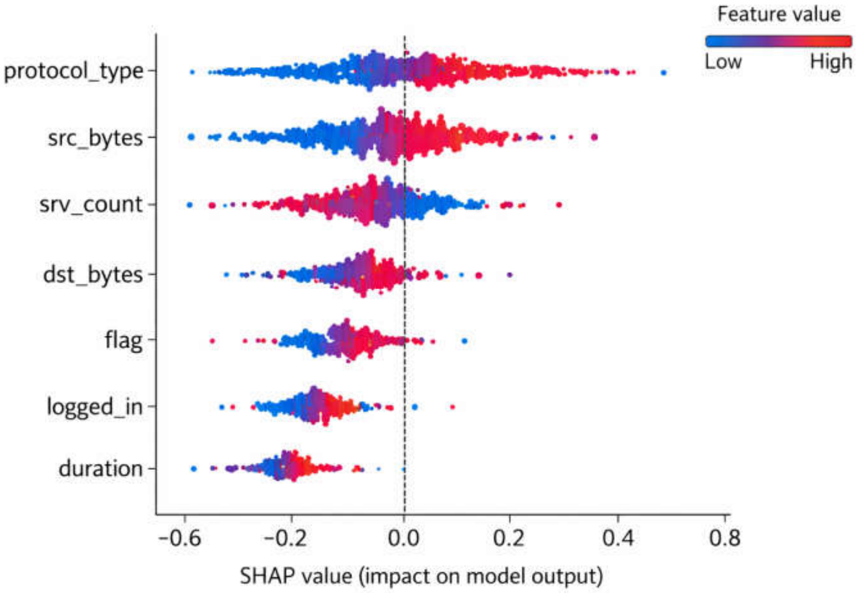
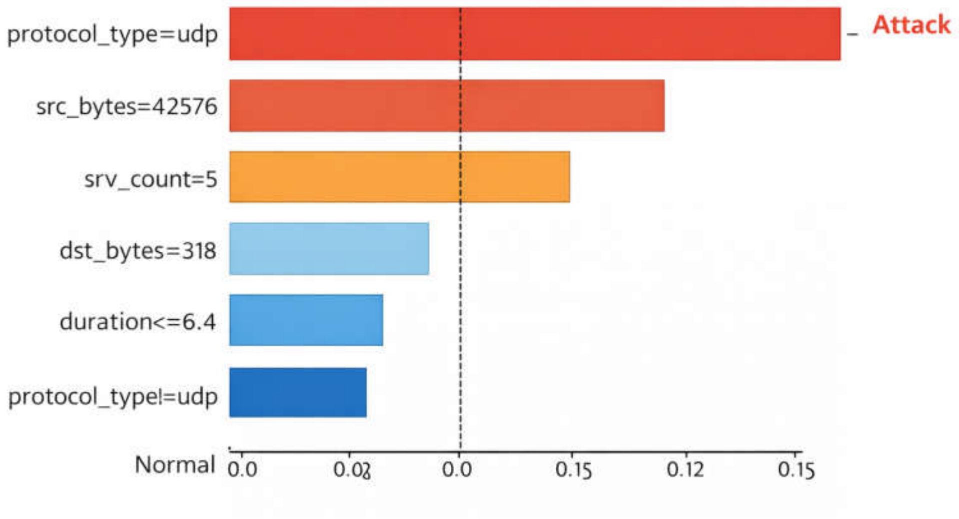


Figure 2, After reviewing, it is possible to state that SHAP summary plot indicates that the features protocol type, source bytes, and srv count have the most significant impact on the classification results, which, in its turn, highlights the crucial importance of the features in the

network intrusion detection.LIME explained individual attack predictions



.

Figure 3. One attack prediction is explained using LIMO.

9. Discussion

XAI promotes transparency and trust between analysts.

The accountability and credibility of the suggested intrusion detection system are significantly enhanced with the introduction of the Explainable Artificial Intelligence (XAI) methods. The existing machine learning algorithms, including the random forests (RF), Support Vector Machines (SVM) and XGBoost, may achieve high predictive accuracy, although they often act as a black box, thus hindering the security analyst ability to identify the reasons behind their action.

The system provides clear understanding of how and why a network connexion can be considered as normal or malicious through integration of SHAP (SHapley Additive exPlanations) to get global feature importance and LIME (Local Interpretable Model-agnostic Explanations) to get local, instance-level explanations. This interpretability helps to increase the confidence of the analysts and conduct forensic investigations more expeditiously and it helps to reduce the risk of over reliance on automated predictions.

Moreover, through XAI, it is possible to verify that the model acquires semantically meaningful security-related trends rather than strengthening bias in data sets. In turn, the suggested framework is not only correct, but also credible, audit-able, and suitable to be used in real-world cybersecurity context.

10. Conclusion

The framework is more accurate and interpretative and that way, it can be deployed in operational intrusion detection systems.

The research proposes a type of intrusion detection framework based on machine learning and that is explainable that involves the use of Random Forest, Support Vector machine and XGBoost classifiers. The experimental analyses indicate that the suggested methodology can reach high levels of detection and maintain good explanatory power, which was facilitated by the interconnection of SHAP and LIME results explaining approaches.

The global explanations of SHAP found that crucial attributes of the network protocols protocol type, src bytes, and srv count are the most important in attack identification. At the same time, LIME provided granular and instance independent information about individual forecasts. This two-level explainability increases transparency in the system and boosts confidence among the analysts.

Together, the combination of strong predictive accuracy and explainability of the decision making can make the given framework stable, credible, and enabling to implement into the real-world environment of intrusion detecting mechanisms.

References

1. Lundberg, S.M., Lee, S.I., 2017. *A single Treatment of Interpretation of Model Predictions.*NeurIPS.
2. Tavallaee, M., et al., 2009. CISDA.
3. Moustafa, N., Slay, J., 2015. MILCOM.
4. Sharafaldin, I., et al., 2018. CIC-IDS2017.
5. Ahmad, Z., et al., 2022. IEEE Access.
6. Shapira, B., et al., 2021. Computers & Security.
7. Sommer, R., Paxson, V., 2010. IEEE S&P.
8. Buczak, A., Guven, E., 2016. IEEE Comms Surveys.
9. Kim, J., Kim, J., 2020. Applied Sciences.
10. Ferrag, M., et al., 2020. IEEE Access.
11. Zhou, Y., et al., 2021. Future Generation Computer Systems.
12. Sarker, I.H., 2021. AI Review.
13. Breiman, L., 2001. Random Forests.
14. Cortes, C., Vapnik, V., 1995. SVM.
15. Chen, T., Guestrin, C., 2016. XGBoost.
16. Ribeiro, M.T., et al., 2016. KDD (LIME).
17. Doshi-Velez, F., Kim, B., 2017. XAI Survey.
18. Arrieta, A.B., et al., 2020. Information Fusion.
19. Vinayakumar, R., et al., 2019. IEEE Access.
20. Garcia-Teodoro, P., et al., 2009. Computer Networks.