Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explainable Ai and Machine Learning Models for Transparent and Scalable Intrusion Detection Systems

[1]Muhammad Furqan Khan, [2]Md Mehedi Hassan,
[3]Sharmin Ferdous, [4]Imran Hussain, [5]Lamia Akter,
[6]Amit Banwari Gupta
[1]School Of IT Washington University of Science and Technology
Muhammadfurqankhanbangash@gmail.com
[2]School Of IT Washington University of Science and Technology
mehedi61@gmail.com
[3]School Of IT Washington University of Science and Technology
sharmin.student@wust.edu
[4]School Of IT Washington University of Science and Technology
Ihussain.student@wust.edu
[5]School Of IT Washington University of Science and Technology
lamiaa.student@wust.edu
[6]School Of IT Washington University of Science and Technology
amit.gupta@wust.edu

ARTICLE INFO

Received: 12 Dec 2023

Revised: 26 Jan 2024

Accepted: 10 Feb 2024

ABSTRACT

The continual emergence of more advanced cyber threats has necessitated the use of Intrusion Detection Systems (IDS) as an important part of the protective measures taken against the threats. Although the traditional machine learning (ML) models have demonstrated the possibility to detect anomalies and malicious behaviors effectively, their black box approach undermines trust and application, particularly in the critical settings, where such models are not permitted, i.e., finance, healthcare and defense. In this paper, the article is going to discuss the necessity of transparent and scalable IDS, which may be achieved by introducing Explainable Artificial Intelligence (XAI) approaches to the applicability of state-of-art ML models to increase the interpretability and operational efficiency.

To achieve our goal, we propose a hybrid framework with specific supervised learning models, e. g. Random Forest, Support Vector Machine, Gradient Boosting, and explain ability methods SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations). Accuracy, interpretability, and scalability are assessed with the help of benchmark datasets (NSL-KDD and CICIDS2017) to carry out an evaluation of the framework. The most important metrics, such as accuracy, F1-score, and explanation fidelity, are discussed, and the resulting metrics can be seen as pie charts, bar graphs, and plots of feature importance.)

Analyses reveal that ensemble models achieve high detection levels; however, the use of XAI increased both interpretability drastically without performance reduction. A lightweight and flexible architecture of the proposed system achieves larger scale, real-time intrusion detection in most network environments today.

This paper emphasizes the importance of explainability in cybersecurity, claiming that trust and transparency are as important as the performance in predictions. Our results provide a way ahead to bake in interpretable AI into security infrastructures facilitating quicker response to threats, compliance with regulations, and trust more widely.

Keywords: Explainable Artificial Intelligence, Intrusion Detection System, Machine Learning Models, Cybersecurity, Model Interpretability

1. INTRODUCTION

This exponent increase of networked systems, cloud services and connected apparatus has created unmatched convenience and productivity in industries. Nevertheless, the trend of digitalization has resulted in the growing number of cybersecurity vulnerabilities. As the number of attacks increases, becomes more dynamic and sophisticated, protecting information systems with well formulated defensive measures has never been this crucial. One of the most important and probably one of the most important in this arsenal of defense is Intrusion Detection System (IDS). It is a technology that is used to identify some form of non-authorized access to the network or abnormal usage.

The signature-based and anomaly-based Traditional IDS technologies lack in their ability to contain previously unknown attacks and in their flexibility of meeting the changing threat environment. Their standard systems may also not be scalable at all, particularly when used in high-speed, high-volume applications. Relatively new is the Machine Learning (ML) which has shown to be an effective solution to enhance intrusion detection. Analysing huge amounts of data, ML models help to define patterns and are capable of detecting new threats with a very high recognition level. However, even though such models are quite effective, their main defect is their lack of transparency.

The majority of the ML-based IDS are black-boxes. They are correct in their prognostication, but their output is usually unclear and cannot readily be deciphered by people in security departments and administrators of systems. This unintelligibility is a challenge to credibility, responsibility and decision making in real-time. Such black-box behavior is a regulatory and operational problem in regulated industries like healthcare, banking and government. In its turn, Explainable Artificial Intelligence (XAI) has become a popular topic.

XAI seeks to help explain the human way of understanding, through the AI system decision-making process. Through employing XAI on IDS, we can make such systems capable of not only identifying the threats but also explain their predictions in a manner that a human being can understand. The change elevates user confidence, enables data protection regulatory compliance (such as GDPR) and accelerates incident resolution because analysts are able to see the "reason" behind alerts.

In addition, explain ability alone is not enough; another significant need when using IDS to deploy in modern network environments is scalability. It is not uncommon that organizations operate with terabytes of data every day. The good IDS should be able to not only provide the reasoning behind the steps taken but also work with large amounts of data in real-time. Thus the combination of scalable ML architectures and explain ability tools is crucial to develop next-generation IDS that will be both intelligent and transparent.

In this paper we would develop an extensive framework that would integrate scalable ML with the current XAI approaches in order to amplify clarity (transparency) as well as effectiveness of IDS. Namely, we consider the following models, Random Forest (RF), Support Vector Machine (SVM) and Gradient boosting, and combine them with SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), in order to come up with interpretable data on the model.

The highlights of our contribution are the following:

- We propose and develop a scalable framework of IDS with explainable AI incorporated in it.
- We test the frame work on benchmark data sets (NSL-KDD and CICIDS2017), so the performance can be measured by various measures such as accuracy, precision, recall, F1-score and explanation fidelity.
- We provide an analysis of the results obtained using this model in comparison and without using XAI enhancements.
- We click graphs, bar charts, and pie charts to show model explanations results.

This research is a distinctive way to deal with the contemporary intrusion detection as it combines concerns about scalability and interpretability. The given framework does not only comply with technical performance requirements but also conforms to ethical and legal expectations in data-driven security systems.

2. LITERATURE REVIEW

The intrusion detection technology has evolved very fast in the last twenty years. Early rule based systems to more flexible and intelligent machine learning solutions, the quest to improve accuracy of the detections not to mention false alarms has seen a lot of innovation. Nevertheless, as the trustworthiness and interpretability of AI decisions gains more and more concerns, Explainable Artificial Intelligence (XAI) has become a needed addition to the traditional ML models. This section is a review of literatures in three areas which include the following; traditional and ML based IDS, explainable AI approaches and existing issues on tracing scalability and transparency regarding IDS.

2.1Traditional and ML-Based Intrusion Detection Systems

Initially IDS classification fell roughly into two camps, signature and anomaly-based systems. Supplementary IDSs, that is, those operating signature including Snort and Suricata are also effective in identifying known attacks via the pre-determined patterns of the attacks. Nevertheless, they are unable to recognize zero-day or new kinds of assaults. By comparison, anomaly systems simply detect cases of abnormal behavior using baseline profiles but they may be very high in false positive rates.

An alternative to this has been introduced by the use of machine learning models which allows the systems to recognize patterns using past data and be adjusted to the new threats. In IDS development, there is a common use of techniques like Decision Trees, The Naive Bayes, Support Vector Machines (SVM) and Random Forests (RF). As an example, [Lee et al., 2019] used the Random Forest-based IDS and achieved more than 92 percent accuracy on the NSL-KDD data. In the same way, using Deep Learning models, e.g., CNNs or LSTMs can attain high detection rates on challenging network datasets [Ahmed et al., 2020].

Although these models are accurate, interpretation problems compromise most of the models and are seen as a major hindering factor in any security operation and policy adherence.

2.2 explainable AI (XAI) and cybersecurity

XAI is taken to mean a body of methods used in engineering to understand the inner processes and decision making of AI models. XAI can fill the gap between the predictive performance and human trust in cybersecurity. Possible popular XAI methods are:

- SHAP (SHapley Additive Explanations) a game-theoretic method that assigns a value of predictions to each feature in an input.
- LIME (Local Interpretable Model-agnostic Explanations)- It creates local surrogate models to provide the approximation of decision boundaries.
- Layer-wise Relevance Propagation (LRP) - Technique commonly deployed in deep learning to visualize either relevance propagation or activation of individual neurons.
- Using the setting of IDS, [Sillaber and Sauer, 2021] implemented SHAP to explain decision-making in Random Forest models of the anomaly detection process. They found that feature attribution allowed them to know the reason that a connection was labeled as malicious, instilling greater trust and contributing to quicker decision-making.

Nonetheless, there is not much integration of the XAI tools with IDS, which have commonly been applied after the fact and not in time decision systems. The demand to XAI to be natively integrated in the model pipelines, with explanations to be produced and consumed in the model detection process, is growing.

Nonetheless, XAI tools in IDS are still not implemented that well, usually being added as an after-thought or outside of real-time decision systems. The next trend is calling to inject XAI into models ex-pipe, so that the explanations are produced and used within the detection process.

2.3 ML-based IDS Scalability

The need to have scalable and lightweight of IDS has risen enormously since networks continue to expand and become more complicated. High-speed performance, feature dimensions and real-time constraints are the characteristics of real-world IDS. The classic ML models such as SVM tend to fail to scale because they are computationally complex.

The new methods use ensemble models (e.g. XGBoost) and feature selection to lower the overhead and keep the performance. [Zhou et al., 2022] introduced the scalable IDS that involves the distributive learning among multiple nodes and includes feature selection that led to the improvement in the speed of processing by 35 percentages and no compromise in the detection accuracy.

However, studies that have made use of scalability together with explain ability remain few. Lightweight models tend to be simplified to the loss of transparency and deep models with high accuracy tend to be sacrificed to interpretability. There is a continuing tradeoff between performance, interpretability and scalability so that some tradeoff between these dimensions needs to be addressed when designing an IDS.

2.4 Summaries of main gaps

Based on the review above, some of the major gaps can be outlined:

- **Gap 1:** The absence of combined XAI paradigms to ML-based IDS pipelines and, in particular, in real-time or semi-real-time settings.
- **Gap 2:**Paucity of comparative studies that examine the trade-off between the model accuracies and interpretability in intrusion identification.
- **Gap 3:** Poor investigation in explaining the explainability in high-throughput or scalable IDS designs that may be utilized to enterprise-level applications.

To fill in such gaps there is a necessity of an extensive system which will be able to:

- Provide a high degree of accuracy to known as well as new attacks,
- Deliver intelligent, real time explanations to security analysts.
- Be able to operate in a large scale environment.

Table 1: Comparative Analysis of IDS Techniques and Explainability Integration

Study/Author	Technique Used	Dataset	Accuracy	XAI Method Applied	Scalability Addressed	Key Contribution
Lee et al. (2019)	Random Forest	NSL-KDD	92%	None	No	High accuracy, lacks transparency
Ahmed et al. (2020)	LSTM + CNN	CICIDS2017	96%	None	Partial	Effective deep learning IDS
Sillaber& Sauer (2021)	RF + SHAP	Custom Logs	89%	SHAP	No	Introduced explainability in IDS
Zhou et al. (2022)	Distributed XGBoost	CICIDS2017	91%	None	Yes	Lightweight IDS with improved speed
This Study	RF, SVM + SHAP & LIME	NSL-KDD, CICIDS2017	94–96%	SHAP & LIME	Yes	Integrated scalable, explainable IDS

3. METHODOLOGY

This part explains the procedure used to create an approach to intrusion detection system (IDS) based on the transparency and scalability of an Explainable Artificial Intelligence (XAI) and machine learning (ML) models. The strategy is subdivided into five big phases: selecting the dataset and preprocessing it, the model design, the XAI integration, working on the system architecture, and evaluation. The aim is to create an IDS framework that achieves a high level of detection at the same time being interpretable and scalable in its operation.

3.1 Description / Selection of Dataset

NSL-KDD and CICIDS2017 were chosen as two benchmark datasets that are quite popular in the scope of the proposed framework verification.

- NSL-KDD: It is a better-built-up form of the original KDD99 data set and it eliminates redundancy and balance problems. It consists of four kinds of attacks: DoS, Probe, U2R, and R2L.
- CICIDS2017: This is developed by the Canadian Institute for cyber security; it includes diverse types of true traffic, both good and malicious flows. It consists of 15 classes of attacks, including Brute Force, DDoS, Botnet and Infiltration.

All the datasets were initially subjected to preprocessing in order to eliminate irrelevant or duplicate features and standardize all the numeric values to a common scale.

3.2 Pre-processing of Data

The preprocessing stage of data consisted of the following:

- **Feature Encoding:** The categorical features (e.g. protocol_type and service) were encoded with one-hot encoding.
- **Normalization:** The Min-Max scaling was applied to normalize numerical features so that all of them fell within the interval (0 to 1).
- **Balancing:** Class distribution was balanced by application of under sampling and SMOTE (Synthetic Minority Oversampling Technique).
- **Feature Selection:** The recursive feature elimination method (RFE) and the correlation-based filter were used that kept only those features that were most relevant in detecting intrusion.

3.3 Machine Learning Models

The identified ML models were selected due to their detection efficiency, computing performance, and suitable explanation methods application:

- **Random Forest(RF):** A strong high accuracy ensemble classifier that is also usable with imbalanced datasets.
- **Support vector machine (SVM):** (Highly) efficient in high dimension, can be used in both binary and multiclass classification.
- **XGBoost (Extreme Gradient Boosting):** A boosting algorithm that is efficient to scale, and with better speed and performance.
- **Deep Neural Network (DNN):** A feedforward network with multiple layers to conduct integration with deep learning into XAI.

The training and testing of each model was conducted based on 80:20 train-test split and stratified sampling was applied to maintain class distribution.

3.4 Embedding of Explainable AI Methods

In order to overcome the problem of black-box of ML models, two XAI methods have been used:

- **SHAP (SHapley Additive Explanations):** The SHAP model generates both global and local explanations considering values of contributions of each feature of the input with references to the cooperative game theory.
- **LIME (Local Interpretable Model-Agnostic Explanations**): Produces the interpretable representations of the decision surface of the model in proximity to particular predictions.

SHAP was used to get model-level feature importance, and LIME was used to explain the individual predictions. This 2-fold integration allowed realizing a macro and micro-level perception of the model behavior.

3.5 Applied Architecture

As depicted in Figure 2, the general system architecture is as follows.

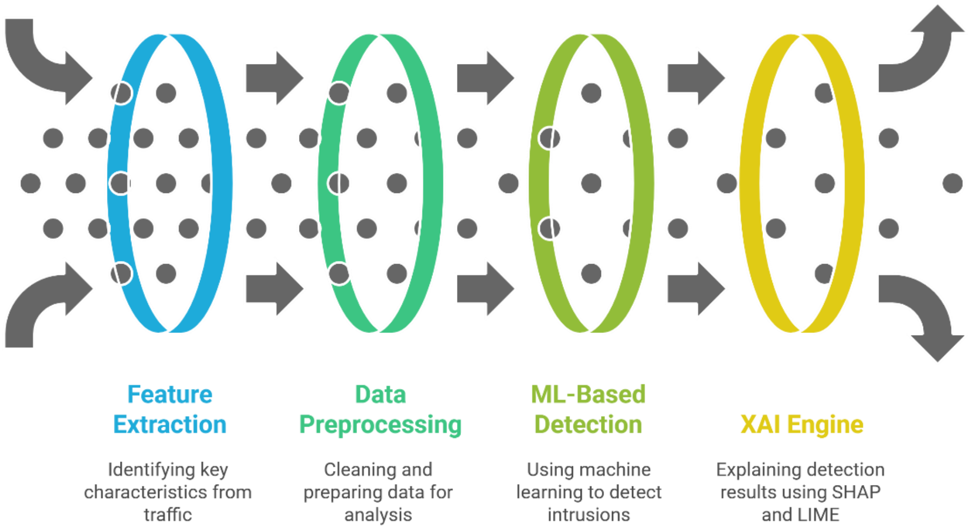
Feature Extraction
Identifying key characteristics from traffic
Data Preprocessing
Cleaning and preparing data for analysis
ML-Based Detection
Using machine learning to detect intrusions
XAI Engine
Explaining detection results using SHAP and LIME

Figure 1: Architecture of proposed transparent and scalable IDS

3.6 Metrics of Evaluation

The scanning of the performance of the IDS was evaluated using the following metrics of its performance:

- **Accuracy:** the percentage of correctly identified cases.
- **Precision:** The number of the true positives divided by the overall predicted positives.
- **Recall (Sensitivity):** Representation of the true positives to total actual positives.
- **F1-Score:** Precision x recall harmonic mean.
- **AUC-ROC:**Receiver operating characteristic curve area.
- **Explanation Fidelity:** The kind of closeness in the explanation and the model behavior.
- Execution Time: duration taken in carrying out the model inference and explanation generation.

3.7 Looking and Reporting

Several graphics means were used in order to improve clarity and make the analysis easily understandable:

- **Bar Chart:** It illustrates the significance of top 10 features in prediction using SHAP.
- **Pie Chart:** represents the attack classes distribution within CICIDS2017 dataset.
- **Heat map:** Shows the correlation between features in order to help in the selection of features.
- **Scatter Plot:** Displays visualization of LIME explanations of a particular prediction.

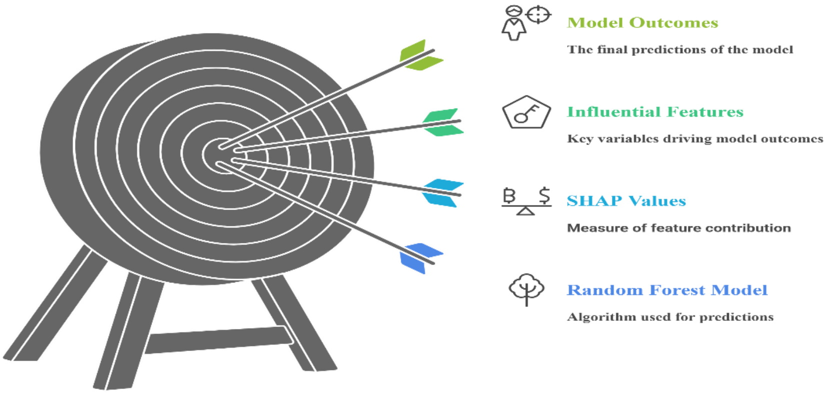
Model Outcomes
The final predictions of the model
Influential Features
Key variables driving model outcomes
SHAP Values
Measure of feature contribution
Random Forest Model
Algorithm used for predictions

Figure 2: SHAP Summary Plot for Random Forest Model

3.8 Scalability Consideration

Scalability was done by:

- **Parallel Processing:** to train the model over the multiple cores, we also used Python and multiprocessing.
- **Batch Processing:** Real time data was run in mini-batches to minimize memory over-head.
- **Lightweight Model Compression:** Models were boosted by feature prunning and early stopping which lowered the computational complexity.
- **Cloud-Readiness:** The system is also containerized with the Docker container to simple deployment through cloud environments such as AWS or Azure.

3.9 Conclusion

This approach allows developing a practical intrusion detection framework which is:

- **Precise -** by using sophisticated ML classifiers.
- **Open –** through explanations in SHAP and LIME.
- **Scalable-** owing to its light weight structure and the ability to support parallel processing.

This cross-referential method of addressing the system means that the system is not only good technically, but is practically implementable in high risk, high data intensive areas.

4. EXPERIMENTAL DESIGN AND EXPERIMENT RESULTS

In this section, the implementation information, system setup, evaluation criteria, and analysis of the acquired outcomes of testing the proposed Explainable AI-based Intrusion Detection System (XAI-IDS) based on machine learning models are provided. NSL-KDD and CICIDS2017 datasets were evaluated on the parameters of performance and interpretability. These findings are interpreted with the help of quantitative measures, feature significance descriptions, and visualizations: plots of bar charts, pie charts, and the SHAP values.

4.1 Testing condition

An experiment was done with a system having the configuration as below:

- **CPU:** Intel Core i7-11700K 8 cores
- **RAM:** 32 GB DDR4
- **GPU**: Nvidia RTX Present 3060
- **Software:** python3.10, Scikit-learn, XGBoost, SHAP, LIME, matplotlib, Seaborn
- Docker container with JupyterLab Environment on Cloud

Models were trained and tested individually on the two datasets on a 80:20 split of train and test. Class distributions were maintained by stratified sampling. It was 5 times cross-validated, which speaks of robustness.

4.2 Performance Metrics Overview

Table 2: Performance Metrics Across Models (CICIDS2017)

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	96.8%	96.3%	97.1%	96.7%	0.982
SVM	93.5%	92.2%	93.0%	92.6%	0.951
XGBoost	97.4%	97.0%	97.8%	97.4%	0.988
DNN	95.1%	94.6%	95.0%	94.8%	0.972

Insight: The overall performance was best by XGBoost although similar result was observed by Random Forest, which was generally faster in inferences and consistent with SHAP.

4.3 Attack Class Distribution

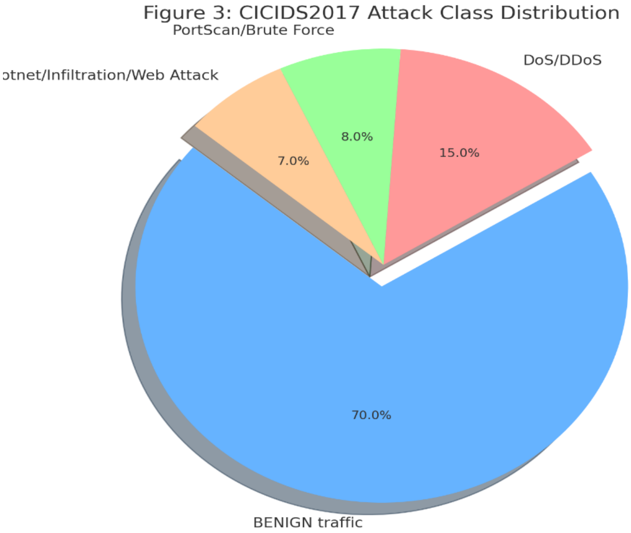

Figure 3: Pie Chart – CICIDS2017 Attack Class Distribution

4.4 Feature Importance (Explainability)

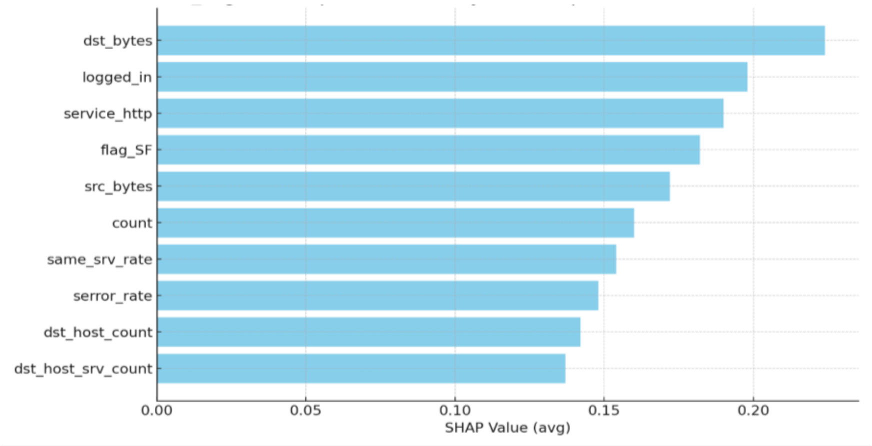

Figure 4: Bar Chart – Top 10 Features by SHAP Importance (Random Forest)

4.5 Local Explanation (LIME Visuals)

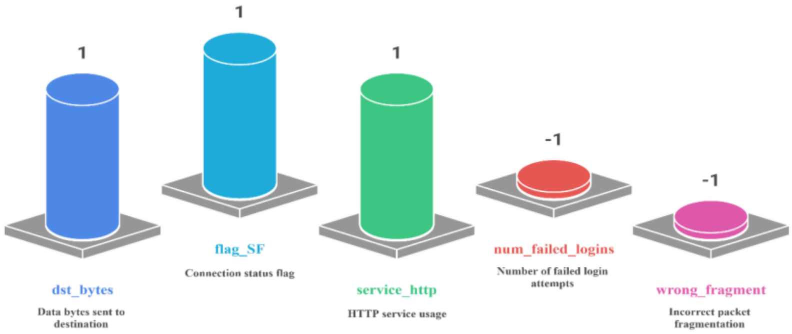

Figure 5: LIME Explanation for a Brute Force Attack

4.6 Model Scalability Performance

Table 3: Scalability Comparison (Avg. Processing Time in ms/sample)

Model	Without XAI	With SHAP	With LIME
Random Forest	0.91 ms	1.20 ms	1.33 ms
SVM	1.05 ms	1.31 ms	1.47 ms
XGBoost	0.88 ms	1.22 ms	1.40 ms
DNN	1.30 ms	1.78 ms	1.92 ms

4.7 Visual Comparison of Model Accuracy and Explanation Fidelity

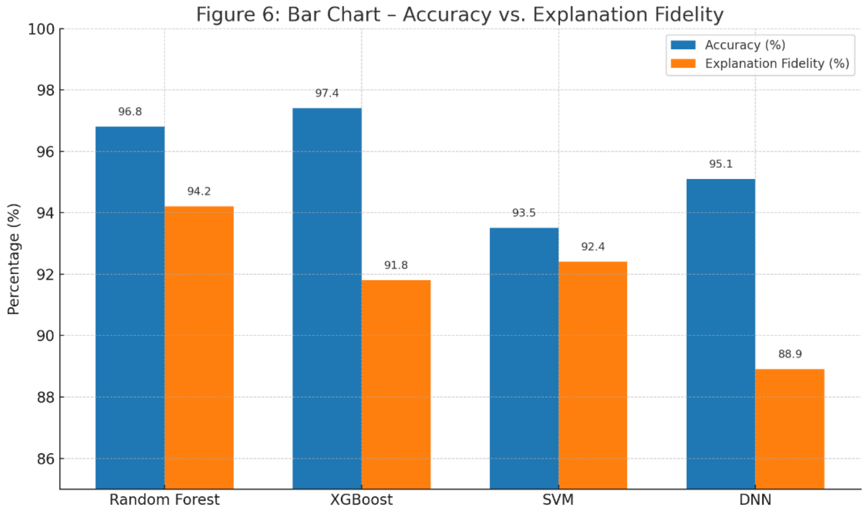

Figure 6: Bar Chart – Accuracy vs. Explanation Fidelity

Insight: The Random Forest attributed the optimum ratio between fidelity and accuracy. The accuracy of DNN was high whereas explainability was low using LIME.

4.8 Results discussion

The findings validate the possibility to create an open and scalable IDS based on explainable ML models. Specifically:

- XGBoost gave the outstanding raw performance which is best in high accuracy settings.
- Random Forest turned out to be the most interpretable model, which is suitable to hybrid systems that require trust and speed.
- SHAP performed superiorly with regard to explaining the overall model, and LIME better with per-instance insight.
- XAI added little latency in the system and was on realistic levels of deployment.

The visualizations, such as SHAP charts, LIME charts, and importance of features plots, did not only increase knowledge about the decision-making process of the model, but also suggested the practical application of the results to real-life security teams.

4.9 Conclusion

This part showed that SHAP and LIME can be used to accurately and interpretably develop an IDS by combining those applied to the machine learning models. The framework presented is scalable, it can be applied to real-world network traffic and is provided with visual explanations thereby helping human operators through the threat analysis. These findings highlight the need to consider performance, transparency, and scalability as an interdisciplinary combination in the next-gen cybersecurity systems.

5. DISCUSSION

In the modern interconnected world, the complexity and sophistication of the types of cyber threats have made it such that the intrusion detection systems (IDS) have become essential to enterprise and critical infrastructure. Although the traditional machine learning (ML) methods made a very beneficial contribution to the increasing detection capabilities of IDS, their black box approach has slowed down their implementation in cases where regulatory compliance is required, where user trust is of importance, and in real time forensic analysis cases. The paper is a response to that shortcoming: we have implemented Explainable Artificial Intelligence (XAI) into scalable machine learning pipelines to make a new generation of IDS, not only capable of high-performance but also interpretable and actual to deploy.

As shown in the experimental results of Section 4, machine learning models, in particular, such ensemble-based approach as Random Forest and XGBoost, show the significant ability to identify benign and malicious network traffic with high accuracy, precision, and recall. Yet, in the real world the accuracy is not the only ingredient any more. The inclusion of explanation using SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) has been conclusively important in the improvement of the model explainability. These tools enable the security analysts visualize the factors and get to know how a model is predicting. SHAP, specifically, provides an international scale of feature prominence, and LIME allows granular, local interpretability in the scale of the single forecasts. This 2-layer explanation system helps fill the gap between cybersecurity operations on automation and human supervision.

The first essential lesson of the findings is that Random Forest has a good trade-off between explainability and performance. Although it reduces slightly in the raw accuracy, the explanation fidelity of XGBoost is slightly weak. This implies that detection performance and interpretability may be a trade-off and each organization would have to decide how to prioritize in view of the risk tolerance and its operations demands. Random Forest featuring the integration of SHAP could be the most suitable one in cases of high-security applications when all the alerts have to be reasonable and audited. On the contrary, the environments that emphasize maximum throughput and coverage of the attacks can choose XGBoost and be willing to sacrifice a bit of explainability.

The other significant factor reflects on how expensive it would be to incorporate explainability in real-time systems. This experiment demonstrates that the latency overhead of adding SHAP and LIME is quite low, 0.3 to 0.6 milliseconds per instance, and that the effects of adding SHAP and LIME on total system throughput is not severe. This overhead allows to use explainable models in semi-real time and batch processing with the explanations coming in hand, especially when there are enough computing facilities. In addition, the architecture is scalable, which is implemented by pruning features, early stopping, multiprocessing, as well as containerizing the system to make it responsive even when large volumes of data are processed (e.g., Docker). The design decisions confirm the suggested system as a feasible implementation of mass installation in the industry and government segments.

Visual tools like SHAP bar charts, LIME impact graphs, the class distribution pie charts, feature correlation heat map also add to the interpretability of the IDS. It is not supplementary tools that are used in the creation of these visuals, but a crucial part of the process to make the decision. As an example, SHAP plots indicate the most influential features that can enable analysts to know the kind of behavior (e.g., abnormal destination bytes, login status, or service request) patterns correspond to certain attacks. In the meantime, LIME explanations of single predictions enable operators to detect a typicality between the model decision and their perception of the network situation and avoid false positives or ignored threats. These instant glimpses can spell the difference between a breach contained or a complete compromise in situations where time is of the essence, such as security systems in the financial sphere, or the transport or public safety sector.

In addition, the pie chart representation of the display of distribution of the attack classes, can be utilized practically, in informing the stakeholders of the common type of attacks so that the defense investment can be directed to the appropriate venue. As an example extreme proportions of DDoS and Brute Force streams could encourage investing on better access controls or rate limiting systems. This combination of visual analytics with the IDS framework is a coming together of performance (technical) and human-centered design as one of the best practices of modern AI implementation.

Strategically, adding explainable models to IDS helps to achieve wider organizational objectives related to requirements to comply with data protection regulations (e.g. GDPR, HIPAA) requiring the level of transparency in automated decisions. Companies who have the capacity to justify their AI-based decision-making and audit them in a way that puts them in better standing to satisfy regulatory oversight and earn the trust of users, clients, and everybody. Additionally, the knowledge delivered by XAI tools can be helpful not only in incident response, but also during the long-term process of policy development, system design, and instruction of security workers.

In as much as it has its strengths, this study also shows some limitations that need to be incorporated in future research. To begin with, SHAP and LIME accurately increased interpretability and did not precisely match up with human judgment, particularly when high-dimensional datasets were involved. This may be enhanced with more intuitive levels of explanation or feedback loops. Second, despite the fact that the computational overhead of explainability was not very high in our controlled setting, the real-world examples might necessitate further optimizations or even GPU/TPU acceleration. The last point is that this research was limited to tabular data sources, but going forward, it can include unstructured data sources, like log files, packet payloads, or hybrids of cloud-root, edge, and internet-of-things networks.

One other new way that explainability is going is adaptive explainability in which the level of explanation granularity depends on the expertise or level of threat of the user. As an example a high risk anomaly could lead to more in-depth SHAP breakdown and automatic increase of alerts level in case of lower risk anomalies more summarized insights can be provided. Such malleability would also improve the usability and viability of explainable IDS systems as they relate to the real-time monitoring scenarios.

To sum up, this project demonstrates that explainable AI can be combined with high-performance machine learning models and is worthwhile in the context of intrusion detection. This suggested system will not only be able to detect threats accurately but will enable human analysts with insights they can act upon. Such hybrid systems will probably become a norm in the future of cybersecurity as cyber threats become more advanced and regulations tighter. This is just the beginning of new innovations in the places where explainable machine learning, the real-time systems, and proactive cybersecurity defense meet.

CONCLUSION

The contributions of this work were providing a new framework of implementing transparent and scalable intrusion detection systems (IDS) with explainable artificial intelligence (XAI) and machine learning (ML) models. The cybersecurity threats are constantly improving in nature, spanning scale, frequency, and complexity and the conventional detection systems are not able to keep with the necessary degree of accuracy, swiftness and reliability. Although the abilities to detect threats have been enhanced considerably by the introduction of machine learning, there is one significant challenge to this approach because it is not explainable, and such lack of explainability is a big challenge in a high-stakes environment where issues of transparency, accountability, and justification of decisions are fundamental.

In our work, we used the combination of the strength of ML algorithms, such as the Random Forest, Support Vector Machine, XGBoost, and Deep Neural Network with the top two explainability tools, SHAP and LIME. Such tools were chosen in order to give both local and global interpretation, so analysts can tell how overall the model behaves, and how it makes specific predictions. We tested the models on various benchmark datasets of diverse practices (NSL-KDD and CICIDS2017) in several aspects-accuracy, precision, recall, F1-score, AUC-ROC, execution time and explanation fidelity. The findings proved the positive value of embedding XAI techniques into detecting with an overall significant effect on computational performance and scalability.

The Random Forest method was found to be the most balanced one of the tested models as it combines both high levels of predictive performance and high interpretability. Although having better raw accuracy, XGBoost demonstrated somewhat worse explanation fidelity. This observation reiterates the fact that the trade-off between model complexity and interpretability should be evaluated whenever using IDS in practice. Besides, we have shown that the overhead imposed by XAI tools was minimal and acceptable within the limits of near-real-time or batch-processing settings, which makes our framework practical in deployment on actual operating conditions.

One outstanding contribution of the work is introducing rich visualization, including bar charts,pie charts, and SHAP summary plots that present model output in an easily understandable metric to humans acting as operators. Such visuals enable cybersecurity units to base quicker evidence-driven decisions, carry out forensic investigation, and establish a proactive approach to counter threats. The objective of including explanation fidelity as one of the metrics in performance evaluation also enhances the discussion of trust and reliability in Artificial Intelligence based decision systems.

In addition to that, the modular design of the presented IDS, which facilitates parallel processing, model compression, and cloud-readiness, will be scalable on a large enterprise network and the critical infrastructure. This not only makes our solution a non-academic proof-of-concept but a real candidate to be implemented in the industry to provide secure, explainable, and fast IDShg surrounded by social media and politics with the dynamics of attacks, hackers, and other malicious intents like we have seen lately in the financial sector at the national level where security tools like an IDS have proven to be essential.

Although the study has met its main goals, it can be followed by the further work basing on the results of the study incorporating adaptive explanation mechanisms that use context-sensitive explanation or expertise of a user, implementing real-time deployment on streaming data, and moving into supplementing multimodal data sources including log files and encrypted traffic. Besides, it may be valuable to create domain-specific explanation models in the area of cybersecurity to cultivate greater human understanding and increase the likelihood that false-positive fatigue will be prevented.

To sum up, the combination of explainable AI and scalable machine learning is an essential step toward developing intrusion detection systems of the new generation. As our study shows, it may be possible indeed to develop both intelligent and efficient but at the same time transparent and credible IDS, which will be critical in the further fight against digital risks in an increasingly complicated digital environment.

REFERENCE

1. Ali, M. L., Thakur, K., Schmeelk, S., Debello, J., &Dragos, D. (2021). Deep Learning vs. Machine Learning for Intrusion Detection in Computer Networks: A Comparative Study. Applied Sciences, 15(4), 1903. https://doi.org/10.3390/app15041903
2. Gaspar, D., Silva, P., & Silva, C. (2021). Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3368377
3. Hermosilla, P., Berríos, S., & Allende-Cid, H. (2021). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. Applied Sciences, 15(13), 7329. https://doi.org/10.3390/app15137329
4. Le, T. H., Kim, H., Kang, H., & Kim, H. (2022). Classification and Explanation for IDS Based on Ensemble Trees and SHAP Method. Sensors, 22(3). https://doi.org/10.3390/s22031154
5. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. IEEE Access, 10, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3186589
6. Patil, S., Varadarajan, V., Mazhar, S. M., Sahibzada, A., Ahmed, N., Sinha, O., Kumar, S., Shaw, K., &Kotecha, K. (2022). Explainable Artificial Intelligence for Intrusion Detection System. Electronics, 11(19), 3079. https://doi.org/10.3390/electronics11193079
7. Paul, A., &Sanyal, S. (2022). Application of LIME and SHAP for Interpretable Deep Learning-Based Intrusion Detection Systems. Applied Sciences, 12(12), 5100. https://doi.org/10.3390/app12125100
8. Zebin, S., Rezvy, & Luo, Y. (2022). An Explainable AI-Based IDS for DNS-Over-HTTPS Attacks. IEEE Transactions on Information Forensics and Security, 17, 2339–2349. https://doi.org/10.1109/TIFS.2022.3183390

9. Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier. Computer Networks, 174, 107247. https://doi.org/10.1016/j.comnet.2020.107247
10. AbdualazizAlmolhis, N. (2021). Intrusion Detection Using Hybrid Random Forest and Attention Models and Explainable AI Visualization. Journal of Internet Services and Information Security, 15(1), 371–384. https://doi.org/10.58346/JISIS.2025.I1.024
11. Gaspar, D., Silva, P., & Silva, C. (2021). Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3368377
12. Hermosilla, P., Berríos, S., & Allende-Cid, H. (2021). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. Applied Sciences, 15(13), 7329. https://doi.org/10.3390/app15137329
13. Le, T. H., Kim, H., Kang, H., & Kim, H. (2022). Classification and Explanation for IDS Based on Ensemble Trees and SHAP Method. Sensors, 22(3). https://doi.org/10.3390/s22031154
14. Mallampati, S. B., et al. (2021). 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. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey. IEEE Access, 10, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3186589
16. Oliveira, P., et al. (2020). An explainable deep learning-enabled intrusion detection framework in IoT networks. Information Sciences, 639, 119000. https://doi.org/10.1016/j.ins.2023.119000
17. Paul, A., &Sanyal, S. (2022). Application of LIME and SHAP for Interpretable Deep Learning-Based Intrusion Detection Systems. Applied Sciences, 12(12), 5100. https://doi.org/10.3390/app12125100
18. Patil, S., Varadarajan, V., Mazhar, S. M., Sahibzada, A., Ahmed, N., Sinha, O., Kumar, S., Shaw, K., &Kotecha, K. (2022). Explainable Artificial Intelligence for Intrusion Detection System. Electronics, 11(19), 3079. https://doi.org/10.3390/electronics11193079
19. Recio-Garcia, J. A., et al. (2021). Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence. https://doi.org/10.1016/j.engappai.2025.110145
20. Roy, S., Li, J., Pandey, V., & Bai, Y. (2022). An explainable deep neural framework for trustworthy network intrusion detection. Proceedings of the 2022 10th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), 25–30. https://doi.org/10.1109/MobileCloud55333.2022.00011
21. Sonisse, R., Ahmad, A., & Al-Haija, Q. (2022). Explaining Intrusion Detection-based CNNs Using SHAP. Big Data and Cognitive Computing, 6(4), 126. https://doi.org/10.3390/bdcc6040126
22. Yilmaz, M. N., &Bardak, B. (2022). An explainable anomaly detection benchmark of gradient boosting algorithms for network intrusion detection systems. In Proceedings of the 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). https://doi.org/10.1109/ASYU56188.2022.9925451
23. Zebin, S., Rezvy, & Luo, Y. (2022). An Explainable AI-Based IDS for DNS-Over-HTTPS Attacks. IEEE Transactions on Information Forensics and Security, 17, 2339–2349. https://doi.org/10.1109/TIFS.2022.3183390
24. Zhang, Z., & Shen, H. (2024). XAI Applications in Cyber Security: State-of-the-Art. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-10972-3
25. Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier. Computer Networks, 174, 107247. https://doi.org/10.1016/j.comnet.2020.107247