International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 12 Dec. 2024, pp: 229-240 www.ijaem.net ISSN: 2395-5252

Explainable AI for cyber security. Improving transparency and trust in intrusion detection systems

Akpan Itoro Udofot, Omotosho Moses Oluseyi, Edim Bassey Edim

Department of Computer Science, Federal School of Statistics, Amechi Uno, Awkunanaw, Enugu, Enugu State
Department of Computer Science, Federal School of Statistics, Sasha Ajibode Road Ibadan, Oyo State Nigeria
Department of Computer Science, Faculty of Physical Sciences, University of Calabar, Cross-River State, Nigeria

Date of Submission: 10-12-2024 Date of Acceptance: 20-12-2024

ABSTRACT
In recent years, the integration of Artificial Intelligence (AI) in cybersecurity has significantly enhanced the capabilities of Intrusion Detection Systems (IDS) to detect and mitigate sophisticated cyber threats. However, the increasing complexity and opaque nature of AI models have led to challenges in understanding, interpreting, and trusting these systems. This paper addresses the critical issue of transparency and trust in IDS by exploring the application of Explainable AI (XAI) techniques. By leveraging XAI, we aim to demystify the decision-making processes of AI-driven IDS, enabling security analysts to comprehend and validate the system's outputs effectively. The proposed framework integrates model-agnostic XAI methods, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), with state-of-the-art IDS algorithms to improve both interpretability and performance. Through comprehensive experiments on benchmark datasets, we demonstrate that our approach not only maintains high detection accuracy but also enhances the explainability of the model's decisions, thereby fostering greater trust among end-users. The findings of this study underscore the potential of XAI to bridge the gap between AI's advanced capabilities and the human need for understanding, ultimately contributing to more secure and reliable cyber defense systems.

Keywords: Explainable AI (XAI), Intrusion Detection Systems (IDS), Cybersecurity, Transparency, Trust, Model-agnostic Explanations, LIME, SHAP

I. INTRODUCTION

Background and Motivation

In today's digital era, the proliferation of cyber threats has necessitated the deployment of advanced security measures to protect sensitive data and critical infrastructure. Artificial Intelligence (AI) has emerged as a transformative tool in cybersecurity, enabling the development of sophisticated Intrusion Detection Systems (IDS) capable of identifying and mitigating potential security breaches in real-time. AI-driven IDS leverage machine learning algorithms to detect anomalies and patterns indicative of malicious activities, thereby enhancing the speed and accuracy of threat detection (Hussain et al., 2021). However, the increasing reliance on complex, "black-box" AI models has raised significant concerns regarding their transparency and trustworthiness, particularly in critical applications such as cybersecurity (Zhang et al., 2020).

Problem Statement

Despite the remarkable advancements AI has brought to IDS, one of the major challenges that persist is the opaqueness of these systems. Traditional AI models used in IDS, such as deep neural networks, often operate as black-boxes, providing little to no insight into how decisions are made. This lack of transparency undermines the trust of security analysts and end-users, making it difficult to justify and validate the decisions made by these systems, especially in high-stakes environments (Samek, Wiegand, and Müller, 2017). Furthermore, the inability to understand the reasoning behind an IDS's decision can lead to challenges in compliance with regulatory standards, which increasingly demand explain ability in AI systems (Rudin, 2019).

Objective

This paper aims to address these challenges by exploring the integration of Explainable AI (XAI) techniques into IDS to improve their transparency and trustworthiness. The objective is to demonstrate that XAI can provide meaningful insights into the decision-making processes of AI-driven IDS, thereby enabling security analysts to interpret and trust the outcomes of these systems. The research focuses on the application of model-agnostic XAI methods, such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), in enhancing the interpretability of IDS without compromising their performance.

Structure of the Paper

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the literature on AI in cybersecurity, XAI techniques, and their application in IDS. Section 3 outlines the methodology used in this research, including the proposed framework for integrating XAI into IDS and the datasets used for evaluation. Section 4 presents the results of the experiments conducted, showcasing the performance metrics of the proposed model, including accuracy, precision, recall, and interpretability scores. Tables and figures are used to illustrate these metrics, highlighting the benefits of XAI in improving IDS transparency. Section 5 discusses the implications of the findings, limitations of the study, and potential avenues for future research. Finally, Section 6 concludes the paper by summarizing the key contributions and the potential impact of XAI on enhancing trust in AI-driven cybersecurity systems.

II. LITERATURE REVIEW

AI in Cyber Security

The rapid evolution of cyber threats has necessitated the adoption of advanced technologies, particularly Artificial Intelligence (AI), to enhance cybersecurity measures. AI has been instrumental in developing sophisticated Intrusion Detection Systems (IDS) capable of detecting and mitigating a wide range of cyberattacks. These systems use machine learning (ML) and deep learning (DL) algorithms to analyze vast amounts of data, identify patterns indicative of malicious activity, and respond in real-time (Kumar et al., 2020). Traditional signature-based IDS, which rely on predefined rules, have become less effective against zero-day attacks and advanced persistent threats (APTs), as they cannot adapt to new, unseen threats. In contrast, AI-driven IDS can learn from data, making them more adaptable and capable of detecting novel attack patterns (Chollet and Allaire, 2018).

One of the key advantages of AI in cybersecurity is its ability to automate threat detection, thereby reducing the reliance on human expertise and improving response times. For instance, deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been successfully applied to network traffic analysis, achieving high detection accuracy and low false positive rates (Shamshirband et al., 2020). Additionally, unsupervised learning techniques, such as clustering and anomaly detection, have been used to identify outliers in network traffic, potentially flagging new types of attacks (Lopez-Martin et al., 2019). However, while AI has significantly advanced the capabilities of IDS, it has also introduced new challenges, particularly concerning the transparency and interpretability of these systems.

Explainable AI (XAI)

Explainable AI (XAI) is an emerging field that seeks to address the opaqueness of AI models by making their decision-making processes more transparent and understandable to humans. As AI systems become more complex and are increasingly deployed in critical domains, such as healthcare, finance, and cybersecurity, the need for explainability has grown. XAI aims to provide insights into how AI models arrive at their predictions, thereby enhancing trust and enabling users to validate the system's outputs (Arrieta et al., 2020).

There are various approaches to achieving explainability in AI models, ranging from intrinsic methods, which involve designing inherently interpretable models, to post-hoc techniques, which seek to explain the decisions of black-box models after they have been made (Guidotti et al., 2018). Among the most widely used post-hoc techniques are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME explains the predictions of any classifier by approximating it locally with an interpretable model, such as a linear model or decision tree (Ribeiro, Singh, and Guestrin, 2016). SHAP, on the other hand, is based on cooperative game theory and provides a unified measure of feature importance by attributing each feature's contribution to a prediction (Lundberg and Lee, 2017).

These techniques are model-agnostic, meaning they can be applied to any AI model, regardless of its complexity. This makes them particularly useful in cybersecurity, where the interpretability of models used in IDS is critical for ensuring that security analysts can trust and act on the system's recommendations.

Current Challenges in IDS

Despite the advances brought about by AI, several challenges remain in the development and deployment of IDS, particularly regarding the transparency and trustworthiness of these systems. One of the primary concerns is the black-box nature of many AI models used in IDS. Models such as deep neural networks are often highly accurate, but their internal workings are not easily understood by humans, making it difficult to explain why a particular decision was made (Samek et al., 2017). This lack of transparency poses significant risks in cybersecurity, where the stakes are high, and decisions need to be justifiable.

Another challenge is the potential for AI models to exhibit biases, which can lead to unfair or incorrect decisions. For example, an IDS trained on imbalanced data may be more likely to flag certain types of network traffic as malicious while ignoring others, leading to false positives or negatives (Zhang et al., 2020). Additionally, the dynamic nature of cyber threats means that IDS must continuously adapt to new attack vectors, requiring models that are not only accurate but also interpretable and explainable.

Furthermore, the use of AI in IDS raises concerns about accountability and compliance with regulatory standards. As regulations such as the General Data Protection Regulation (GDPR) increasingly demand transparency and explainability in automated decision-making systems, there is a growing need to ensure that AI-driven IDS can meet these requirements (Rudin, 2019).

XAI in Cyber Security

The integration of XAI into IDS has been the subject of several recent studies, highlighting the potential of XAI to address the transparency and trust issues associated with AI-driven cybersecurity systems. For instance, Akerkar and Badr (2020) proposed a hybrid approach combining XAI techniques with traditional IDS to improve both the interpretability and effectiveness of the system. Their research demonstrated that using SHAP to explain the output of a deep learning-based IDS allowed security analysts to better understand and trust the model's decisions, leading to improved threat detection performance.

Another study by Kumar et al. (2021) explored the use of LIME to explain the decisions of a random forest-based IDS. Their findings indicated that LIME not only provided valuable insights into the model's behavior but also helped identify potential biases in the training data, which could be addressed to enhance the system's overall accuracy and fairness. Additionally, the study highlighted that integrating XAI into IDS could reduce the cognitive load on security analysts by providing clear, interpretable explanations, thereby improving decision-making efficiency.

Despite these advancements, there are still gaps in the research that need to be addressed. Most studies have focused on applying XAI to relatively simple AI models, and there is a need for further exploration of how XAI can be applied to more complex, deep learning-based IDS. Additionally, while XAI techniques such as LIME and SHAP have shown promise in improving transparency, their computational overhead and scalability in real-time applications remain areas of concern (Bhatt et al., 2020).

XAI Techniques in Cybersecurity: A Comparative Analysis

The implementation of XAI techniques in cybersecurity, particularly in IDS, has been gaining momentum, with various studies exploring different methods to enhance explainability without compromising performance. This section provides a comparative analysis of the key XAI techniques—LIME, SHAP, and other model-agnostic methods—in the context of cybersecurity, focusing on their effectiveness, efficiency, and applicability in real-world scenarios.

Local Interpretable Model-agnostic Explanations (LIME) has been widely adopted due to its versatility and ease of use. LIME works by perturbing the input data and observing changes in the output, allowing it to build a local, interpretable model around each prediction. In cybersecurity, LIME has been successfully applied to explain the decisions of various classifiers, including random forests and support vector machines (SVMs) used in IDS (Kumar et al., 2021). However, while LIME provides valuable insights, it has limitations, particularly in its scalability to large datasets and deep learning models. Additionally, LIME's reliance on local approximations means that its explanations may not always be consistent or stable across different runs, leading to potential trust issues among users (Ribeiro et al., 2016).

SHapley Additive exPlanations (SHAP), based on cooperative game theory, has emerged as a powerful tool for providing consistent and interpretable explanations for complex models. SHAP assigns a unique value to each feature, representing its contribution to the prediction, and offers a unified measure of feature importance. In cybersecurity, SHAP has been particularly effective in explaining the output of deep learning models, such as neural networks, used in IDS (Lundberg and Lee, 2017). Studies have shown that SHAP not only improves transparency but also helps in identifying biases in the training data, making it a valuable tool for enhancing the fairness and reliability of IDS (Bhatt et al., 2020). However, SHAP's computational complexity can be a drawback, especially when applied to large-scale datasets or real-time applications.

Model-agnostic Techniques such as partial dependence plots (PDPs), accumulated local effects (ALE), and feature importance scores have also been explored in cybersecurity. These techniques offer global explanations, providing insights into the overall behavior of the model rather than focusing on individual predictions. For instance, PDPs can illustrate the relationship between a particular feature and the predicted outcome, helping analysts understand how changes in input features influence the model's decisions (Molnar, 2019). ALE, on the other hand, addresses some of the biases inherent in PDPs by accounting for feature interactions and providing more accurate global explanations (Apley and Zhu, 2020). While these methods are useful for understanding model behavior, they may not always be sufficient for explaining complex, high-dimensional data typical in cybersecurity applications.

Comparative Performance Metrics: The following table (Table 1) summarizes the key performance metrics of the discussed XAI techniques when applied to IDS, including their scalability, computational complexity, interpretability, and suitability for different types of AI models.

XAI Technique	**Scalability**	**Computational Complexity**	**Interpretability**	**Suitable AI Models**	**Real-time Applicability**
LIME	Moderate	Moderate	High (local)	Random Forests, SVMs	Limited
SHAP	Low	High	Very High (global)	Deep Learning Models	Limited
PDP	High	Low	Moderate (global)	Various	High
ALE	High	Moderate	High (global)	Various	High

Key Findings and Research Gaps: While XAI techniques like LIME and SHAP have shown promise in improving transparency and trust in IDS, there are still challenges to be addressed. The computational overhead associated with these techniques limits their applicability in real-time cybersecurity scenarios, where quick decision-making is crucial. Moreover, most existing studies have focused on the technical aspects of XAI, with less attention given to the human factors involved in interpreting and trusting these explanations (Kaur et al., 2020). Future research should explore ways to optimize XAI techniques for real-time applications and investigate how different stakeholders, including security analysts, managers, and end-users, perceive and utilize these explanations in their decision-making processes.

Conclusion of Literature Review

The integration of AI in cybersecurity, particularly in the development of IDS, has greatly enhanced the ability to detect and respond to complex cyber threats. However, the black-box nature of many AI models presents significant challenges in terms of transparency and trust, which are critical for ensuring the reliability and accountability of these systems. Explainable AI (XAI) techniques, such as LIME, SHAP, and other model-agnostic methods, offer promising solutions to these challenges by providing interpretable explanations for AI-driven decisions. While current research has demonstrated the effectiveness of these techniques in improving IDS transparency, there remain gaps that need to be addressed, particularly regarding the scalability and real-time applicability of XAI in cybersecurity. Addressing these gaps will be crucial for advancing the field and ensuring that AI-driven cybersecurity systems can be both powerful and trustworthy.

III. METHODOLOGY

Proposed Model

This study proposes an innovative model that integrates Explainable AI (XAI) techniques with Intrusion Detection Systems (IDS) to enhance transparency and trust. The proposed model leverages both classical machine learning algorithms and modern deep learning techniques, augmented by XAI methods such as SHAP and LIME, to provide interpretable outputs.

The proposed model is trained and tested on two benchmark datasets widely used in cybersecurity research: **KDD Cup 99** and **NSL-KDD**.

- **KDD Cup 99**: This dataset is a well-known benchmark for IDS and contains approximately 4.9 million instances with 41 features. It includes various types of attacks, such as Denial of Service (DoS), Probe, and User to Root (U2R) attacks (Tavallaee et al., 2009). Despite criticisms for its redundancy and outdated attack patterns, KDD Cup 99 remains relevant for evaluating IDS due to its extensive use in the literature.
- **NSL-KDD**: A refined version of KDD Cup 99, NSL-KDD addresses some of the criticisms by removing duplicate records and ensuring a more balanced distribution of attack types. It contains 125,973 training instances and 22,544 testing instances (Moustafa and Slay, 2015). NSL-KDD is used alongside KDD Cup 99 to ensure the generalizability and robustness of the proposed model.

The data preprocessing step involves standardization and feature selection, with redundant features removed to enhance the model's efficiency. Table 2 provides a summary of the datasets used in this study.

Data Collection

Dataset	Training Instances	Testing Instances	Features	Attack Types
KDD Cup 99	4,898,431	311,029	41	5
NSL-KDD	125,973	22,544	41	5

Table 2: Summary of datasets used for model training and testing

Algorithms and Techniques

The proposed model utilizes a combination of **Random Forest (RF)** and **Convolutional Neural Networks (CNNs)** as the core algorithms for intrusion detection.

- **Random Forest (RF)**: RF is an ensemble learning method that combines multiple decision trees to improve predictive performance. It is chosen for its interpretability and robustness against overfitting (Breiman, 2001). RF is particularly effective in handling high-dimensional data and is widely used in IDS (Li et al., 2019).
- **Convolutional Neural Networks (CNNs)**: CNNs are deep learning models known for their ability to automatically learn features from raw data. CNNs are employed to capture complex patterns and relationships in the network traffic data that may be missed by traditional methods. The architecture includes several convolutional layers followed by pooling and fully connected layers, optimized using the Adam optimizer (Kingma and Ba, 2015).

The **XAI techniques** applied to these models are:

- **SHapley Additive exPlanations (SHAP)**: SHAP values are computed for the RF and CNN models to provide global and local explanations for each prediction. SHAP's consistency and ability to assign unique importance values to features make it ideal for understanding the contribution of each feature to the prediction (Lundberg and Lee, 2017).
- **Local Interpretable Model-agnostic Explanations (LIME)**: LIME is applied to

generate local explanations by perturbing the input data and observing changes in the model's output. LIME is particularly useful for explaining individual predictions, making it a complementary technique to SHAP in this study (Ribeiro et al., 2016).

Evaluation Metrics

The performance of the proposed model is evaluated using a combination of traditional metrics and novel interpretability scores to assess both detection accuracy and the quality of the explanations provided by the XAI techniques.

1. **Accuracy, Precision, Recall, and F1-Score**: These standard metrics are used to evaluate the effectiveness of the IDS component. Accuracy measures the overall correctness of the model, while precision, recall, and F1-score provide insights into the model's ability to correctly identify intrusions versus normal traffic (Manning et al., 2008).
2. **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: The AUC-ROC score is used to evaluate the model's ability to discriminate between attack and non-attack instances. A higher AUC-ROC indicates better model performance (Fawcett, 2006).
3. **Interpretability Scores**: The explanations generated by SHAP and LIME are evaluated for interpretability using qualitative and quantitative measures. Qualitative measures involve expert assessments of the explanations, while quantitative measures include stability, fidelity, and consistency scores, as suggested by Arya et al. (2019).
4. **Trustworthiness Metric**: A novel metric is introduced to quantify the trust level of users in the model's predictions. This metric is derived from user studies where security analysts rate their trust in the explanations provided by the XAI module (Doshi-Velez and Kim, 2017).

Table 3 provides a summary of the evaluation metrics used in this study.

Metric	Description	Purpose
Accuracy	Overall correctness of the IDS	Evaluate detection performance
Precision	Proportion of true positives among detected positives	Measure model precision
Recall	Proportion of true positives among actual positives	Measure model recall
F1-Score	Harmonic mean of precision and recall	Assess balance between precision and recall
AUC-ROC	Discrimination capability of the IDS	Evaluate model's discriminatory power
Interpretability Score	Expert and quantitative assessment of explanation quality	Evaluate explanation quality
Trustworthiness Metric	User-rated trust in model predictions	Assess user trust in model explanations

Table 3: Summary of evaluation metrics

Experimental Setup

The experiments are conducted on a high-performance computing environment with the following specifications:

- **Processor**: Intel Xeon E5-2670 v3 @ 2.30GHz
- **RAM**: 128 GB
- **GPU**: NVIDIA Tesla K80
- **Software**: Python 3.8, TensorFlow, Scikit-learn, SHAP, and LIME libraries.

The models are trained on the training sets of the KDD Cup 99 and NSL-KDD datasets, with hyperparameters tuned using grid search. The models are validated using 5-fold cross-validation to avoid overfitting and to ensure generalizability.

Hyperparameter Tuning

Hyperparameter tuning is crucial for optimizing model performance. For the **Random Forest** model, the following hyperparameters are tuned:

- **Number of Trees**: The number of decision trees in the forest is varied from 50 to 500 in increments of 50.
- **Max Depth**: The maximum depth of the trees is varied from 10 to 100.

- **Min Samples Split**: The minimum number of samples required to split an internal node is varied from 2 to 10.
- **Max Features**: The number of features to consider when looking for the best split is varied from 'auto', 'sqrt', and 'log2'.

For the **Convolutional Neural Network (CNN)**, the following hyperparameters are optimized:

- **Learning Rate**: Varied from 0.001 to 0.1.
- **Batch Size**: Varied from 32 to 256.
- **Number of Convolutional Layers**: The number of convolutional layers is varied from 2 to 5.
- **Number of Filters**: The number of filters in each convolutional layer is varied from 32 to 128.
- **Dropout Rate**: Varied from 0.1 to 0.5 to prevent overfitting.

The best-performing hyperparameters are selected based on the highest F1-Score obtained during cross-validation.

Explainability Analysis

After training the models, SHAP and LIME are applied to the trained models to generate explanations for their predictions. The analysis focuses on:

1. **Global Explanations**: SHAP values are computed across the entire dataset to understand the overall impact of each feature on the model’s predictions. This provides insights into which features are most influential in detecting intrusions.
2. **Local Explanations**: LIME is used to generate explanations for individual predictions, especially for instances classified as attacks. This allows security analysts to understand why a particular instance was flagged as suspicious.

In addition to SHAP and LIME, feature importance metrics such as Gini importance for the Random Forest model and activation maps for the CNN are analyzed to complement the interpretability assessment.

User Study for Trust Evaluation

To evaluate the trustworthiness of the XAI-enhanced IDS, a user study is conducted involving 30 cybersecurity professionals. The participants are provided with a set of predictions along with the corresponding explanations generated by the XAI module. They are asked to rate their trust in the system’s decisions on a Likert scale from 1 (low trust) to 5 (high trust).

The results of the user study are analyzed using statistical methods such as mean trust scores, standard deviation, and correlation analysis to assess the relationship between explanation quality and trust.

Table 4 presents the average trust scores for different types of explanations generated by SHAP and LIME.

Explanation Type	Mean Trust Score	Standard Deviation
SHAP (Global)	4.2	0.5
SHAP (Local)	4.1	0.6
LIME (Local)	3.8	0.7

Table 4: Average trust scores for different types of explanations.

The findings from the user study are used to refine the XAI module and improve the quality of the explanations, ultimately enhancing the overall trust in the IDS

IV. EXPERIMENTS AND RESULTS

4.1 Experimental Setup

The experiments were conducted in a controlled environment equipped with high-performance computing resources to ensure reliable and reproducible results. The specifications of the experimental setup are as follows:

- **Hardware**:
  - ○ **Processor**: Intel Xeon E5-2670 v3 @ 2.30GHz
  - ○ **RAM**: 128 GB
  - ○ **GPU**: NVIDIA Tesla K80
- **Software**:
  - ○ **Operating System**: Ubuntu 20.04 LTS
  - ○ **Python Version**: 3.8
  - ○ **Libraries and Frameworks**: TensorFlow, Scikit-learn, SHAP, LIME, Matplotlib, Pandas
- **Configuration**:
  - ○ **Dataset**: KDD Cup 99 and NSL-KDD
  - ○ **Cross-Validation**: 5-fold cross-validation to mitigate overfitting
  - ○ **Hyperparameter Tuning**: Grid search for optimal hyperparameters in Random Forest (RF) and Convolutional Neural Network (CNN) models

The models were trained on the training sets of the datasets, with hyperparameters tuned as described in the Methodology section. The evaluation was carried out using both traditional performance metrics (e.g., accuracy, precision, recall) and interpretability metrics derived from SHAP and LIME.

4.2 Results

The results of the experiments are presented in this section, highlighting the performance of the proposed model in terms of accuracy, precision, recall, F1-score, and interpretability.

Table 5 presents the performance metrics of the proposed model on both the KDD Cup 99 and NSL-KDD datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Interpretability Score
KDD Cup 99	RF	98.4%	97.2%	96.8%	97.0%	8.5
	CNN	99.2%	98.5%	98.0%	98.2%	7.8
NSL-KDD	RF	97.6%	96.3%	95.9%	96.1%	8.4
	CNN	98.8%	97.6%	97.2%	97.4%	7.9

Table 5: Performance metrics of the proposed model on KDD Cup 99 and NSL-KDD datasets.

As shown in **Table 5**, the CNN model outperforms the RF model in terms of accuracy and F1-score across both datasets, but the RF model exhibits a higher interpretability score due to its simpler structure and the more straightforward application of SHAP.

4.3 Comparative Analysis

The proposed model's performance was compared with several state-of-the-art models, including Support Vector Machines (SVM), Gradient Boosting Decision Trees (GBDT), and Deep Belief Networks (DBN). The comparison is based on the same datasets, and the results are summarized in **Table 6**.

Model	Dataset	Accuracy	Precision	Recall	F1-Score	Interpretability Score
SVM	KDD Cup 99	96.5%	95.4%	94.7%	95.0%	6.2
GBDT	KDD Cup 99	97.8%	96.8%	96.4%	96.6%	7.1
DBN	KDD Cup 99	98.1%	97.0%	96.7%	96.8%	7.0
Proposed RF	KDD Cup 99	98.4%	97.2%	96.8%	97.0%	8.5
Proposed CNN	KDD Cup 99	99.2%	98.5%	98.0%	98.2%	7.8

Table 6: Comparative analysis of proposed models with other state-of-the-art models.

4.4 Case Studies

To validate the effectiveness of the proposed XAI-enhanced IDS in real-world scenarios, two case studies were conducted using real network traffic from a financial institution and a healthcare provider.

- **Case Study 1: Financial Institution**

In this scenario, the proposed model was deployed to monitor network traffic in a large financial institution. The XAI module provided clear explanations for each flagged intrusion, helping security analysts quickly identify and mitigate threats. The average trust score from the analysts was 4.3, reflecting high confidence in the system’s decisions.

Figure 6 shows the SHAP values for a specific instance where a SQL injection attack was detected.

- **Case Study 2: Healthcare Provider**

The second deployment was at a healthcare provider's network, where the system monitored patient data transmissions for potential breaches. The XAI-enhanced IDS detected an anomalous access pattern that was later confirmed to be an insider threat. The LIME explanations were instrumental in tracing the origin of the breach, and the average trust score was 4.5.

The case studies demonstrate that the proposed model not only improves detection accuracy but also enhances transparency and trust, making it a valuable tool for real-world cybersecurity applications.

V. DISCUSSION

5.1 Interpretation of Results

The results from our experiments demonstrate that integrating Explainable AI (XAI) techniques into Intrusion Detection Systems (IDS) significantly enhances both performance and user trust. As shown in Table 5, the Convolutional Neural Network (CNN) model outperforms the Random Forest (RF) model across key performance metrics, including accuracy, precision, and recall. Specifically, the CNN achieved an accuracy of 99.2% on the KDD Cup 99 dataset, compared to 98.4% for the RF model. This suggests that the CNN is more effective at detecting intrusions with a lower rate of false positives and false negatives.

The interpretability scores, however, indicate a trade-off between accuracy and transparency. The CNN, while more accurate, has a lower interpretability score (7.8) compared to the RF model (8.5). This is in line with findings from Ribeiro et al. (2016), who noted that complex models like CNNs often sacrifice interpretability for performance (Ribeiro et al., 2016). This trade-off underscores the importance of balancing accuracy with the ability to provide understandable explanations for model decisions.

Table 5: Performance Metrics of Proposed Model

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Interpretability Score
KDD Cup 99	RF	98.4%	97.2%	96.8%	97.0%	8.5
	CNN	99.2%	98.5%	98.0%	98.2%	7.8
NSL-KDD	RF	97.6%	96.3%	95.9%	96.1%	8.4
	CNN	98.8%	97.6%	97.2%	97.4%	7.9

Figure 4: Confusion Matrix of CNN Model on NSL-KDD Dataset

The confusion matrix (Figure 4) for the CNN model on the NSL-KDD dataset highlights its high performance in distinguishing between attack and normal traffic. The ability to effectively identify intrusions while minimizing false alarms is critical for the operational effectiveness of IDS.

5.2 Implications for Cyber Security

The integration of XAI into IDS has profound implications for cybersecurity. By providing interpretable explanations for detection results, XAI facilitates better decision-making and trust among security analysts. As highlighted by Singh et al. (2021), explainability in AI models can lead to faster identification of threats and more efficient response strategies (Singh et al., 2021). In practical terms, this means that security teams can more confidently rely on IDS outputs and make informed decisions about mitigating potential threats.

The case studies conducted further validate these benefits. In the financial institution case study, the XAI-enhanced IDS provided clear explanations for detected intrusions, leading to a high trust score from security analysts. Similarly, the healthcare provider case study demonstrated that XAI can help in tracing and addressing insider threats more effectively.

Figure 6: SHAP Values for SQL Injection Detection

Figure 7: LIME Explanation for Insider Threat Detection

These findings align with the broader trend in cybersecurity towards incorporating AI models that are not only effective but also transparent. As cybersecurity threats become more sophisticated, the need for systems that can explain their decision-making processes becomes increasingly critical (Zhang et al., 2022).

5.3 Limitations and Future Research

Despite the promising results, several limitations must be acknowledged. Firstly, the trade-off between model accuracy and interpretability remains a significant challenge. While CNN models provide higher accuracy, their lower interpretability scores may limit their practical applicability in scenarios where explanations are crucial. Future research could explore hybrid models that aim to combine high accuracy with enhanced interpretability.

Additionally, the study was conducted using well-established datasets (KDD Cup 99 and NSL-KDD), which may not fully represent the diverse range of modern cyber threats. Future research should consider evaluating the proposed XAI-enhanced IDS on more recent and varied datasets to assess its robustness in real-world scenarios.

Finally, while this study focused on XAI techniques like SHAP and LIME, there are other explainability methods that could be investigated. For instance, integrating model-specific interpretability techniques with XAI could offer further insights into model behavior (Doshi-Velez & Kim, 2017).

In summary, while the integration of XAI into IDS represents a significant advancement in improving transparency and trust, ongoing research and development are needed to address existing limitations and enhance the overall effectiveness of these systems.

VI. CONCLUSION

6.1 Summary of Findings

This study explored the integration of Explainable AI (XAI) techniques into Intrusion Detection Systems (IDS) to enhance transparency and trust in cybersecurity. Our experiments demonstrated that the Convolutional Neural Network (CNN) model, while achieving superior accuracy (99.2% on the KDD Cup 99 dataset) compared to traditional Random Forest (RF) models, presents a trade-off between performance and interpretability. The CNN model achieved lower interpretability scores (7.8) than the RF model (8.5), indicating that while it provides more accurate intrusion detection, it is less transparent.

The application of SHAP and LIME methods offered valuable insights into the decision-making processes of the IDS. Our case studies in real-world settings, such as financial institutions and healthcare providers, validated the practical benefits of XAI, showing that interpretable explanations enhance trust and efficiency in threat detection and response. The results emphasize the critical role of XAI in improving the usability of IDS by providing clear, understandable reasons for detected anomalies and threats.

Table 5: Performance Metrics of Proposed Model

Dataset	Model	Accuracy	Precision	Recall	F1-Score	Interpretability Score
KDD Cup 99	RF	98.4%	97.2%	96.8%	97.0%	8.5
	CNN	99.2%	98.5%	98.0%	98.2%	7.8
NSL-KDD	RF	97.6%	96.3%	95.9%	96.1%	8.4
	CNN	98.8%	97.6%	97.2%	97.4%	7.9

Figure 4: Confusion Matrix of CNN Model on NSL-KDD Dataset

6.2 Contributions

This paper makes several key contributions to the field of cybersecurity and XAI:

1. **Integration of XAI in IDS**: By incorporating XAI techniques like SHAP and LIME into IDS, this research enhances the interpretability of complex models, bridging the gap between high-performance machine learning models and their usability in practical cybersecurity applications (Ribeiro et al., 2016).
2. **Real-World Validation**: The application of the proposed XAI-enhanced IDS in real-world case studies provides empirical evidence of its effectiveness. These case studies demonstrate that XAI can significantly improve analysts' trust and confidence in the system's outputs, thus supporting better decision-making and threat response (Singh et al., 2021).
3. **Performance Metrics and Comparative Analysis**: The comprehensive performance metrics and comparative analysis presented in the paper offer valuable insights into the trade-offs between accuracy and interpretability. This aids practitioners in selecting the appropriate models based on their specific needs and constraints (Zhang et al., 2022).

Table 6: Comparative Analysis of Proposed Models

Model	Dataset	Accuracy	Precision	Recall	F1-Score	Interpretability Score
SVM	KDD Cup 99	96.5%	95.4%	94.7%	95.0%	6.2
GBDT	KDD Cup 99	97.8%	96.8%	96.4%	96.6%	7.1
DBN	KDD Cup 99	98.1%	97.0%	96.7%	96.8%	7.0
Proposed RF	KDD Cup 99	98.4%	97.2%	96.8%	97.0%	8.5
Proposed CNN	KDD Cup 99	99.2%	98.5%	98.0%	98.2%	7.8

6.3 Future Work

Several avenues for future research emerge from this study:

1. **Hybrid Models**: Investigate hybrid models that combine the high accuracy of deep learning techniques with enhanced

interpretability. This could involve integrating attention mechanisms or interpretable neural network architectures to improve transparency without significantly compromising performance (Doshi-Velez & Kim, 2017).

2. **Diverse Datasets**: Evaluate the proposed XAI-enhanced IDS on more contemporary and diverse datasets to assess its robustness against a broader range of cyber threats. This includes datasets representing emerging attack vectors and real-time network traffic (Li et al., 2019).
3. **Advanced XAI Techniques**: Explore additional XAI techniques beyond SHAP and LIME to further enhance model interpretability. This includes developing new methods that can offer more granular explanations and better align with user requirements (Arya et al., 2019).
4. **User-Centric Evaluations**: Conduct studies focused on user experience and interaction with XAI systems in IDS. Understanding how security analysts interpret and use the explanations provided can lead to improvements in the design and implementation of XAI techniques (Moustafa & Slay, 2015).

In conclusion, integrating XAI into IDS presents a promising approach to improving transparency and trust in cybersecurity. This study lays a foundation for future research and practical implementations, with the potential to enhance the effectiveness and usability of IDS in real-world applications.

REFERENCES

[1]. Akerkar, R., and Badr, Y. (2020) ‘Explainable artificial intelligence for cybersecurity: State of the art and challenges’, IEEE Access, 8, pp. 170356-170373.

[2]. Apley, D.W., and Zhu, J. (2020) ‘Visualizing the effects of predictor variables in black box supervised learning models’, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82(4), pp. 1059-1086.

[3]. Arrieta, A.B., et al. (2020) ‘Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 58, pp. 82-115.

[4]. Arya, V., et al. (2019) ‘One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques’, arXiv preprint arXiv:1909.03012.

[5]. Arya, V., et al., 2019. ‘One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques’. arXiv preprint arXiv:1909.03012.

[6]. Bhatt, U., et al. (2020) ‘Explainable machine learning in deployment’, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 2020. New York: ACM, pp. 648-657.

[7]. Breiman, L. (2001) ‘Random forests’, Machine Learning, 45(1), pp. 5-32.

[8]. Choi, Y., et al., 2020. ‘A deep learning approach to intrusion detection systems based on the NSL-KDD dataset’. Computers, Materials & Continua, 65(2), pp. 1247-1261.

[9]. Chollet, F., and Allaire, J.J. (2018) Deep Learning with R. Shelter Island: Manning Publications.

[10]. Doshi-Velez, F. and Kim, B., 2017. ‘Towards a rigorous science of interpretable machine learning’. arXiv preprint arXiv:1702.08608.

[11]. Doshi-Velez, F., and Kim, B. (2017) ‘Towards a rigorous science of interpretable machine learning’, arXiv preprint arXiv:1702.08608.

[12]. Fawcett, T. (2006) ‘An introduction to ROC analysis’, Pattern Recognition Letters, 27(8), pp. 861-874.

[13]. Guidotti, R., et al. (2018) ‘A survey of methods for explaining black box models’, ACM Computing Surveys (CSUR), 51(5), pp. 1-42.

[14]. Hussain, F., Abbas, H., and Khan, M.A. (2021) ‘Artificial intelligence-based intrusion detection systems in the era of industrial internet of things’, Computers & Security, 105, p. 102235.

[15]. Kaur, H., et al. (2020) ‘Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning’, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, USA, April 2020. New York: ACM, pp. 1-14.

[16]. Kingma, D.P., and Ba, J.L. (2015) ‘Adam: A method for stochastic optimization’, arXiv preprint arXiv:1412.6980.

[17]. Kumar, R., et al. (2021) ‘Towards explainable AI in cybersecurity through feature-based anomaly detection’, IEEE

Transactions on Dependable and Secure Computing, 18(3), pp. 1060-1074.

[18]. Kumar, S., et al. (2020) ‘Machine learning algorithms for cybersecurity applications: Challenges and opportunities’, ACM Computing Surveys (CSUR), 53(4), pp. 1-36.

[19]. Kumar, S., et al. (2020) ‘Machine learning algorithms for cybersecurity applications: Challenges and opportunities’, ACM Computing Surveys (CSUR), 53(4), pp. 1-36.

[20]. Li, Y., et al. (2019) ‘A new intrusion detection system based on GBDT optimized by PSO’, Computers & Security, 88, p. 101645.

[21]. Liu, L., et al., 2023. ‘A comparative study of machine learning techniques for intrusion detection systems’. Journal of Information Security and Applications, 68, p. 103027.

[22]. Lopez-Martin, M., et al. (2019) ‘A neural network IDS for the detection of Android malware based on the characteristic difference between benign and malicious traffic’, Computers & Security, 87, p. 101561.

[23]. Lundberg, S.M., and Lee, S.I. (2017) ‘A unified approach to interpreting model predictions’, Advances in Neural Information Processing Systems, 30, pp. 4765-4774.

[24]. Lundberg, S.M., and Lee, S.I. (2017) ‘A unified approach to interpreting model predictions’, Advances in Neural Information Processing Systems, 30, pp. 4765-4774.

[25]. Manning, C.D., Raghavan, P. and Schütze, H., 2008. Introduction to Information Retrieval. Cambridge: Cambridge University Press.

[26]. Manning, C.D., Raghavan, P., and Schütze, H. (2008) Introduction to Information Retrieval. Cambridge: Cambridge University Press.

[27]. Moustafa, N. and Slay, J., 2015. ‘UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)’. Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS). Canberra, Australia, November 2015. IEEE, pp. 1-6.

[28]. Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. ‘"Why should I trust you?" Explaining the predictions of any classifier’. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, August 2016. New York: ACM, pp. 1135-1144.

[29]. Rudin, C. (2019) ‘Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead’, Nature Machine Intelligence, 1(5), pp. 206-215.

[30]. Samek, W., et al. (2017) ‘Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models’, ITU Journal: ICT Discoveries, 1(1), pp. 1-10.

[31]. Samek, W., Wiegand, T., and Müller, K.R. (2017) ‘Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models’, ITU Journal: ICT Discoveries, 1(1), pp. 1-10.

[32]. Shamshirband, S., et al. (2020) ‘Deep learning-based anomaly detection in smart grids: A systematic review’, Future Generation Computer Systems, 113, pp. 171-190.

[33]. Singh, S., et al., 2021. ‘On the performance of explainable AI for cyber security: Insights from real-world deployments’. IEEE Transactions on Information Forensics and Security, 16, pp. 1840-1853.

[34]. Tavallaee, M., et al. (2009) ‘A detailed analysis of the KDD CUP 99 data set’, Proceedings of the 2009 IEEE Symposium on Computational Intelligence

[35]. Zhang, C., et al (2020) ‘A framework for applying explainable artificial intelligence in cyber-physical systems security’, IEEE Transactions on Industrial Informatics, 16(10), pp. 6562-6570.

[36]. Zhang, J., et al (2022). ‘Explainable AI techniques for network security: A review and a novel framework’. Computers & Security, 107, p. 102389.