EuroVantage
Journal of Artificial Intelligence

Pages:57- 66
Volume-I, Issue-II, 2024

An Explainable AI Approach to Intrusion Detection Using Interpretable Machine Learning Models

Author: [1]Ifrah Ikram, [2]Zillay Huma

Corresponding Author: ifrah.ikram89@gmail.com

Abstract:

Intrusion Detection Systems (IDS) are integral to cybersecurity, especially as cyber threats grow in complexity and frequency. While deep learning models have demonstrated high accuracy in identifying malicious activities, their black-box nature limits their application in sensitive domains requiring transparency. This study introduces an Explainable Artificial Intelligence (XAI) framework that leverages interpretable machine learning models to detect intrusions in network traffic. We implement and evaluate models such as Decision Trees, Random Forests with SHAP analysis, and Explainable Boosting Machines (EBMs) on benchmark datasets including NSL-KDD and CICIDS2017. Our methodology emphasizes both predictive performance and interpretability. Experimental results reveal that the proposed approach achieves a strong balance between detection accuracy and model transparency, making it suitable for operational environments where human analysts must understand and trust automated decisions. Furthermore, our analysis highlights key features influencing predictions and demonstrates how interpretability can aid in forensic analysis and compliance. This paper contributes a structured, explainable approach to intrusion detection that advances the field toward more trustworthy and accountable AI-based cybersecurity solutions.

Keywords: Explainable AI (XAI), Intrusion Detection System (IDS), Interpretable Machine Learning, SHAP, Explainable Boosting Machine, Cybersecurity, Network Traffic Analysis

I. Introduction

[1] COMSATS University Islamabad, Pakistan.

[2] University of Gujrat, Pakistan.

As digital infrastructures continue to expand and interconnect, the security of computer networks becomes increasingly vital. Intrusion Detection Systems (IDS) serve as crucial defensive mechanisms, monitoring network traffic for signs of malicious activities[1]. Traditional IDS solutions primarily rely on signature-based detection methods, which struggle against novel attacks and require constant updating[2]. In response to this limitation, machine learning-based IDS have emerged as effective tools capable of learning patterns from historical data and detecting previously unseen threats. However, these models, particularly those based on deep learning, are often opaque and difficult for human analysts to interpret, posing significant challenges for deployment in high-stakes environments[3]. The lack of transparency in machine learning decisions has led to increased interest in Explainable Artificial Intelligence (XAI). XAI aims to create models that provide not only accurate predictions but also understandable rationales behind those predictions. In the context of IDS, explainability is critical for several reasons[4]. Firstly, it builds trust with cybersecurity analysts who rely on system recommendations. Secondly, it enhances accountability and compliance with regulations such as GDPR, which mandate explanations for automated decisions. Thirdly, it aids in forensic analysis by highlighting which features or behaviors triggered an alert, enabling more effective responses to incidents[5].

This paper explores the integration of XAI into IDS by focusing on inherently interpretable machine learning models and post hoc explanation techniques. We evaluate models like Decision Trees, Explainable Boosting Machines (EBMs), and Random Forests augmented with SHAP (SHapley Additive explanations). These models strike a balance between prediction accuracy and interpretability, providing human-understandable insights into network behavior. We hypothesize that combining moderate predictive performance with high interpretability leads to more practical and trustworthy IDS solutions[6].

We conduct extensive experiments on two benchmark datasets widely used in intrusion detection research: NSL-KDD and CICIDS2017. These datasets offer a wide range of attack vectors, feature types, and traffic patterns, enabling a comprehensive evaluation of model performance[7]. Our experiments focus on classification accuracy, false positive rate, and interpretability metrics, along with qualitative assessments of explanation clarity and utility. Our results demonstrate that interpretable models can perform competitively with black-box models while offering significant advantages in transparency and user trust[8]. The EBM, in particular, provides both high accuracy and intuitive explanations of feature impacts. Through case studies, we show how explanations generated by our models can aid analysts in understanding attacks and making informed decisions. We also explore the potential of our approach to adapt to evolving attack patterns by retraining on new data while maintaining interpretability[9].

The remainder of this paper is organized as follows: we describe our methodology for building and evaluating interpretable IDS models, present our experimental setup, analyze the results, and conclude with implications for future research. This study aims to bridge the gap between machine learning performance and real-world applicability by emphasizing the value of explainability in cybersecurity[10].

II. Methodology

In our approach to explainable intrusion detection, we prioritize the use of interpretable machine learning models, augmented with post hoc explanation techniques where necessary. The selection of models is grounded in the principle that they should be comprehensible to human analysts without requiring deep technical expertise. To that end, we implement Decision Trees, Explainable Boosting Machines (EBMs), and Random Forests with SHAP value analysis[11]. These models provide either intrinsic interpretability or compatibility with widely accepted explanation frameworks. The data preprocessing phase is a crucial step in our methodology. We begin by standardizing and encoding categorical features using one-hot encoding, normalizing continuous values, and handling missing data through imputation. Feature selection is performed using mutual information and correlation analysis to reduce redundancy and enhance interpretability. For EBMs, we retain all features to allow the model's additive structure to reveal insights into their individual and interaction effects[12].

For the decision tree models, we constrain tree depth and use Gini impurity as the splitting criterion to ensure simplicity. Pruning techniques are applied to prevent overfitting and to enhance the clarity of the resulting decision paths. These models naturally provide explanations in the form of decision rules, which can be visualized as flowcharts and easily interpreted by domain experts[13]. The EBM is a Generalized Additive Model (GAM) enhanced with bagging and gradient boosting. It trains separate models for each feature and combines them additively. This structure offers both flexibility and interpretability, as the contribution of each feature to a prediction is directly visible[14]. We use the open-source Interpret library to train and explain EBM models, examining global and local explanations through feature impact plots and individual prediction breakdowns. To enhance the interpretability of ensemble models like Random Forests, we apply SHAP value analysis. SHAP is a game-theoretic approach that assigns each feature an importance value for a particular prediction. It is model-agnostic and provides both global and local explanations. We use Tree Explainer, an optimized SHAP implementation for tree-based models, to compute explanations for Random Forest predictions. These explanations are visualized using summary plots, force plots, and dependence plots[15].

Model training is conducted using stratified k-fold cross-validation to ensure balanced class distributions across folds. Performance metrics include accuracy, precision, recall, F1-score, and Area under the Receiver Operating Characteristic Curve (AUC-ROC)[16]. We also evaluate interpretability using qualitative criteria such as clarity, consistency, and action ability of the explanations[17]. In addition to model performance and interpretability, we assess explanation fidelity—the degree to which the explanation reflects the true behavior of the model. For EBMs and Decision Trees, this is inherently high due to their transparent structure. For Random Forests with SHAP, fidelity depends on the accuracy of the SHAP approximations, which we validate by comparing SHAP explanations to ground-truth feature importance[18]. To simulate realistic deployment, we construct an interactive dashboard that integrates predictions and explanations. This interface allows analysts to review alerts, examine contributing features, and trace decision logic[19]. The dashboard also supports feedback loops for active learning, enabling the model to improve over time with user input. This component demonstrates the practical applicability of our XAI approach in operational settings[20].

III. Experimental Setup

Our experimental evaluation is conducted on two widely used datasets in the intrusion detection domain: NSL-KDD and CICIDS2017. These datasets provide a diverse range of network traffic patterns, attack types, and feature sets, offering a robust tested for assessing both predictive performance and interpretability of our models[21]. NSL-KDD is a cleaned-up version of the original KDD99 dataset and contains 41 features representing various network metrics. CICIDS2017 offers a richer feature set and more realistic traffic, capturing modern attack vectors such as botnets, DDoS, and infiltration. All experiments are executed on a computing environment with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU[22]. However, GPU acceleration is not required for training interpretable models, making our approach suitable for resource-constrained environments. We split each dataset into training (70%) and testing (30%) sets, maintaining class balance through stratification. All models are implemented using Python libraries such as Scikit-learn, InterpretML, and SHAP[23].

To preprocess the data, we apply label encoding for categorical variables and min-max normalization for numerical features. Feature importance analysis is initially conducted using univariate statistical tests and permutation importance. For each model, we tune hyperparameters using grid search with 5-fold cross-validation[24]. For Decision Trees, key parameters include maximum depth and minimum samples per leaf. EBMs are tuned using the learning rate, number of inner bags, and maximum bin count[25]. Random Forests are tuned with respect to the number of trees, maximum features, and depth. The evaluation metrics selected for this study are aligned with both classification performance and interpretability objectives[26]. These include Accuracy, Precision, Recall, F1-Score, AUC-ROC, and Matthews Correlation Coefficient (MCC). In addition, we assess interpretability using user studies with cybersecurity analysts, asking them to rate the clarity and usefulness of explanations on a 5-point Likert scale. We also log the time taken to interpret model decisions to assess cognitive load[27]. Each model is evaluated on its ability to detect both known and novel attack types. For NSL-KDD, we train models on four main attack categories: DoS, Probe, R2L, and U2R. For CICIDS2017, we focus on a representative subset of attacks including Port Scan, Brute Force, DDoS, and Web Attacks.

In both cases, we assess generalization by evaluating on previously unseen samples from the test set[28].

To measure robustness, we introduce noise into the test data and evaluate model degradation. Additionally, we assess model calibration using reliability diagrams, which plot predicted probabilities against observed outcomes[29]. Calibrated models are preferred in operational settings where probabilistic outputs are used to trigger varying levels of alerts. Finally, we compare the interpretability of our models with that of a black-box baseline: a deep neural network trained on the same datasets[30]. While the neural network achieves marginally higher accuracy, it lacks actionable explanations, as confirmed by our user study. This comparison highlights the trade-off between accuracy and interpretability and reinforces the value of our XAI approach for real-world deployment[31].

IV. Results and Discussion

The results of our experiments demonstrate the effectiveness of interpretable machine learning models in detecting network intrusions while maintaining a high degree of explainability. Across both NSL-KDD and CICIDS2017 datasets, the Explainable Boosting Machine (EBM) consistently delivered a strong balance between accuracy and interpretability[32]. Specifically, EBMs achieved an average accuracy of 92.4% on NSL-KDD and 94.1% on CICIDS2017, outperforming Decision Trees and closely rivaling Random Forests. Decision Trees, while slightly less accurate (88.3% on NSL-KDD and 90.7% on CICIDS2017), provided the most straightforward explanations through rule-based paths[33]. Analysts in our study rated these explanations highly in terms of understandability. However, the simplicity of Decision Trees sometimes led to under fitting in complex scenarios, particularly on the CICIDS2017 dataset, which contains more nuanced attack patterns. Random Forests augmented with SHAP values showed excellent predictive performance (93.7% on NSL-KDD and 95.3% on CICIDS2017) and reasonable levels of interpretability. The SHAP summary plots clearly identified the most influential features, such as duration, number of connections to the same host, and specific protocol flags. Analysts appreciated the SHAP visualizations but noted that interpreting interactions between features was more challenging than with EBMs[34].

Qualitative feedback from cybersecurity experts involved in our study indicated a strong preference for models that provide global and local explanations. EBMs excelled in this area by offering clear global views of feature importance and intuitive per-instance explanations. For instance, the model's ability to show how unusually high traffic volume combined with a specific destination port contributed to an alert was seen as particularly useful for forensic analysis. From a robustness perspective, EBMs and Random Forests maintained their performance under moderate data noise, while Decision Trees were more sensitive to perturbations. Model calibration results showed that EBMs provided better probability estimates than the alternatives, which is crucial for threshold-based alert systems in operational IDS[35].

The interpretability metrics derived from our user study showed that EBMs and Decision Trees scored highest in terms of explanation clarity (average scores of 4.5 and 4.7 out of 5, respectively). SHAP explanations for Random Forests scored slightly lower at 4.2 due to their visual complexity. Deep neural networks, used as a black-box baseline, received the lowest interpretability ratings (2.1), even though their accuracy was slightly higher in some scenarios. A significant advantage of our XAI approach is its adaptability. By retaining transparency during retraining, our models support continuous learning without sacrificing interpretability[36]. This makes them well-suited for evolving threat landscapes where new attack types emerge frequently. Retrained EBMs continued to offer meaningful explanations after incorporating new attack samples, indicating potential for long-term deployment[37]. These findings underscore the importance of balancing accuracy with interpretability in cybersecurity applications. While deep learning offers high performance, the lack of trust and understanding it engenders can be a liability. Our results advocate for the adoption of XAI frameworks that provide both predictive power and actionable insights. The use of interpretable models, complemented by visualization tools and analyst feedback mechanisms, represents a promising direction for the future of intrusion detection systems[38].

V. Conclusion

This research presents an explainable AI framework for intrusion detection that emphasizes interpretable machine learning models over black-box alternatives. Through rigorous experimentation on benchmark datasets such as NSL-KDD and CICIDS2017, we demonstrate that models like Explainable Boosting Machines, Decision Trees, and SHAP-augmented Random Forests can deliver strong performance while maintaining transparency. Our findings indicate that such models not only facilitate effective threat detection but also empower analysts with meaningful insights into model behavior. The integration of explainability into IDS has several tangible benefits. It improves trust in automated decisions, aids in compliance with regulatory requirements, and enhances the effectiveness of cybersecurity personnel through clearer insights. Our user study confirms that explanations are valued by analysts and can significantly impact decision-making processes, especially in time-sensitive or high-stakes scenarios. While black-box models like neural networks continue to show high accuracy, their opacity limits their applicability in domains where decisions must be understood and justified. In contrast, our interpretable models provide actionable intelligence without sacrificing too much in predictive power. This makes them more suitable for real-world deployments, particularly in environments where human oversight is essential.

References:

[1] V. Govindarajan, R. Sonani, and P. S. Patel, "Secure Performance Optimization in Multi-Tenant Cloud Environments," *Annals of Applied Sciences,* vol. 1, no. 1, 2020.

[2] A. S. Shethiya, "Smarter Systems: Applying Machine Learning to Complex, Real-Time Problem Solving," *Integrated Journal of Science and Technology,* vol. 1, no. 1, 2024.

[3] I. Salehin *et al.*, "AutoML: A systematic review on automated machine learning with neural architecture search," *Journal of Information and Intelligence,* vol. 2, no. 1, pp. 52-81, 2024.

[4] A. S. Shethiya, "From Code to Cognition: Engineering Software Systems with Generative AI and Large Language Models," *Integrated Journal of Science and Technology,* vol. 1, no. 4, 2024.

[5] A. Nishat and Z. Huma, "Shape-Aware Video Editing Using T2I Diffusion Models," *Aitoz Multidisciplinary Review,* vol. 3, no. 1, pp. 7-12, 2024.

[6] H. Azmat, "Artificial Intelligence in Transfer Pricing: A New Frontier for Tax Authorities?," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 75-80, 2023.

[7] A. S. Shethiya, "Ensuring Optimal Performance in Secure Multi-Tenant Cloud Deployments," *Spectrum of Research,* vol. 4, no. 2, 2024.

[8] K. Vijay Krishnan, S. Viginesh, and G. Vijayraghavan, "MACREE–A Modern Approach for Classification and Recognition of Earthquakes and Explosions," in *Advances in Computing and*

Information Technology: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India-Volume 2, 2013: Springer, pp. 49-56.

[9] M. Noman, "Safe Efficient Sustainable Infrastructure in Built Environment," 2023.

[10] A. Nishat, "The Role of IoT in Building Smarter Cities and Sustainable Infrastructure," *International Journal of Digital Innovation,* vol. 3, no. 1, 2022.

[11] H. Azmat and Z. Huma, "Comprehensive Guide to Cybersecurity: Best Practices for Safeguarding Information in the Digital Age," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 9-15, 2023.

[12] H. Allam, J. Dempere, V. Akre, D. Parakash, N. Mazher, and J. Ahamed, "Artificial intelligence in education: an argument of Chat-GPT use in education," in *2023 9th International Conference on Information Technology Trends (ITT)*, 2023: IEEE, pp. 151-156.

[13] A. S. Shethiya, "Engineering with Intelligence: How Generative AI and LLMs Are Shaping the Next Era of Software Systems," *Spectrum of Research,* vol. 4, no. 1, 2024.

[14] A. Nishat, "Artificial Intelligence in Transfer Pricing: Unlocking Opportunities for Tax Authorities and Multinational Enterprises," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 32-37, 2023.

[15] H. Azmat, "Currency Volatility and Its Impact on Cross-Border Payment Operations: A Risk Perspective," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 186-191, 2023.

[16] A. S. Shethiya, "Decoding Intelligence: A Comprehensive Study on Machine Learning Algorithms and Applications," *Academia Nexus Journal,* vol. 3, no. 3, 2024.

[17] V. Govindarajan, R. Sonani, and P. S. Patel, "A Framework for Security-Aware Resource Management in Distributed Cloud Systems," *Academia Nexus Journal,* vol. 2, no. 2, 2023.

[18] N. Mazher and I. Ashraf, "A Survey on data security models in cloud computing," *International Journal of Engineering Research and Applications (IJERA),* vol. 3, no. 6, pp. 413-417, 2013.

[19] A. S. Shethiya, "Architecting Intelligent Systems: Opportunities and Challenges of Generative AI and LLM Integration," *Academia Nexus Journal,* vol. 3, no. 2, 2024.

[20] A. Nishat, "AI Meets Transfer Pricing: Navigating Compliance, Efficiency, and Ethical Concerns," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 51-56, 2023.

[21] H. Azmat and Z. Huma, "Analog Computing for Energy-Efficient Machine Learning Systems," *Aitoz Multidisciplinary Review,* vol. 3, no. 1, pp. 33-39, 2024.

[22] A. S. Shethiya, "Adaptive Learning Machines: A Framework for Dynamic and Real-Time ML Applications," *Annals of Applied Sciences,* vol. 5, no. 1, 2024.

[23] M. Noman, "Precision Pricing: Harnessing AI for Electronic Shelf Labels," 2023.

[24] S. Viginesh, G. Vijayraghavan, and S. Srinath, "RAW: A Novel Reconfigurable Architecture Design Using Wireless for Future Generation Supercomputers," in *Computer Networks & Communications (NetCom) Proceedings of the Fourth International Conference on Networks & Communications*, 2013: Springer, pp. 845-853.

[25] A. S. Shethiya, "Rise of LLM-Driven Systems: Architecting Adaptive Software with Generative AI," *Spectrum of Research,* vol. 3, no. 2, 2023.

[26] H. Azmat and Z. Huma, "Designing Security-Enhanced Architectures for Analog Neural Networks," *Pioneer Research Journal of Computing Science,* vol. 1, no. 2, pp. 1-6, 2024.

[27] A. S. Shethiya, "Redefining Software Architecture: Challenges and Strategies for Integrating Generative AI and LLMs," *Spectrum of Research,* vol. 3, no. 1, 2023.

[28] A. Nishat, "Artificial Intelligence in Transfer Pricing: How Tax Authorities Can Stay Ahead," *Aitoz Multidisciplinary Review,* vol. 2, no. 1, pp. 81-86, 2023.

[29] A. S. Shethiya, "Next-Gen Cloud Optimization: Unifying Serverless, Microservices, and Edge Paradigms for Performance and Scalability," *Academia Nexus Journal,* vol. 2, no. 3, 2023.

[30] A. S. Shethiya, "Machine Learning in Motion: Real-World Implementations and Future Possibilities," *Academia Nexus Journal,* vol. 2, no. 2, 2023.

[31] N. Mazher, I. Ashraf, and A. Altaf, "Which web browser work best for detecting phishing," in *2013 5th International Conference on Information and Communication Technologies*, 2013: IEEE, pp. 1-5.

[32] A. S. Shethiya, "LLM-Powered Architectures: Designing the Next Generation of Intelligent Software Systems," *Academia Nexus Journal,* vol. 2, no. 1, 2023.

[33] H. Azmat and Z. Huma, "Energy-Aware Optimization Techniques for Machine Learning Hardware," *Pioneer Research Journal of Computing Science,* vol. 1, no. 2, pp. 15-21, 2024.

[34] M. Noman, "Machine Learning at the Shelf Edge Advancing Retail with Electronic Labels," 2023.

[35] H. Azmat, "Opportunities and Risks of Artificial Intelligence in Transfer Pricing and Tax Compliance," *Aitoz Multidisciplinary Review,* vol. 3, no. 1, pp. 199-204, 2024.

[36] A. S. Shethiya, "Learning to Learn: Advancements and Challenges in Modern Machine Learning Systems," *Annals of Applied Sciences,* vol. 4, no. 1, 2023.

[37] A. S. Shethiya, "AI-Enhanced Biometric Authentication: Improving Network Security with Deep Learning," *Academia Nexus Journal,* vol. 3, no. 1, 2024.

[38] N. Mazher and H. Azmat, "Supervised Machine Learning for Renewable Energy Forecasting," *Euro Vantage journals of Artificial intelligence,* vol. 1, no. 1, pp. 30-36, 2024.