Intrusion Detection using Explainable Machine Learning Techniques

1st Rishikant Mallick
Department of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, India

2nd Smriti Rout
Department of Computer Science
Siksha'O'Anusandhan
Bhubaneswar, India

3rd Soumyabrata Biswas
Department of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, India

4th Lalit Vashishtha
Department of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, India

5th Santosh Kumar Sahu
GEOPIC
Oil and Natural Gas Corporation
Dehradun, India

2023 OITS International Conference on Information Technology (OCIT) | 979-8-3503-5823-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/OCIT59427.2023.10430641

***Abstract*—With the rise in complex cyber threats, there is a pressing need for accurate and interpretable methods to detect intrusions effectively. This research investigates the fusion of explainable machine learning techniques with intrusion detection, aiming to improve both detection accuracy and the ability to interpret model decisions. The study involves the utilization of various explainable machine learning algorithms, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations), to create models that not only predict intrusions but also provide insights into the features influencing these predictions. The proposed approach is tested on a benchmark intrusion detection dataset, and its performance is compared with traditional machine learning methods. The results demonstrate that the explainable machine learning-based intrusion detection approach achieves competitive detection accuracy while also offering valuable explanations for each prediction. This enhanced interpretability aids cybersecurity experts in comprehending why certain instances are flagged as intrusions. By combining accuracy and transparency, this research contributes to the development of more reliable and understandable intrusion detection systems, thereby bolstering cyber defense strategies in an increasingly intricate digital landscape.**

***Index Terms*—Explainable Artificial Intelligence, Intrusion Detection, Cyber Security**

I. Introduction

In the realm of modern cybersecurity, the integration of artificial intelligence (AI) has ushered in a new era of threat detection and mitigation. However, as AI algorithms become increasingly intricate, their decision-making processes often retreat into obscurity, raising concerns about accountability and interpretability. This has given rise to the burgeoning field of Explainable AI (XAI), which seeks to illuminate the inner workings of these algorithms and enable informed decision-making within the context of cybersecurity. This research paper embarks on a comprehensive exploration of the pivotal role that XAI plays in enhancing the effectiveness of diverse classification and regression methods within the cybersecurity domain. In an era where cyber threats have grown in sophistication, understanding the underpinnings of AI-generated predictions becomes paramount. To this end, our study encompasses a diverse array of classification and regression techniques, ranging from traditional models to state-of-the-art deep learning approaches. Each method is meticulously dissected, not only for its predictive prowess but also for its interpretability. By employing various XAI techniques, we aim to shed light on the decision-making processes of these AI models, unraveling the complex features and data interactions that drive their predictions. Furthermore, we delve into the practical implications of such interpretability, focusing on the empowerment it grants cybersecurity professionals in rapidly identifying and mitigating emerging threats.

II. Literature Review

A novel approach employing a deep quantum neural network based on single-qubit encoding for efficient quantum image classification as approached by [1] . This approach mirrors traditional convolutional neural networks (CNNs) used in classical deep learning while reducing parameter requirements compared to previous quantum image classification models. The study showcases the viability of this proposal by achieving classification accuracies of 94.6%, 89.5%, and 82.5% on subsets of MNIST, Fashion-MNIST, and ORL face datasets, respectively, in noisy simulation environments similar to the NISQ era. Similarly, Artificial Intelligence (AI) pervades daily life, but its opacity raises concerns, particularly in CyberSecurity where trusting unexplainable AI decisions poses risks.

A survey on the integration of Explainable Artificial Intelligence (XAI) in the realm of cyber security [2]. It examines how AI, specifically Machine Learning (ML) and Deep Learning (DL), is employed for tasks like intrusion detection and malware identification. However, the opacity of many AI models hinders understanding their decisions. XAI principles aim to make AI more transparent and interpretable, enhancing user trust. This survey fills a gap by focusing on XAI's

application in cyber security, offering insights into challenges, frameworks, and datasets.

A two-stage pipeline for network intrusion detection, aiming to enhance the system's accuracy and interpretability [3]. In the first stage, they used an XGBoost model for supervised detection of malicious network traffic. To explain this model's decisions, they employed the SHAP framework. In the second stage, they designed an anomaly-based system using a deep autoencoder to identify deviations from the model's behavior during training.

A review on the integration of quantum and classical machine learning while addressing the rising importance of network security due to increased cyber network usage [4] . Intrusion detection systems (IDSs) are essential for securing networks, and the paper proposes a SHAP-based framework to enhance their interpretability. This framework provides local and global explanations for IDS predictions, aiding cybersecurity experts in understanding and building trust.

The study compares quantum and classical machine learning algorithms and explores the potential of quantum computing for improving machine learning tasks [5]. Additionally, the paper proposes an XAI model for in-vehicle intrusion detection systems (IV-IDS) to enhance trust and transparency. The proposed model, VisExp, uses the SHAP method for explanation, and a user survey shows increased trust in the AI-based IV-IDS with explanatory insights.

The study introduces a Quantum Support Vector Machine (QSVM) model for optimizing urban services like mobility, security, and healthcare [6]. This model utilizes quantum computing capabilities to enhance tasks such as identifying DDoS attacks in smart micro-grids. Real DDoS attack data validates the model's effectiveness. The paper concludes by discussing the potential of merging quantum computing and machine learning, while acknowledging existing challenges.

The cybersecurity community is adopting Machine Learning (ML) to counter evolving threats [7]. To ensure the successful integration of these models, it's crucial for domain experts and users to understand and trust their functioning. As black-box models are increasingly used for critical predictions, the demand for transparency and explainability grows. This is especially important in cybersecurity, where detailed insights are needed beyond binary outputs. Recent research has focused on enhancing explainability methods, attacking interpreters in white-box settings, and defining explanation properties.

The study investigates quantum computing's parallelism for faster machine learning, focusing on quantum algorithms' potential in real-world applications like classification and clustering [8]. Explainable AI (XAI) is explored in diverse disciplines, with historical shifts in explanation focus. XAI methods are classified by timing and provide global/local insights. Security and reliability are crucial for XAI adoption. Real-world XAI tests reveal the importance of explanations, despite challenges including security risks and adversarial attacks. The study offers a comprehensive security analysis of explanation use, covering attacks like membership inference, model extraction, and poisoning. Unified framework, real-world datasets/models validate findings, highlighting cybersecurity risks tied to counterfactual explanations.

A review on literature from 2017 to 2022, exploring quantum machine learning in intrusion detection systems (IDS) [9] . They focused on quantum algorithms, especially hybrid models like quantum support vector machines and quantum neural networks.Their findings showcased quantum's advantages, like quicker training and better accuracy in spotting malicious network activity. This indicates quantum computing's potential for improving machine learning in intrusion detection.As internet complexity grows, cyberattacks on DNS rise. Traditional methods fall short, leading to AI solutions. Initially, rule-based, case-based, and ML approaches were used. Advanced ML models improved predictions but lacked transparency.

As IoT becomes more pervasive, cybersecurity challenges escalate due to constant connectivity and resource limitations [10] . The research proposes an XAI-powered framework combining Deep Learning (DL) and XAI techniques (SHAP, RuleFit, LIME) to improve the interpretability of DL-based Intrusion Detection Systems for IoT. Testing on real datasets validates its effectiveness against various attacks. Key contributions include a novel framework, DL-based architecture for IoT security, integration of XAI methods, and validation on NSL-KDD and UNSW-NB15 datasets. The paper's structure encompasses related works, the proposed framework, performance evaluation, and conclusions.

III. Methodology

Explainable AI (XAI) is imperative in cybersecurity to unravel the intricate decisions made by AI systems. As AI increasingly drives threat detection and response, XAI offers transparency, helping cybersecurity experts understand, validate, and trust these automated processes. The transparency provided by XAI enhances accountability, aids in detecting potential biases or vulnerabilities in models, and empowers decision-making. In a landscape where rapid, accurate responses are vital, XAI bridges the gap between advanced AI technologies and the need for comprehensible, reliable cybersecurity measures.

1) LIME (Local Interpretable Model-Agnostic Explanations): It is a machine learning technique that provides local, easily understandable explanations for model predictions by approximating a complex model's behavior in a simplified manner, helping to make black-box models more interpretable.

2) SHAP (SHapley Additive exPlanations): It is a technique in Explainable Artificial Intelligence (XAI) that quantifies the contribution of each feature to a model's prediction. It offers a unified, game-theoretic approach for explaining complex model outcomes in a comprehensible manner.

Machine learning techniques involve creating algorithms that enable computers to learn and make predictions from data. Explainable AI (XAI) focuses on making AI decisions

interpretable to humans. This is crucial due to the complexity of modern AI models, ensuring accountability, complying with regulations, detecting biases, improving models, fostering human-AI collaboration, and facilitating education. XAI techniques simplify complex models, reveal decision factors, and enhance trust, making AI systems transparent, ethical, and effective across various domains.

3) Support Vector Machines (SVMs):: Support Vector Machines (SVM) is a supervised machine learning technique suitable for classification and regression. It excels at binary classification by determining the optimal decision boundary, or hyperplane, between data points. Especially effective in high-dimensional spaces, SVM identifies support vectors—closest points to the decision boundary. The margin, or space around the hyperplane, is established based on these vectors.

4) K Nearest Neighbour(KNN): KNN is a supervised machine learning algorithm that can be used for classification and regression tasks. It works by finding the k most similar data points to a new data point and then assigning the new data point to the class of the majority of the k nearest neighbors. KNN is a simple and effective machine learning algorithm that can be used for a variety of tasks. It is particularly well-suited for problems where the data is not linearly separable or where the use of other machine learning algorithms is not feasible.

5) Random Forest: It is a machine learning ensemble method. It combines multiple decision trees to enhance accuracy and mitigate overfitting. Each tree is trained on a random subset of the data and provides a prediction. The final outcome is determined by aggregating the predictions of all trees. This approach is valuable for classification and regression tasks due to its robustness and generalization ability.

6) Gradient Boosting: It is a machine learning ensemble technique that sequentially builds a strong predictive model by combining the outputs of multiple weak models. It does this by emphasizing the correct prediction of instances that previous models struggled with. It minimizes errors by adjusting weights during training, creating a robust final model. Gradient boosting is widely used for tasks like classification and regression due to its ability to handle complex relationships within data.

7) Adaptive Boosting: AdaBoost is a boosting algorithm that creates a strong classifier by combining weak classifiers. It works by training a series of weak classifiers and then weighting the predictions of each classifier according to its accuracy. The final prediction is made by combining the predictions of all the classifiers.

A. NSL-KDD Dataset

In the field of network security and intrusion detection research, the NSL-KDD dataset is a crucial resource. NSL-KDD, which is derived from the KDD Cup 1999 dataset, addresses the shortcomings of its forerunner by getting rid of redundant information and adding a wider variety of attack scenarios. It includes information about network traffic that reflects both legitimate and illicit activity and is essential for the creation and evaluation of intrusion detection systems. The dataset is divided into training and testing subsets and provided with binary labels indicating the presence of attacks.

B. CICIDS 2017 Dataset

The CIC IDS 2017 dataset is derived from actual network activity and is divided into two sections: one with four attack-launching machines and the other with ten victim machines. 50 GBytes of raw data are included in this dataset in PCAP files, along with 84 features that are listed in CSV files. It includes a sizable 2,830,743 instances and captures network activity over a 5-day period. The dataset divides network traffic into 15 groups, including both regular traffic and 14 different attack methods.

C. UNSWNB15 Dataset

Raw network packets from the UNSW-NB15 dataset were painstakingly assembled in the Cyber Range Lab of UNSW Canberra, fusing real-world contemporary activities with artificially generated attack behaviors using the IXIA PerfectStorm program. The tcpdump tool was used to capture 100 GB of raw traffic, producing Pcap files. The dataset includes class labels produced by twelve algorithms and 49 characteristics. With 175,341 and 82,332 records, respectively, a segmentation produces a training set and testing set that include various attack and typical instance records.

IV. Implementation

In the implementation phase of our research, we utilized three distinct datasets: CICIDS 17, NSL-KDD, and UNSWNB15. To ensure the reliability of our subsequent analyses, we initiated the process with comprehensive data cleaning to remove any inconsistencies or errors present in the datasets. Subsequently, we performed rigorous normalization to standardize the data, ensuring that it is on a consistent scale and format. Additionally, we strategically partitioned the data, possibly into training, validation, and testing sets, to ensure that the model's performance assessment is accurate and representative of real-world scenarios.

Within this framework, we incorporated eXplainable Artificial Intelligence (XAI) techniques, including LIME and SHAP. LIME allowed us to provide local, interpretable explanations for individual model predictions, helping us understand how our machine learning algorithms arrived at specific outcomes. SHAP, on the other hand, offered a holistic view by quantifying the feature contributions to model predictions. These steps ensured our research produced interpretable and transparent results, enhancing insights into the decision-making processes of these models. This comprehensive approach exemplifies the sequence of steps that our study followed, showcasing how these initial data preparation steps, coupled with XAI techniques, fit within the broader process.

In our pursuit of meaningful results, we tailored our approach to evaluation techniques. Specifically, we selected evaluation methods, denoted as, that align with the distinctive characteristics of each dataset. This strategy ensured that our

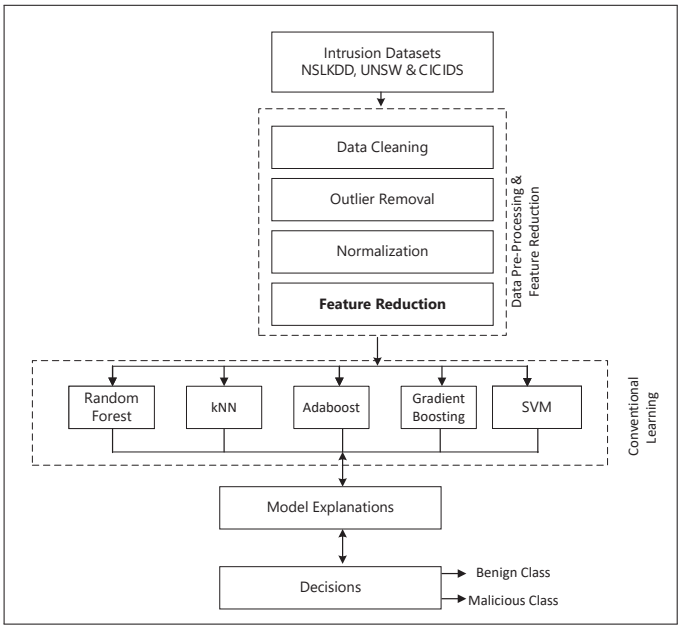
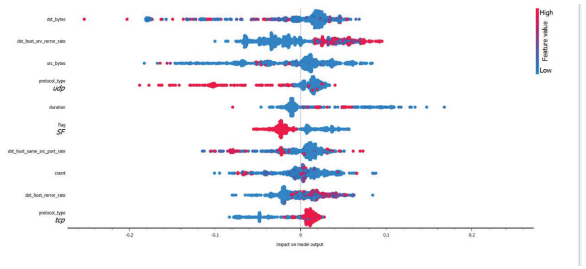


Fig. 1: Data flow diagram of the study

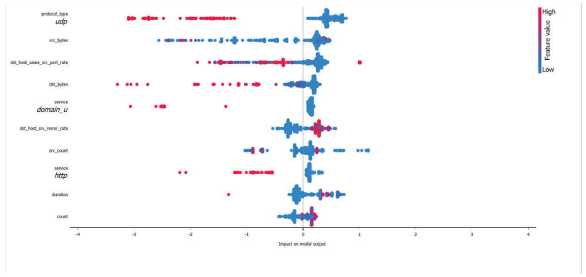
findings are both statistically robust and contextually relevant within the context of our data cleaning, normalization, and incorporation of XAI techniques such as LIME and SHAP.

V. Results and Discussions

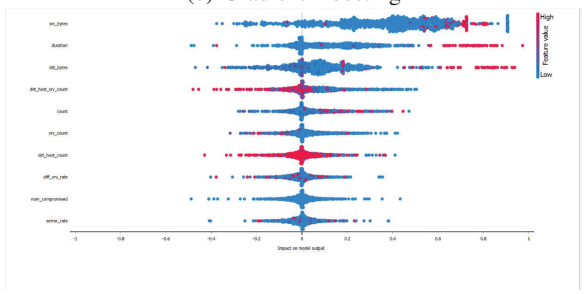
In light of the growing threat of cyberattacks, our research has effectively tackled the vital requirement for precise and transparent intrusion detection systems. Our work provides a promising path in the field of cybersecurity by fusing the capabilities of interpretable models like LIME and SHAP with the power of machine learning. In addition to maintaining high detection accuracy, this integration outperforms traditional black-box techniques by providing insight into the variables affecting model predictions. It highlights the possibility of a paradigm change in intrusion detection and stresses the significance of accuracy and openness. Although there are still issues with scalability and model complexity, our work advances our understanding of intrusion detection. The combination of transparency and accuracy can yield more effective cyber defense measures, which in turn can strengthen cybersecurity and allow for a proactive response to the increasingly complex and dynamic character of contemporary cyber threats.



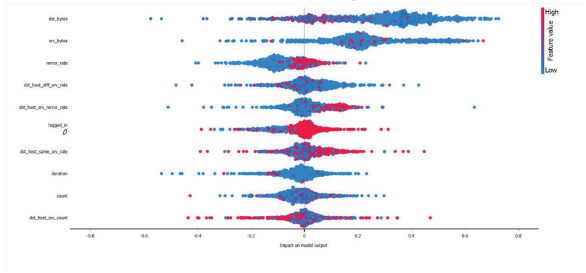
(a) Random Forest



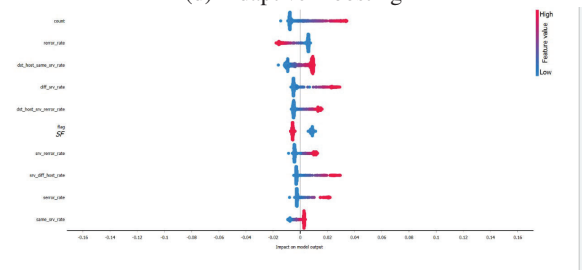
(b) Gradient Boosting



(c) K-Nearest Neighbors

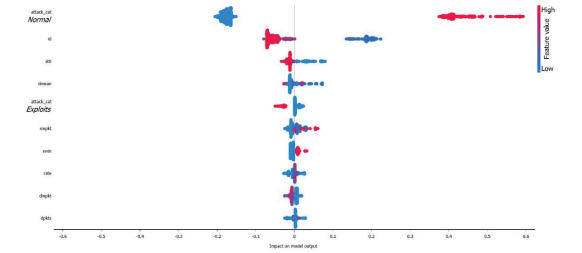


(d) Adaptive Boosting

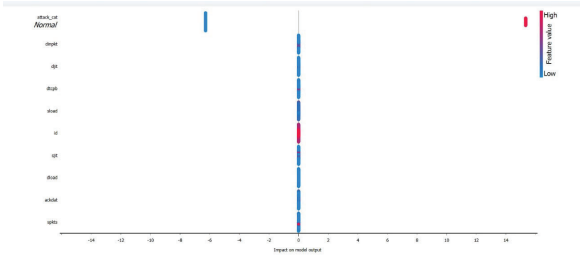


(e) Support Vector Machine

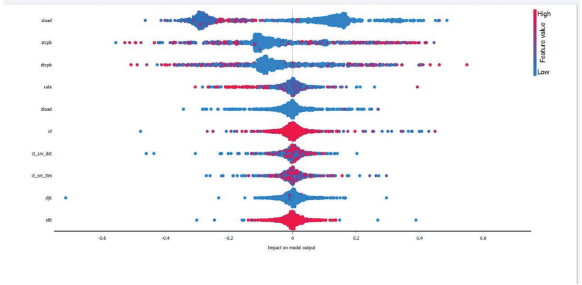
Fig. 2: Results of (a) RF, (b) GB, (c) kNN,(d) AdaBoost, (e) SVM using NSLKDD dataset.



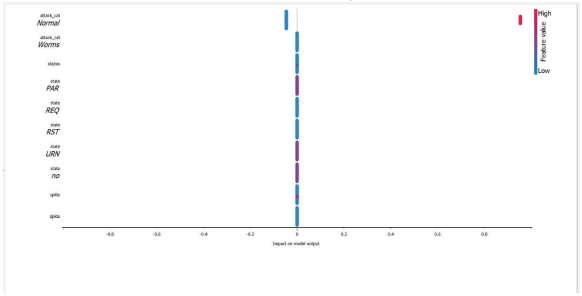
(a) Random Forest



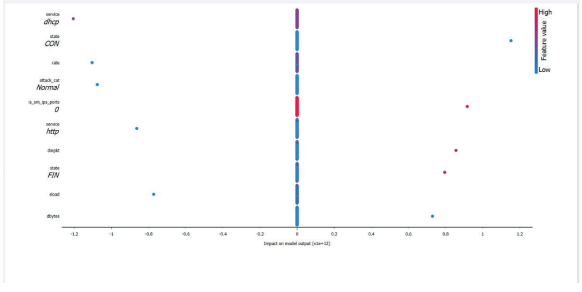
(b) Gradient Boosting



(c) K-Nearest Neighbors

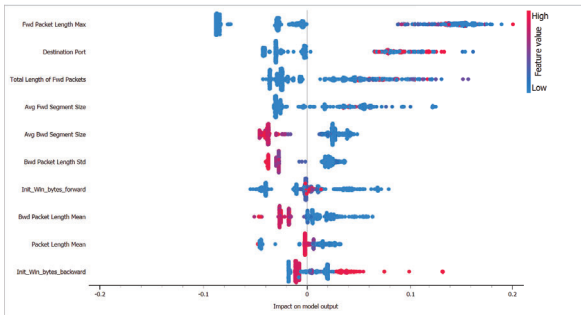


(d) Adaptive Boosting

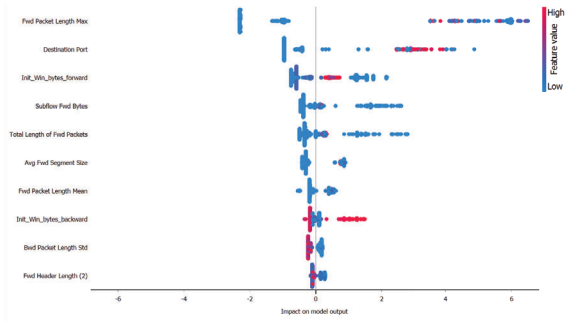


(e) Support Vector Machine

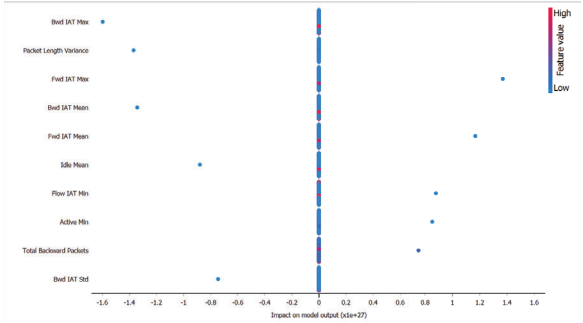
Fig. 3: Results of (a) RF, (b) GB, (c) kNN, (d) AdaBoost, (e) SVM using UNSW-NB15 dataset.



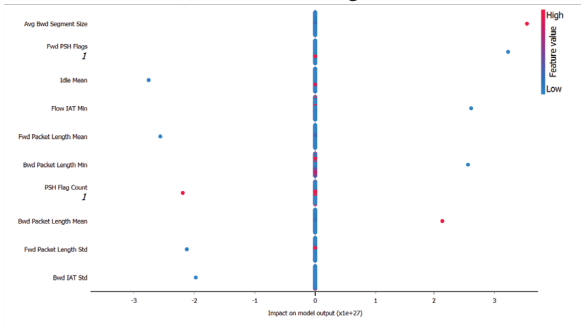
(a) Random Forest



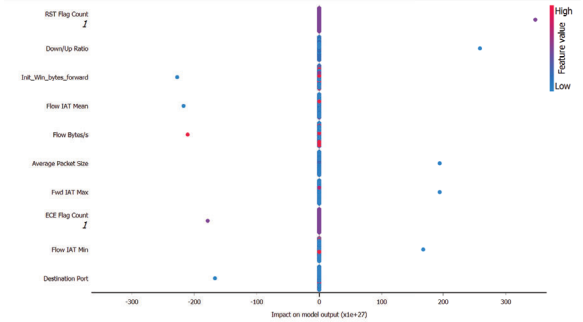
(b) Gradient Boosting



(c) K-Nearest Neighbors



(d) Adaptive Boosting



(e) Support Vector Machine

Fig. 4: Results of (a) RF, (b) GB, (c) kNN, (d) AdaBoost, (e) SVM using CICIDS-17 dataset.

VI. Conclusion

The primary objective of addressing the imperative need for accurate, reliable, and transparent intrusion detection systems in the face of escalating cyber threats. By marrying the power of machine learning with the interpretability of explainable models, this research has illuminated a promising path forward in the field of cybersecurity.

The investigation showcased the successful integration of explainable machine learning algorithms, such as LIME and SHAP, into intrusion detection models. This integration not only maintained competitive levels of detection accuracy but also transcended traditional black-box approaches by providing insights into the factors driving model predictions. This breakthrough significantly enhances the trustworthiness of intrusion detection systems, as cybersecurity experts can now comprehend the decision-making process behind flagged intrusions.

The results underscore the potential of explainable machine learning techniques to revolutionize intrusion detection strategies. However, it's important to acknowledge the ongoing challenges, including fine-tuning model complexity, scalability, and adaptation to evolving threat landscapes.

Ultimately, this study contributes to the paradigm shift from mere detection to comprehensive understanding. The synergy between accuracy and transparency holds the key to more effective cyber defense mechanisms. As organizations strive to safeguard their digital assets, the adoption of these explainable machine learning techniques could play a pivotal role in fortifying their cybersecurity posture and fostering a proactive response to the dynamic and sophisticated nature of modern cyber threats.

References

[1] Nicola Capuano; Giuseppe Fenza; Vincenzo Loia; Claudio Stanzione, "Explainable Artificial Intelligence in CyberSecurity," IEEE Access, 2022, pp. 93575 - 93600 doi:10.1109/ACCESS.2022.3204171.

[2] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani and Chan Yeob Yeun, "Explainable Artificial Intelligence Applications in Cyber Security," in IEEE Access, vol:10, 2022, pp. 93104 - 93139. doi: 10.1109/ACCESS.2022.3204051.

[3] Pieter Barnard, Nicola Marchetti and Luiz A. DaSilva, "Robust Network Intrusion Detection Through Explainable Artificial Intelligence (XAI)", pp. 167 - 171, Vol: 4 , doi: 10.1109/LNET.2022.3186589.

[4] E. H. Houssein, Z. Abohashima, M. Elhoseny, and W. M. Mohamed, "An Explainable Machine Learning Framework for Intrusion Detection Systems," IEEE Access, vol. 10. Oct , 2022, doi: 10.1109/ACCESS.2022.3208573.

[5] Carolina Sanchez Hernandez; Samuel Ayo; Dimitrios Panagiotakopoulos, "Experimental Analysis of Trustworthy In-Vehicle Intrusion Detection System Using eXplainable Artificial Intelligence (XAI)," 2021, IEEE/AIAA 40th Digital Avionics Systems Conference (DASC),doi: 10.1109/DASC52595.2021.9594341.

[6] Carolina Sanchez Hernandez, Samuel Ayo And Dimitrios Panagiotakopoulos, "An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools," Energies (Basel), Nov. 2021, doi: 10.1109/DASC52595.2021.9594341.

[7] Aditya Kuppa; Nhien-An Le-Khac, "Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security," 2020. doi: 10.1109/IJCNN48605.2020.9206780 .

[8] Aditya Kuppa, and Nhien-An Le-Khac, " Adversarial XAI Methods in Cybersecurity " IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4924 - 4938, 2021, doi: 10.1109/TIFS.2021.3117075.

[9] Nida Aslam,Fatima M. Anis ,Irfan Ullah Khan , Samiha Mirza , Alanoud AlOwayed , G Reef M. Aljuaid and Reham Baageel, "Interpretable Machine Learning Models for Malicious Domains Detection Using Explainable Artificial Intelligence (XAI)," in 2022, visit: https://doi.org/10.3390/su14127375 .

[10] Zakaria Abou El Houda , Bouziane Brik and Lyes Khoukhi, " Why Should I Trust Your IDS? " IEEE OPEN JOUNAL OF THE COMMUNICATIONS SOCIETY, vol. 3, pp. 1164 - 1176, July. 2022, doi: 10.1109/OJCOMS.2022.3188750.